eman ta zabal zazu

Universidad del País Vasco
Euskal Herriko Unibertsitatea

Facultad de Economía y Empresa
Departamento de Economía Aplicada III
(Econometría y Estadística)

La emigración cubana a Estados Unidos desde una perspectiva econométrica

Aleida Cobas Valdés

Directores
Ana Fernández Sainz
Javier Fernández Macho

Bilbao

6 de abril de 2017

«*No puede haber historia donde no existen recuerdos a los cuales aferrarse*»

Reinaldo Arenas

Agradecimientos

Después de un largo camino, por fin puedo escribir estas páginas. He dejado el apartado de Agradecimientos para el final como el reto que se deja para, una vez alcanzado, sentir el alivio de que se han superado todas las pruebas. No ha sido para mí la tesis doctoral una meta profesional en sí misma, sino un ejercicio de constancia, de perseverancia y de querer aprender cada día.

Agradecer en primer lugar a mis directores de tesis Ana y Javier, por la confianza depositada y por haberme dado la oportunidad de adquirir conocimientos y habilidades en la labor de investigación que sin ellos hubiese sido imposible alcanzar. No existen muchos voluntarios a querer dirigir una tesis doctoral, gracias Ana por aceptar emprender este camino cuando ya había tirado la toalla y por darme la oportunidad de volver a creer que era posible; gracias por compartir conmigo tu tiempo, tus conocimientos y darme siempre tu apoyo y tu fuerza. Gracias Javier por aceptar trabajar con nosotras y por permitirme sumar a mi trabajo tu valiosa experiencia. Les estaré siempre agradecida. Gracias también por la paciencia demostrada, porque lidiar con una cubana no es tarea fácil.

A todos y cada uno de mis compañeros del Departamento de Economía Aplicada III, siempre me habéis ofrecido vuestra ayuda desinteresada y habéis estado pendientes de la evolución de esta tesis. No quiero poner nombres, tendría que escribir los de todos los que están actualmente en el departamento y de aquellos que ya no están porque se han jubilado o han cambiado de trabajo. Gracias también por todo lo que aprendí junto a vosotros en los años en que fui docente. Agradecer además al Departamento como institución por el apoyo brindado en la difusión de los trabajos realizados en esta tesis.

A Isabel Urrutia e Inés García por la constante preocupación y por brindarme la mano cada vez que lo he necesitado. A todas mis compañeras y compañeros en el edifico Zubiria y a los más cercanos como, Brenda, María Jesús, Clara, Federica, Marisa, Enara y Roberto. A los profesores del Departamento de Fundamentos del Análisis Económico II y a la profesora Azucena Vicente, del Departamento de Economía Financiera II, por sus constantes muestras de cariño y apoyo.

A mi esposo y a mi hijo por acompañarme en esta aventura y por padecer mis malos momentos en el transcurso de la tesis. A mi hijo en especial, porque ha sido mi principal motivo para terminar. Gracias mi niño, por ofrecerte tantas veces a ayudarme, pronto vas a saber más de R y de Látex que tu mamá. Siempre estaré en deuda con vosotros.

A mis padres, porque si he llegado hasta aquí ha sido por ellos, por sus valiosas e

insustituibles enseñanzas en valores y en sacrificio. A mi hermana por su inestimable ayuda en cada momento. A mi tío Juanito, que nunca dudó que prodría lograrlo. A ese ángel de la guarda que me ha ayudado muchísimo sin pedir reconocimiento a cambio, que sepas que no necesito escribir tu nombre en estas líneas porque le tengo escrito en el corazón. Gracias.

Al Econometrics Research Group por el apoyo institucional para presentar parte de mi trabajo en Congresos Internacionales. A todos los participantes en los diferentes Congresos y Jornadas en los que he presentado mis propuestas, por sus comentarios encaminados a mejorar y enriquecer mi trabajo.

A Benjamin Hofner por responderme siempre con rapidez y guiarme en la utilización del paquete mboost. A Blaise Melly por haberme facilitado el código R con el que pude hacer el contraste de independencia en el modelo de regresión cuantílica con selección muestral.

A Manuel Arellano, por aclararme más de una duda sobre la selección muestral en la regresión cuantílica. A James Albrecht, Alejandro Badel y anónimos evaluadores de los artículos publicados. A Liam Tze Lim, por poner a disposición de todos su experiencia en Látex.

A mis amigas y amigos, los que he ganado en el País Vasco y los que dejé en Cuba. En especial a Cristina, Miguel, Betty, Henry, Azucena, Alexandra, Alicia, Aidita, Gertrudis, Massiel, Mayte y Vanessa. A todos mis queridos vecinos de Cuba, que son mi familia.

Índice general

Índice de figuras

Índice de cuadros

A mi familia, a los que están y a los que se fueron. A todos los cubanos que murieron en el intento de alcanzar sus sueños

Parte I

Introducción

Introducción

Se estima que cerca de 244 millones de personas, en el año 2015, vivían en un país diferente al que les vio nacer; concentrándose el 58 % de ellas en los países desarrollados. El fenómeno de la emigración acapara interés desde diferentes disciplinas, que intentan explicar los tipos de procesos migratorios, sus causas y repercusiones para los países emisores y de acogida.

Según la teoría neoclásica, las migraciones son el resultado de decisiones individuales, tomadas por actores racionales que buscan aumentar su bienestar al trasladarse a lugares, donde la recompensa por su trabajo es mayor; en una medida lo suficientemente alta que les permita compensar los costes de su desplazamiento (Harris and Todaro, 1970; Sjaastad, 1962). Esta teoría, dominante hasta los años 70 del siglo XX, coincidió con etapas de crecimiento económico, la internacionalización de la actividad económica, el proceso de descolonización y el desarrollo emergente de algunos países subdesarrollados (Arango, 2003).

Otra teoría basada en el modelo de desarrollo de Lewis (1954) sostiene que las migraciones permiten el desarrollo de las economías tanto para los países receptores como emisores, una vez que los excedentes de mano de obra se desplazan a aquellos países más desarrollados, permitiéndoles una expansión, sin que ello traiga consigo un aumento de los salarios. Por otro lado, a los países emisores les permite desprenderse de ese excedente y avanzar en una relaciones capital-producto más altas.

En los años 60 y 70 del siglo pasado, surgió además una corriente con matices marxistas, que postulaba la idea de que el capitalismo había traído consigo un núcleo formado por países industrializados y una periferia de países pobres, vinculados a ese núcleo a través de relaciones asimétricas, provocando esas diferencias los desplazamientos de la población de los países pobres hacia los ricos (Balibar and Wallerstein, 1991).

En las últimas décadas, la composición de los flujos migratorios así como sus características, han experimentado cambios. América Latina, Asia y África han pasado a sutituir a Europa como principales emisores y se han incrementado las restricciones a la libre circulación de

personas, lo que ha incentivado la emigración ilegal y el tráfico de seres humanos.

Las grandes brechas existentes en las condiciones de vida y en las expectativas de los individuos entre los diferentes países, las guerras, las persecuciones por motivos políticos y religiosos, factores climáticos y medioambientales, son algunas de las actuales causas de la migración.

A su vez, los países receptores aceptan la llegada de personas de otros países por razones demográficas (al contar con bajas tasas de natalidad y un alto grado de envejecimiento de su población), por atraer a aquellos con una alta cualificación o por motivos humanitarios.

La migración abarca diferentes etapas: la toma de la decisión de emigrar con la consiguiente preparación; el acto en sí de emigrar; el asentamiento que se entiende va desde que el emigrante llega al país de acogida hasta que resuelve los problemas mínimos de subsistencia y la integración que conlleva un proceso de inmersión e incorporación a la nueva cultura, asumiéndola como propia.

Entre 1990 y 2015 el número de personas que emigró de su país de origen se incrementó en 91 millones de personas, lo que representó un aumento del 60 %. El mayor receptor de inmigración es Europa (76 millones), seguido de Asia (75 millones) y de Estados Unidos (54 millones). Lejos se ubican Africa (21 millones), América Latina y el Caribe (9 millones) y Oceanía (8 millones) (UN, 2015).

Dentro de este flujo migratorio creciente, han estado presentes, desde los años 30 del siglo pasado, los cubanos; momento a partir del cual, Cuba cambió su patrón migratorio al dejar de ser un país de inmigrantes, para convertirse en un país de emigrantes. La migración desde Cuba ha estado orientada históricamente hacia Estados Unidos, país donde viven más de 2 millones de personas de origen cubano.

En Cuba plantearse emigrar del país, es sinónimo de pensar en cómo llegar a Estados Unidos, por cualquier vía, ya sea legal o ilegal, por mar, tierra o aire. En los momentos en los que escribo estas líneas, varios cientos de cubanos se encuentran varados en México o en algún país de Centroamérica, en un peregrinaje que hasta hace pocas semanas les conducía a la residencia en Estados Unidos.

Otros muchos cubanos hubiesen estado en medio del Estrecho de la Florida, de no ser por la derogación de la ley denominada “pies secos, pies mojados” por el Presidente Barack Obama, el 12 de Enero del presente año.

El tema central de esta tesis es la emigración cubana a Estados Unidos, en un intento de contribuir a explicar este proceso, desde un punto de vista empírico; utilizando para

ello las herramientas que la Estadística y la Econometría modernas nos ofrecen. Es un ejercicio que constituía para mí una deuda con los cubanos, como testigo de tantos hechos relacionados con la emigración y sobre todo del anhelo diario de tantas personas por salir de Cuba y llegar al "otro lado".

La historia de la emigración cubana a Estados Unidos en estas últimas décadas ha estado influenciada por las relaciones hostiles entre ambos países desde el triunfo de la Revolución cubana en 1959, produciéndose conocidas oleadas migratorias que se han diferenciado entre sí por la composición de las personas que han participado en ellas.

Intentar entender las causas de la emigración desde Cuba, no es objetivo de esta tesis y es un tema que de por sí requeriría una dedicación exclusiva, por su complejidad y por la cantidad de hechos y circunstancias presentes a lo largo de los años.

Desde Cuba, el fenómeno de la emigración se plantea como resultado de la política de descrédito de Estados Unidos y por el mantenimiento del bloqueo comercial, lo que ha erosionado la capacidad productiva del país.

Estados Unidos, en su política de bloqueo a Cuba, implementó la Ley Torricelli en 1992 y la Ley Helms-Burton en 1996, privando a Cuba durante años del acceso a fuentes financieras, al mantenimiento de relaciones comerciales con empresas norteamericanas y de otras partes del mundo. Además, los cubanos son tratados en Estados Unidos como refugiados y, amparados en la Ley de Ajuste Cubano de 1966, reciben la residencia permanente al año y un día de arribar a Estados Unidos.

Sin embargo, en Cuba además del bloqueo norteamericano, ha estado presente una política económica basada en el modelo socialista, que lejos de favorecer el crecimiento económico del país, ha agravado su situación. Con el fin de la era socialista en Europa, se produjo un gran deterioro del nivel de vida de los cubanos, el agravamiento de las limitaciones de consumo y de las libertades individuales, la pérdida de confianza en el proyecto revolucionario, un marcado descenso de la productividad, el retraso tecnológico y el declive de sectores insignias de la Revolución cubana como la Salud Pública y la Educación.

La Libertad Económica de un país se entiende por el derecho natural de la persona a ser dueña del valor que genera, que para los países socialistas es el principio: "de cada cual según su capacidad, a cada cual según su trabajo"(Dixon and Macarov, 2016). El índice de Libertad Económica (Miller et al., 2015) muestra el desempeño de un país en cuanto a crecimiento económico, ingresos per cápita, atención médica, educación, protección del medio ambiente, esfuerzos en la reducción de la pobreza y bienestar en general. A pesar de los grandes esfuerzos desarrollados en la Educación (Fox and Byker, 2015) y en la Salud

Pública (Atun et al., 2015), Cuba ocupa el puesto 177 de este índice calculado para 178 países, sólo delante de Corea del Norte.

El conjunto de estas circunstancias, unido a lo que algunos sociólogos denominan la "dimensión subjetiva de la pobreza", que hace que los cubanos se sientan aún más pobres de lo que son, al compararse con otros individuos de la región (Pumar, 2013) agravan el fenómeno migratorio cubano.

Los cubanos son considerados un grupo de éxito en Estados Unidos, siendo su enclave principal el estado de la Florida y dentro de éste, Miami. Son los cubanos un grupo consolidado en los aspectos culturales, políticos y económicos dentro de ese país (Portes, 2010).

La emigración hacia Estados Unidos ha sido considerada, en general, fuente de crecimiento poblacional, económico y de cambio cultural. A pesar de su impacto y tratamiento en la literatura, no es en Estados Unidos, donde más población ha nacido en el extranjero, en términos relativos. En el año 2013, el 43.8 % de la población de Luxemburgo había nacido en el extranjero, seguida de Australia con un 28.3 %, Israel con un 22.6 %, Nueva Zelanda con un 22.4 % y Canadá con un 20 %. Estados Unidos se colocaba en el centro de la distribución con un 13.1 %; cifra parecida a la de Alemania y España.

América Latina, región cercana geográficamente a Estados Unidos, donde viven cerca de 600 millones de personas, ha sido en las últimas décadas emisor natural de emigración a ese país, siendo México el país que más emigración genera hacia Estados Unidos, con una migración que se autoselecciona de forma negativa, en cuanto a nivel de educación (Cuecuecha, 2010), lo que se traduce en que son las personas con menos cualificación las que toman la decisión de salir de su país.

Los mexicanos representan el 28 % del total de inmigrantes en Estados Unidos y los cubanos el 3 %. Sin embargo, comparados con el total de población en el país de origen, los mexicanos son el 9.4 % y los cubanos el 20 % de sus respectivas poblaciones.

Una parte importante de la fuerza laboral de Estados Unidos es inmigrante. En 2014, por ejemplo, el 48.3 % de la fuerza de trabajo era de origen hispano; siendo los cubanos el tercer grupo en cuanto a tamaño dentro de la fuerza laboral hispana, representando el 5 % de los trabajadores en general (US Bureau of Labor Statistics, 2015).

Es México el principal suministrador de fuerza de trabajo de bajo coste salarial para la economía norteamericana. Los cubanos, en cambio, al contar con niveles de capital humano más elevados y al estar beneficiados por la Ley de Ajuste Cubano, han creado un dinámico enclave empresarial en Miami, logrando alcanzar posiciones sociales y económicas más

ventajosas que el resto de inmigrantes de origen hispano (Portes et al., 2009).

Para lograr entender la emigración cubana, desde el momento de la toma de la decisión de emigrar hasta la integración en Estados Unidos, se ha analizado en esta tesis en ***primera instancia***, quiénes son los que marchan de Cuba; lo que se explica a través de un modelo de elección binaria en el Capítulo 4 centrando el análisis en la autoselección educativa.

La Educación ha sido siempre en Cuba, uno de los pilares de la Revolución socialista, siendo uno de los países en el mundo que más recursos dedica a este sector (Gasperini, 2000), por lo que resulta de interés investigar el nivel de educación de las personas que emigran y cómo influye ese nivel de educación en la probabilidad de emigrar.

Es sabido que si existe una relación positiva entre la probabilidad de emigrar y los años de educación, se produce una autoselección positiva entre los individuos que deciden emigrar. Si este proceso no es puntual sino continuado, representa un pérdida de capital humano para el país de origen (Albo and Díaz, 2011).

El problema de la autoselección de los cubanos por su nivel de educación no ha sido tratado en profundidad en la literatura, sólo aparece en Borjas (1991) en el contexto de la emigración hacia Estados Unidos desde diferentes países, incluida Cuba.

En ***segundo lugar*** se ha querido indagar en la asimilación económica de los cubanos en Estados Unidos a partir del análisis de la distribución de los salarios de los inmigrantes cubanos en ese país; cuantificando el efecto de diferentes variables socioeconómicas sobre el salario por hora, como los años de educación, la experiencia potencial, el dominio del idioma inglés, el estado civil y la edad del individuo en el momento de emigrar, entre otras.

En el estudio del proceso de asimilación económica de los inmigrantes existe un considerable número de trabajos (Borjas, 1985, 2015a,b; Cortes, 2004; Elliott and Lindley, 2008; Chiswick et al., 2008; Hunt, 2012). Otros trabajos han estado encaminados al estudio de la influencia sobre el salario de características específicas como el nivel de inglés (Chiswick and Miller, 2002, 2015; Adsera and Pytlikova, 2015; Di Paolo and Tansel, 2015) o el nivel de educación (Khan, 1997; Li and Sweetman, 2014; Brunner and Pate, 2016).

Para cumplir este objetivo, se han estimado en el Capítulo 6 dos modelos, utilizando la regresión cuantílica lineal. La regresión cuantílica lineal se debe al trabajo de Koenker and Bassett (1978) y permite describir la distribución condicional de una variable respuesta, como función de diferentes covariables en los diferentes puntos o locaciones de esa distribución, ofreciendo un punto de vista mucho más amplio de la relación entre esa variable respuesta, en este caso el salario y las variables explicativas.

La regresión cuantílica resulta de gran utilidad en general cuando el término de error no es independiente e idénticamente distribuido (*iid*), cuando la variable respuesta no sigue una distribución de probabilidades conocida o cuando presenta colas más largas y existen valores extremos (Koenker, 2005).

En un primer modelo, estimado en la Sección 6.1 se utiliza la regresión cuantílica lineal, empleando el concepto de centercepto (Wainer, 2000), con el cual todas las variables continuas son centradas con respecto a su valor mediano, lo que permite una interpretación económica más realista al recoger en el intercepto el valor condicionado de la función cuantílica para el individuo mediano.

Un segundo modelo, planteado en la Sección 6.2, se estima a través de la regresión cuantílica lineal teniendo en cuenta el problema de la selección muestral y utilizando el test de independencia para modelos de regresión cuantílica con selección muestral propuesto por Huber and Melly (2015).

El análisis de la selección muestral en la regresión cuantílica se debe a los trabajos de Buchinsky (1998a, 2002), en los cuales se extiende el estimador de Newey (1991) a los cuantiles.

El problema de la selección muestral surge cuando la variable de interés es sólo observable para una parte de la muestra que no es aleatoriamente seleccionada. Gronau (1974) and Heckman (1974) fueron los primeros en abordarla desde una perspectiva paramétrica. En el análisis de los salarios, la selección muestral emerge del hecho de poder observar los salarios sólo cuando el individuo trabaja, existiendo por tanto un conjunto de características que no son observables y que afectan la decisión de trabajar y por ende influyen sobre los salarios.

En el Capítulo 9 se tiene en cuenta le regresión cuantílica aditiva estructurada para el análisis de los salarios de los inmigrantes cubanos en Estados Unidos. La utilización de la regresión cuantílica aditiva parte de la necesidad de permitir mayor flexibilidad en el planteamiento de la relación entre los regresores y la variable respuesta en el contexto de la regresión cuantílica.

Los Modelos Cuantílicos Aditivos Estructurados (STAQ) se deben a los trabajos de Fahrmeir et al. (2004), Kneib et al. (2009) y Fenske et al. (2012) y se ha estimado el mismo a través del *Component-wise Functional Gradient Descent Boosting* (Bühlmann and Yu, 2003). El Boosting (Schapire, 1990) fue introducido para tratar problemas de clasificación, extendiéndose luego a la Estadística al tratar problemas de regresión.

La conveniencia de preferir la regresión sobre los cuantiles, ya sea de forma lineal como flexible y no sobre la media, se fundamenta en esta tesis por las siguientes razones:

- El área de interés en el análisis de los salarios no se circunscribe a la media, es importante y necesario estudiar la distribución condicional de salarios en cada una de sus locaciones y en específico en los cuantiles más bajos y los más altos, para poder comparar la influencia de las variables socioeconómicas sobre el salario de los individuos que menos y que más salario perciben.
- En el análisis de los salarios es posible la existencia de outliers o valores extremos y en este caso es preferible la regresión cuantílica a la regresión sobre la media (Koenker, 2005), debido a la robustez inherente de los cuantiles.
- No tiene por qué existir una relación lineal entre las covariables seleccionadas y el salario, por lo que un método de regresión más flexible utilizando la regresión cuantílica aditiva, puede ser útil para captar esa relación. En la práctica, la regresión cuantílica lineal no siempre es suficiente para expresar la relación existente entre las covariables y la función cuantílica de la variable respuesta.
- Contar con un predictor aditivo estructurado en la regresión cuantílica nos permite tener en cuenta un conjunto de efectos que no es posible asumir en la regresión cuantílica lineal, como efectos no lineales suaves, efectos de parámetros cambiantes suaves, efectos lineales y espaciales, entre otros.

Con el fin de justificar y explicar las herramientas econométricas utilizadas en el estudio de la emigración cubana; a lo largo de esta tesis se hace un recorrido por la metodología utilizada. Así en la Parte II se explican los modelos de regresión con respecto a la media, desde el modelo lineal clásico hasta los modelos lineales generalizados.

En la Parte III se fundamenta la regresión cuantílica lineal y en la Parte IV se parte de los modelos aditivos generalizados hasta llegar a los modelos cuantílicos aditivos estructurados, profundizando en la técnica Boosting para la estimación de éstos últimos. Finalmente se resumen las principales conclusiones obtenidas y se citan las formas en las que han dado a conocer los resultados de los diferentes análisis desarrollados en esta tesis, señalando las futuras líneas de investigación.

El resto de esta Parte I correspondiente a la Introducción es estructurada en dos secciones. En la primera se presenta una revisión de los modelos utilizados en la literatura en el estudio de la emigración hacia Estados Unidos y en una segunda sección se describen las etapas del proceso migratorio desde Cuba hacia Estados Unidos.

CAPÍTULO 1

Tratamiento de la Emigración hacia Estados Unidos en la Literatura

Antes de la firma del Acta de Inmigración de 1965, por el Presidente Lyndon B. Johnson, en Estados Unidos se promovía el sistema nacional de cuotas en la política de migración, lo que garantizaba visas principalmente a personas provenientes de los países de Europa Occidental, mayoritariamente de Alemania y Reino Unido. Por ejemplo, en 1929 de las 150.000 visas anuales asignadas, 51.227 correspondieron a Alemania, 100 a Grecia y ninguna a China.

Con la nueva ley, se prohibió la discriminación (hasta el momento imperante) en la emisión de visas, basada en "raza, sexo, nacionalidad, lugar de nacimiento o lugar de residencia". Se estableció un tope anual de 170.000 visas para inmigrantes, sin que ningún país pudiese superar las 20.000 visas. Los cónyuges, hijos menores de edad y padres de ciudadanos estadounidenses, estaban exentos de esas cuotas.

A partir de ese año, el sistema comenzó a basarse en la reunificación familiar y la acogida de refugiados, lo que provocó un considerable incremento de la emigración hacia Estados Unidos de ciudadanos asiáticos y latinoamericanos.

Si en 1950 dos tercios de la inmigración provenía de Europa o Canadá, un cuarto de América Latina y sólo un 6 % de Asia, entre 1976 y 1986 el número de inmigrantes venidos de Asia se incrementó un 663 %, dentro de los que destacaron los indios con un incremento del 3.000 %, los coreanos con un aumento del 1.328 % y los pakistaníes con un incremento del 1.600 %.

En 1990 sólo el 17 % de los inmigrantes provenía de Europa o Canadá, la mitad de América Latina y un 30 % de Asia. Se generó, por tanto un cambio importante en la composición

étnica y racial de la población de Estados Unidos.

Los modelos sobre emigración internacional y en específico sobre la emigración hacia Estados Unidos, han estado centrados en el estudio de las características de los flujos migratorios y en el análisis del proceso de asimilación de los emigrantes así como el impacto económico de este proceso.

1.1. Modelos Gravitacionales

En el estudio de los flujos migratorios, el modelo más extendido ha sido el modelo gravitacional, el cual parte de los modelos espaciales para estudiar la interrelación entre la oferta y la demanda o el estudio de las corrientes de comercio internacional y de capital. Los modelos gravitacionales más simples, en la teoría de la emigración, relacionan el tamaño de la población de los países de origen y destino y la distancia entre ellos.

Modelos más elaborados incluyen variables relacionadas con los denominados factores *pull and push*,[1] como por ejemplo las oportunidades en el país de destino en cuanto a salarios y bajas tasas de desempleo, las condiciones de seguridad ciudadana, seguridad social, el idioma, entre otros. Estos modelos se basan en la maximización de la utilidad esperada aleatoria (RUM).

En los estudios de emigración, el uso de los modelos gravitacionales se debe al trabajo de Sjaastad (1962) donde se compara la utilidad esperada del individuo en el país de origen con la utilidad esperada en el país de destino.

En este sentido, Borjas (1987, 1989) propuso un modelo de análisis de los flujos de emigración incluyendo características de los países emisores y de los países receptores tales como: actividad económica agregada de los países emisores y receptores, factores políticos, económicos y demográficos, nivel de educación del país emisor, etc.

Con este análisis, Borjas demuestra la influencia de estas variables en el número y la composición de los flujos migratorios. Resume su modelo de emigración en la especificación de tres ecuaciones: una para los salarios en el país de origen, otra para los salarios para el país de destino y una tercera que represente los costes de la migración.

Karemera et al. (2000) examina, a partir de un modelo gravitacional, qué factores políticos, económicos y demográficos condicionan el tamaño y la composición de los flujos migratorios

[1]Los factores push son aquellos que fuerzan la emigración desde el país de origen mientras los factores pull incluye aquellos factores que, en el país de destino, son atractivos para los potenciales emigrantes.

hacia Estados Unidos, considerando variables como la distancia entre los dos países, la población, las tasa de inflación y de desempleo, la inestabilidad política y el índice de libertad económica relativa de Gastil (1987)[2].

Clark et al. (2007) examinan los flujos migratorios desde diferentes regiones del mundo hacia Estados Unidos entre los años 1971 y 1998 obteniendo resultados importantes como el hecho de que tanto la distancia como el nivel de inglés son determinantes en la composición de esos flujos migratorios y que el efecto del nivel de educación sobre la probabilidad de emigrar es positivo para los inmigrantes provenientes de países pobres.

Donato et al. (2010) realizan una comparación de la migración que se realiza desde diferentes países de América Latina hacia Estados Unidos profundizando en las causas de la emigración y las características socioeconómicas que la distinguen entre nivel de educación, sexo, estado civil y familiares que dejan en el país de origen.

Mayda (2010) investiga los factores determinantes en los flujos bilaterales migratorios hacia 14 países de la OECD, incluido Estados Unidos desde 1980 hasta 1995. En este artículo se encuentran evidencias a favor del papel de las oportunidades de ingresos en el país de destino sobre las tasas de emigración y al pequeño y no significativo efecto del nivel de pobreza en los países emisores. Con este trabajo se demuestra que el papel de las políticas migratorias, influenciadas por la actitud de los votantes hacia los inmigrantes, la presión de los grupos políticos, la estructura de las instituciones es determinante en la migración, conllevando a que en períodos con leyes de migración menos restrictivas se condicione a los factores pull a tener un efecto mayor.

En términos generales, los modelos gravitacionales describen la utilidad que el individuo i, que vive en el país j en el momento $t-1$, deriva de optar por el país k dentro de un conjunto de países D en el momento t (Beine et al., 2016):

$$U_{ijkt} = w_{jkt} - c_{jkt} + \epsilon_{ijkt} \tag{1.1}$$

w_{jkt} representa el componente determinista de la utilidad y c_{jkt} denota el coste "*time-specific*" de moverse desde j hasta k. Ambos componentes pueden ser modelados a partir de variables observables; ϵ_{ijkt} es el término estocástico.

Los supuestos que se establezcan sobre ϵ_{ijkt} determinan la probabilidad esperada de optar

[2]Raymond Gastil, especialista en estudios regionales de la Universidad de Washington, desarrolló la metodología para poder diferenciar a los países en estudio según su grado de libertad civil y derechos políticos. Este análisis se lleva a cabo desde 1972 y se presenta anualmente como un libro titulado *Freedom in the World*.

por el país k y representará la opción que, para el individuo i, maximiza su utilidad esperada. Así, si se asume que ϵ_{ijkt} es *iid*, entonces:

$$E\left(p_{jkt}\right) = \frac{\epsilon^{w_{jkt}-c_{jkt}}}{\sum_{l\in D}\epsilon^{w_{jkt}-c_{jkt}}} \tag{1.2}$$

donde $p_{jkt} \in [0,1]$ representa la proporción de individuos del país j que se mueven al país k en el momento t.

Sabiendo que s_{jt} representa el número de población que reside en el país j en el momento t, se puede obtener el flujo de personas que se mueven del país j al país k en el momento t:

$$m_{jkt} = p_{jkt}s_{jt} \tag{1.3}$$

El flujo bruto de emigración esperado del país j al país k se obtiene:

$$E(m_{jkt}) = \frac{\epsilon^{w_{jkt}-c_{jkt}}}{\sum_{l\in D}\epsilon^{w_{jkt}-c_{jkt}}} \cdot s_{jt} \tag{1.4}$$

Si se asume que el componente determinista no varía con respecto al país de origen j, entonces la Ecuación 1.4 se puede re-escribir:

$$E\left(m_{jkt}\right) = \phi_{jkt}\frac{y_{kt}}{\Omega_{jt}}s_{jt} \tag{1.5}$$

donde $y_{kt} = e^{w_{kt}}$, $\phi_{jkt} = e^{-c_{jkt}}$ y $\Omega_{jt} = \sum_{l\in D}\phi_{jlt}y_{lt}$. El flujo esperado de migración en la Ecuación 1.5 está directamente relacionada con: (i) el potencial s_{jt} del país de origen j en generar migración, (ii) el atractivo y_{kt} del país de destino k y (iii) la accesibilidad $\phi_{jkt} \leq 1$ del destino k para los potenciales emigrantes del país j. Y está inversamente relacionada con (iv) Ω_{jt} que captura el componente determinístico de la utilidad de no emigrar.

Bertoli and Moraga (2015) proponen estimar un modelo gravitacional relajando la hipótesis de independencia de ϵ_{ijkt} entre las diferentes alternativas, demostrando la sensibilidad de los flujos migratorios bilaterales ante políticas restrictivas de terceros Estados.

1.2. Modelos para el estudio del proceso de asimilación

Chiswick (1978) define como medida de asimilación económica de los emigrantes en Estados Unidos a la tasa de convergencia salarial entre nativos e inmigrantes, dando paso a los modelos de análisis del proceso de asimilación de los inmigrantes en el país de acogida.

Utilizando datos de sección cruzada, concluye que el salario de los hombres inmigrantes recién llegados es menor que el de los individuos que llevan varios años en Estados Unidos así como que se produce un crecimiento relativamente rápido de los salarios, llegando éstos a igualar e incluso a superar el salario de los nativos en poco tiempo. Trabajos posteriores reafirmaron la tesis de Chiswick (1978) como los artículos de Long (1980), Carliner (1980) y Borjas (1982).

El análisis de Chiswick (1978) se basa en la estimación del siguiente modelo de sección cruzada:

$$ln(w_i) = \boldsymbol{\lambda x}_i + \boldsymbol{\beta} t_i + \boldsymbol{\varepsilon}_i \tag{1.6}$$

siendo w_i el salario del individuo i, $\boldsymbol{x}_i$ es un vector de características socioeconómicas y t_i mide el tiempo que el individuo i lleva residiendo en Estados Unidos.

Borjas (1985, 1995) demuestra, a través de la utilización de cohortes relacionados con los años que los individuos de la muestra llevan en Estados Unidos, que la medida propuesta por Chiswick (1978) no es correcta si se utilizan datos de sección cruzada y por tanto los resultados no son concluyentes. El análisis de Borjas (1985, 1995) se basa en la estimación del siguiente modelo:

$$ln(w_{it}) = \boldsymbol{X}_{it}^{'}\boldsymbol{\theta} + \gamma_1 \boldsymbol{I}_{it} + \gamma_2 \boldsymbol{I}_{it} \times \boldsymbol{Y}_{it} + \gamma_3 \boldsymbol{I}_{it} \times \boldsymbol{Y}_{it}^2 + \delta_1 \boldsymbol{I}_{it}\boldsymbol{C}_{it} + \delta_2 \boldsymbol{I}_{it}\boldsymbol{C}_{it}^2 + \boldsymbol{\varepsilon}_{it} \tag{1.7}$$

siendo $\boldsymbol{I}$ una variable que indica si el individuo i es inmigrante; $\boldsymbol{C}$ es una variable que especifica cada cohorte; $\boldsymbol{Y}$ representa el número de años en Estados Unidos. El vector $\boldsymbol{X}$ incluye diferentes características sociodemográficas; el parámetro γ_1 representa las diferencias salariales entre los nativos y los inmigrantes; los parámetros γ_2 y γ_3 muestran cómo se incrementan esas diferencias cuando el tiempo en el país de acogida, aumenta y

los parámetros δ_1, δ_2 muestran las diferencias salariales en específico para cada cohorte.

Antecol et al. (2006) utilizan el modelo de Borjas (1985, 1995) para comparar el proceso de asimilación en Estados Unidos, Canadá y Australia; obteniendo como resultados principales que el proceso de asimilación en cuanto a salarios, es más rápido en Estados Unidos y más lento en Australia, motivado por los salarios más rígidos en este último país y por ofrecer generosas prestaciones por desempleo. También es en Estados Unidos donde los inmigrantes recién llegados tienen las mayores diferencias salariales en comparación con los nativos y con los inmigrantes que llevan más años en el país.

Hamermesh and Trejo (2013) basan su teoría de la asimilación de los inmigrantes en el uso del tiempo, argumentando que la asimilación es un proceso en el que el inmigrante realiza actividades que los nativos no necesitan realizar con el objetivo de aprender más sobre el nuevo país y de "encajar" mejor en el futuro.

Borjas (2015a) muestra que el proceso de asimilación ha sufrido una importante desaceleración en los emigrantes recientes y concluye que los inmigrantes recientes tienen menos incentivos para invertir en la obtención de las características específicas de los nativos en cuanto a destrezas y habilidades, lo que provoca la disminución de la tasa de asimilación económica de los mismos. Para llegar a esas conclusiones se basa en el siguiente modelo:

$$ln(w_{i\tau}) = \boldsymbol{\phi}_{c\tau} + \boldsymbol{X}'_{i\tau}\boldsymbol{\beta}_{\tau} + \boldsymbol{\varepsilon}_{i\tau} \tag{1.8}$$

siendo $w_{i\tau}$ el salario semanal del individuo i en la sección cruzada τ; $\boldsymbol{X}$ es el vector de variables socioeconómicas incluyendo el número de años que el individuo i lleva residiendo en Estados Unidos, introducida como un polinomio de segundo grado y $\boldsymbol{\phi}$ es un vector de efectos fijos indicando un específico cohorte en la sección cruzada.

Kaushal et al. (2016) analizan el proceso de asimilación a través de datos longitudinales de Canadá y Estados Unidos, utilizando modelos de efectos fijos que permitan controlar las características invariantes en el tiempo. El modelo queda descrito como:

$$ln(w_{it}) = \alpha_{i0} + \boldsymbol{X}_{it}\boldsymbol{\beta} + \alpha_1^*\boldsymbol{T}_t + \alpha_2^*\boldsymbol{IMM}_i^*\boldsymbol{T}_t + \boldsymbol{\varepsilon}_{it} \tag{1.9}$$

donde el vector $\boldsymbol{X}$ denota las características socioeconómicas tales como el nivel de estudios, el estado civil o el número de hijos. Los efectos fijos se incluyen a través de α_{i0} para controlar características que son invariantes en el tiempo, incluyendo aquellas que traen a su llegada al país. La variable IMM es igual a 1 si el individuo es inmigrante y la variable

T es una variable de tendencia.

Para estimar un modelo que tenga en cuenta el tiempo que el individuo lleva en el país de acogida, Kaushal et al. (2016) utilizan la siguiente especificación:

$$ln(w_{it}) = \beta_{i0} + \boldsymbol{X}_{it}\boldsymbol{\Phi} + \beta_1^*\boldsymbol{T}_t + \sum_c \beta_{2c}^*\boldsymbol{YSI}_c^*\boldsymbol{T}_t + \boldsymbol{\varepsilon}_{it} \tag{1.10}$$

reemplazando en esta especificación a la variable IMM por la variable YSI que indica el número de años desde la emigración, tipificada en diferentes categorías: $0-5, 6-10, 11-20$, más de 20. Como resultados en la estimación de estos modelos, Kaushal et al. (2016) encontraron que es muy rápido el proceso de asimilación de los inmigrantes en Estados Unidos motivado por las características del mercado laboral en ese país, manifestando un crecimiento rápido en sus salarios tanto para los individuos con pocos años de estudio, como para aquellos con mayor nivel educativo.

Otro de los principales resultados de este estudio es concluir que el uso de datos de sección cruzada sobreestima la asimilación salarial de los inmigrantes recientes (aquellos que llevan menos de 10 años en Estados Unidos). En cuanto a las mujeres, el proceso de asimilación es más lento y experimentan un decrecimiento en sus salarios, lo que estaría motivado por las limitaciones en las oportunidades de empleo o por el hecho de que las mujeres acepten trabajos más flexibles, con menos salarios que les permita cuidar de sus hijos o nietos.

En resumen, existen controversias en cuanto al grado de asimilación de los inmigrantes desde el punto de vista económico en Estados Unidos, pero el punto de coincidencia de todos los trabajos al respecto, es comprobar que la asimilación depende de la similitud de características socioeconómicas y habilidades entre los nativos y los inmigrantes y que el proceso de incorporación de los inmigrantes a la actividad económica en el país de acogida será mayor cuanto más cerca estén sus habilidades, de las habilidades de los nativos (Elliott and Lindley, 2008; Brunner and Pate, 2016).

En cuanto a los cubanos, algunos trabajos que han utilizado la regresión lineal clásica para datos de sección cruzada, han encontrado que el proceso de asimilación de los cubanos es rápido. En este sentido, Chiswick (1978) encontró que los salarios de los inmigrantes cubanos que arribaron entre 1965-1969 se incrementaba en un 37 % en los primeros 10 años después de la emigración. Sin embargo, Borjas (1985) haciendo el mismo análisis pero utilizando cohortes para los años de la emigración, encontró que esta asimilación no era cierta y que los salarios de los cubanos lejos de aumentar, descendían en un 25 %. Un análisis posterior en Borjas (2015a) para los cubanos que llegaron a Estados Unidos entre

1995 y 1999 da cuentas de que el decrecimiento de los salarios durante los primeros 10 años es de un 4 %. No existen trabajos dedicados a este tema para años posteriores.

CAPÍTULO 2

La Emigración cubana a Estados Unidos

El 17 de diciembre de 2014 marca un antes y un después para los cubanos residentes en Cuba y fuera de ella, cuando los presidentes de Cuba y de Estados Unidos anunciaron, simultáneamente desde La Habana y Washington, el restablecimiento de relaciones entre ambos países.

> *Igualmente, hemos acordado el restablecimiento de las relaciones diplomáticas. Esto no quiere decir que lo principal se haya resuelto. El bloqueo económico, comercial y financiero que provoca enormes daños humanos y económicos a nuestro país debe cesar. Aunque las medidas del bloqueo han sido convertidas en Ley, el Presidente de los Estados Unidos puede modificar su aplicación en uso de sus facultades ejecutivas. Proponemos al Gobierno de los Estados Unidos adoptar medidas mutuas para mejorar el clima bilateral y avanzar hacia la normalización de los vínculos entre nuestros países, basados en los principios del Derecho Internacional y la Carta de las Naciones Unidas. Cuba reitera su disposición a sostener cooperación en los organismos multilaterales, como la Organización de Naciones Unidas. Al reconocer que tenemos profundas diferencias, fundamentalmente en materia de soberanía nacional, democracia, derechos humanos y política exterior, reafirmo nuestra voluntad de dialogar sobre todos esos temas....* ***Raúl Castro, 17 de diciembre de 2014.***

Si hay algo que ha distinguido la relación entre ambos países, desde que el 3 de enero de 1961 se cerraran las embajadas en Washington y La Habana, ha sido la emigración de los cubanos a Estados Unidos.

> *La historia entre Estados Unidos y Cuba es complicada. Yo nací en 1961, justo dos años después de que Fidel Castro tomara el poder en Cuba y unos meses después de la invasión en la Bahía de Cochinos, en la que se intentó derrocar al régimen. En las siguientes décadas, la relación entre nuestros países tuvo lugar frente al transfondo de la Guerra Fría y la firme oposición de Estados Unidos al comunismo.... Entretanto, la comunidad de exiliados cubanos en los Estados Unidos contribuyó enormemente con nuestro país en la política, los negocios, la cultura y los deportes...., los cubanos ayudaron a reconstruir a los Estados Unidos, a pesar de sentir una dolorosa nostalgia por la tierra y las familias que dejaron atrás....* ***Barack Obama, 17 de diciembre de 2014.***

La historia de la emigración cubana a Estados Unidos se remonta al siglo XIX debido a la cercanía entre ambos países y a la influencia política y económica de Estados Unidos sobre Cuba. En aquella época era una emigración formada principalmente por intelectuales y hombres de negocio quienes emigraban debido a las diferencias de orden político con España y a la situación económica imperante en la isla.

Entre 1871 y 1900 emigraron a Estados Unidos alrededor de 60.000 cubanos y se estima que la cifra de cubanos llegados a ese país entre 1871 y 1958 fue de 221.505 personas. Entre 1950 y 1958 unos 65.000 cubanos fueron admitidos como inmigrantes permanentes. En esos años, el flujo de cubanos se debió esencialmente a la búsqueda de empleo en ciudades como Nueva York, Tampa y Miami.

Estados Unidos comenzó a ser el lugar preferido por la mediana y alta burguesía cubana para enviar a sus hijos a estudiar, para vacacionar e invertir en pequeños negocios; coincidiendo con el hecho de que en esos años la emigración hacia ese país se producía fundamentalmente de países con un alto nivel de desarrollo. En el año 1958 la población cubana asentada en Estados Unidos ascendía a 125.000 personas (Sainz-Cano et al., 2015).

La entrada de inmigrantes en Estados Unidos se incrementó de forma considerable a partir de 1965 cuando se realizó la Enmienda a la Ley de Inmigración pasando del sistema de cuotas al sistema que se basa en la reunificación familiar; sistema que ha recibido críticas por parte de autores como Clark et al. (2002) y Borjas (2015a).

Los cubanos que entran en Estados Unidos gozan de un estatus que no posee ningún otro grupo de inmigrantes, aún si la persona es un refugiado político; diferenciándose fundamentalmente de la emigración desde otros países de América Latina, que ha tenido que enfrentarse a fuertes restricciones implementadas por parte de Estados Unidos para

entrar al país. Por otro lado los cubanos son en su mayoría personas de alta cualificación dentro de los inmigrantes, sobre todo entre aquellos de origen latino.

Amparados bajo la Ley de Ajuste Cubano (CAA por sus siglas en inglés) firmada por el presidente Johson, el 2 de diciembre de 1966 y por la política denominada "pies secos, pies mojados" firmada como complemento a la Ley de 1966 en el marco de las negociaciones con el gobierno cubano en 1995, cualquier cubano que llegue a territorio norteamericano aún si es por vía ilegal y reside al menos un año y un día, recibe el estatus de residente permanente.

Los cubanos son el único grupo de inmigrantes que recibe de manera inmediata a su entrada a los Estados Unidos el permiso de trabajo sin necesidad de presentar una declaración jurada de manutención, número de seguridad social y beneficios públicos para alimentos y alojamiento. Además no tienen que regresar a su país de origen a esperar la aprobación del permiso de residencia y tampoco requieren de abogados ni se les cobra por los gastos que conlleva este proceso.

La Ley de Ajuste Cubano provee un procedimiento especial bajo el cual las personas nacidas en Cuba o ciudadanos cubanos y sus cónyuges e hijos que les acompañan, pueden solicitar la Residencia Permanente. La Ley de Ajuste Cubano le permite al Fiscal General de los Estados Unidos la discreción de conceder residencia permanente a los nacidos en Cuba o ciudadanos cubanos si cumplen los requisitos siguientes:

- Han estado físicamente presentes en los Estados Unidos durante al menos 1 año.
- Han sido admitidos o se les ha concedido permiso de entrada.
- Son admisibles como inmigrantes.

Las solicitudes para la Tarjeta Verde (Residencia Permanente) pueden ser aprobadas aún si no cumplen con los requisitos ordinarios decretados bajo la Sección 245 de la Ley de Inmigración y Nacionalidad (INA, por sus siglas en inglés) y no es necesario que los solicitantes sean beneficiarios de una petición de visa de inmigrante.

Además, los nacidos en Cuba o ciudadanos cubanos que lleguen a otro lugar que no es un puerto de entrada a Estados Unidos aún pueden ser elegibles para una Tarjeta Verde si la *U.S Citizenship and Immigration Services* (USCIS) les ha concedido la opción de ser admitidos en los Estados Unidos.

En los procesos migratorios cubanos han estado presentes factores económicos, políticos y sociales con particular énfasis en los momentos de crisis económicas profundas.

Aproximadamente la mitad de los cubanos nacidos en Cuba que han emigrado a Estados Unidos lo han hecho antes de 1980, cada uno con sus diferentes experiencias y por razones fundamentalmente políticas.

Los cubanos que emigraron después de 1980 representan otra Cuba. La mayoría nació bajo la Revolución y han salido de Cuba buscando mejores condiciones de vida, con la esperanza de poder ayudar a las familias que dejaron atrás; por tanto son personas con motivaciones fundamentalmente económicas.

Aunque es de señalar que en el caso cubano, es difícil distinguir entre una motivación y otra para emigrar porque en una economía centralizada como la cubana todo lo que se realiza en el ámbito económico es establecido desde los órganos de poder político.

Las diferencias entre estos grupos de cubanos se manifiesta en aspectos como el nivel de inglés, la marcada disminución de la solicitud de la nacionalidad norteamericana, la intención de voto, el incremento de las remesas enviadas a Cuba y de los viajes que realizan a la isla y otras características socioeconómicas que son objeto de estudio en esta tesis.

Según López (2015) el 57 % de los cubanos residentes en Estados Unidos en el año 2013 ha nacido en Cuba, 6 de cada 10 posee la nacionalidad norteamericana; el 13 % domina el inglés, una tasa baja si se compara con el resto de hispanos (25 %); aún así el 36 % de los cubanos se declara bilingüe. Existen familias cubanas que, aunque lleven más de 20 años residiendo en Estados Unidos, no hablan ni escriben en inglés mientras sus hijos nacidos en territorio norteamericano, no hablan español.

La edad mediana es de 40 años, cuando la edad mediana de los norteamericanos es de 37 años y la de los hispanos de 28 años. El 50 % de los cubanos emigrados está casado frente al 36 % de los nacidos en Estados Unidos. El 25 % de los cubanos de 25 o más años ha obtenido el *Bachelor Degree* comparado con el 14 % de los hispanos y el 30 % de la población en general.

La tasa bruta de natalidad de los cubanos en Estados Unidos ha sido siempre baja, incluso inferior a la tasa de los nativos. Mientras los mexicanos muestran una tasa de natalidad de 16 nacimientos por cada mil mujeres, en los cubanos es de sólo 9 nacimientos por cada mil mujeres (Martin et al., 2015) y esta tasa se ha mantenido por debajo de los 10 nacimientos por cada mil mujeres desde 1989.

Los cubanos están concentrados en la Florida donde vive el 68 % de ellos, siendo el grupo de hispanos más concentrado en cuanto a lugar de residencia. Un 20 % de los cubanos en Estados Unidos vive en la pobreza comparado con el 16 % de la población norteamericana en su conjunto y el 25 % de los hispanos.

En el año 2012 la mediana de ingresos para los cubanos fue de 25.000 dólares anuales, para los hispanos 21.900 y para la población en general 30.000. Un 25 % de cubanos no posee un seguro médico, el 55 % de ellos es propietario de su vivienda frente a un 45 % de los hispanos y el 64 % de la población.

Los cubanos residentes en Estados Unidos se han convertido en uno de los grupos políticos y económicos más influyentes, destacando los congresitas Bob Menéndez, Mario y Lincoln Díaz-Balart, Ileana Ros Lehtinen y los senadores Marco Rubio, Ted Cruz y Mel Martínez (Portes and Puhrmann, 2015).

Además, han sido históricamente considerados como el grupo de inmigrantes hispanos más exitoso de Estados Unidos por tener un ingreso medio por encima del resto de inmigrantes de este grupo y por tener tasas de crecimiento salarial más altas. (Borjas, 1982; Portes and Grosfoguel, 1994).

Cuba ha ocupado desde 1970 uno de los 10 primeros lugares de los grupos emisores de emigración hacia Estados Unidos, siendo en la actualidad el séptimo país. Las cuatro principales comunidades de origen hispano son los mexicanos, seguidos de los puertorriqueños, los salvadoreños y los cubanos. Así, en el año 2015 habían 11.906.325 de inmigrantes nacidos en México, 1.686.258 nacidos en Puerto Rico, 1.382.737 nacidos en El Salvador y 1.225.742 nacidos en Cuba (US Custom and Border Protection, 2016).

A partir de 1959, se pueden diferenciar varias etapas en el proceso de emigración de los cubanos hacia Estados Unidos: 1959-1962, 1965-1973, 1980, 1994, 1995-actualidad. Según datos de la US Custom and Border Protection (2016) actualmente viven en Estados Unidos 2.115.879 personas de origen cubano, inmigración que se ha incrementado de forma considerable a partir del restableciento de relaciones entre ambos países ante el temor a que se derogue la denominada Ley de Ajuste Cubano con el cambio de administración.

2.1. La primera oleada: 1959-1962

Con el triunfo de la Revolución, Estados Unidos comienza a asignarle el estatus de refugiado político a los cubanos que empezaron a llegar a su territorio. En diciembre de 1960 se crea el Centro de Emergencia de Refugiados de Miami (*Cuban Refugge Emergency Center*), financiado por el Fondo de Contingencias del Presidente de Estados Unidos y por fondos privados.

El 3 de febrero de 1961 surge el Programa de Refugiados Cubanos (*Cuban Refugee Program*)

que otorgaba pensiones, créditos, asistencia médica, homologación de estudios, acceso a puestos de trabajos y otras ventajas de las que no disfrutaban inmigrantes venidos de otros países. El 28 de junio de 1962 el presidente Kennedy firma la Ley pública 87-510 conocida como Ley de Asistencia a la Migración y a los Refugiados del Hemisferio Occidental dirigida esencialmente a beneficiar a los cubanos que deseaban emigrar a Estados Unidos. El presidente Kennedy le indica al secretario de Salud, Educación y Bienestar, Ribicoff, las siguientes medidas:

- Proveer toda la asistencia necesaria a las agencias que atienden las necesidades de los refugiados cubanos.
- Obtener la asistencia tanto de agencias privadas como gubernamentales para proveer oportunidades de empleo a los refugiados cubanos.
- Proporcionar fondos suplementarios a los refugiados, incluyendo transporte y costes de adaptación a las nuevas comunidades y el eventual retorno a sus hogares de origen.
- Habilitar asistencia financiera ya sea por canales federales, estatales y locales para cubrir las necesidades básicas de los refugiados cubanos.
- Proveer asistencia sanitaria a través de los programas de asistencia financiera complementada con los servicios de asistencia a menores, los servicios públicos de salud y otras disposiciones que sean necesarias.
- Proveer a las escuelas de los fondos necesarios para afrontar el impacto imprevisto de la llegada de niños refugiados.
- Iniciar las medidas necesarias para incrementar las oportunidades educacionales de los refugiados incluyendo a los médicos, maestros y demás profesionales.
- Proveer la ayuda financiera que sea necesaria para atender las necesidades de los niños que han viajado solos, siendo el grupo de refugiados más indefenso dentro de la población de emigrados.
- Iniciar un programa de distribución de alimentos entre los refugiados a través de agencias públicas y de voluntarios.

Según el Presidente Kennedy, estas medidas eran una expresión de la voluntad y el deseo de los norteamericanos de brindar ayuda a los refugiados cubanos, hasta el momento en que existiesen mejores circunstancias para regresar a Cuba en un clima de confianza y de garantías.

Figura 2.1: *Operación Peter Pan. Diciembre 1960 - Octubre 1962.*

Entre los últimos días de 1958 y octubre de 1962, 248.870 cubanos llegaron a Estados Unidos. Los primeros inmigrantes después del triunfo de la Revolución estaban formados por la élite del régimen de Fulgencio Batista. A partir de Abril de 1961, después de la Invasión de Bahía de Cochinos (Playa Girón para los cubanos) el éxodo de nacionales cubanos se incrementó drásticamente hasta que fueron suspendidos los vuelos regulares entre La Habana y Miami en Octubre de 1962.

El 26 de diciembre de 1960 llegaron a Miami los primeros niños que viajaron bajo el auspicio de la denominada **Operación Peter Pan**, un programa creado por el *Catholic Welfare Bureau* (*Catholic Charities*) de Miami con el fin de que los padres en Cuba enviaran a sus hijos a Estados Unidos, huyendo del carácter socialista de la Revolución.

Este éxodo, ideado por el sacerdote de origen irlandés Bryan O. Walsh y financiado por el gobierno de Estados Unidos, se extendió a lo largo de 22 meses hasta el 22 de octubre de 1962 coincidiendo su fin con la Crisis de Octubre, aunque en realidad se extendió hasta bien entrada la década de los 70, con nuevas variantes, como la de los vuelos de menores de edad a Madrid (Torreira and Buajasán, 2000). Se estima que con este programa salieron de Cuba alrededor de 15.000 niños.

Entre 1961 y 1962 los exiliados eran principalmente de clase media y un 90 % de ellos de raza blanca (Zavodny, 2003). La mayoría de los inmigrantes de esta etapa procedían de La Habana (un 62 %), un 23,5 % de ellos había terminado el bachillerato y un 12,5 % tenía estudios superiores. El 31 % estaba representado por profesionales, directivos y personas

de clase media, otro 31 % correspondía a trabajadores del sector de servicios y a clérigos.

La media de ingresos de estos inmigrantes era de 5.960 pesos cubanos cuando la renta percápita de Cuba en 1957 era de 431 pesos cubanos (Fagen and Brody, 1964). Eran en su mayoría personas provenientes de las altas clases sociales de Cuba, con ingresos comparables a los que se obtenían en Estados Unidos por individuos con la misma cualificación.

Entre octubre de 1962 y septiembre de 1965 llegaron a Estados Unidos unos 69.081 cubanos principalmente a través de terceros países como México y España (58.545 personas), siendo fundamentalmente familiares de los que ya se habían establecido en Estados Unidos y padres de los niños que habían salido de Cuba bajo el auspicio de la operación Peter Pan. Se calcula que de manera irregular, llegaron 10.536 personas durante este período. Los cubanos emigrados entre 1959 y 1964 son conocidos en Estados Unidos como "el exilio histórico", auto definiéndose como emigrantes por razones políticas.

2.2. La segunda oleada: 1965-1973

Entre 1965 y 1973 arribaron a Estados Unidos 300.000 cubanos; coincidiendo con el primer gran éxodo cubano ocurrido en 1965, cuando el gobierno cubano permitió la salida de Cuba a través del puerto marítimo de Camarioca, al norte de la provincia de Matanzas, seguido por el acuerdo establecido entre ambos países de permitir vuelos desde Miami denominados los "Vuelos de la Libertad".

> *....Camarioca fue la primera (crisis migratoria), en octubre de 1965, por lo que le digo que cortaron los viajes, no dejaban salir. Entonces se inician las salidas ilegales, los problemas y la propaganda. Los que estaban allí - ya se habían ido unos cuantos, le dije -, tenían recursos, porque los primeros que se fueron eran los jefes con dinero; los de menos recursos no conocían el camino; se fueron, ya le mencioné, profesionales, médicos, obreros calificados, maestros, etc. Y nosotros aguantando aquí, enfrentando la carencia de ese personal calificado. Pero los norteamericanos cortan la posibilidad de viajar cuando la crisis de octubre, y se empiezan a producir estos problemas de la separación familiar, y las salidas ilegales con el peligro de los accidentes....Entonces dijimos: "No, no hace falta que corran riesgos, vengan a buscarlos" y habilitamos un pequeño puerto, Camarioca, cerca de Varadero....* ***Castro and Ramonet (2006)***.

Del 28 de septiembre al 3 de noviembre de 1965 ocurrió este puente marítimo derivado de presiones internas y externas que llevaron a 4.993 personas a abandonar Cuba en embarcaciones compradas o alquiladas por sus familiares en Miami, llegando a ser de 200.000 la cantidad de peticiones realizadas al gobierno cubano para salir por esta vía.

Esta etapa quedará marcada en la historia de la emigración cubana hacia Estados Unidos por la firma el 2 de noviembre de 1966 por el Presidente Johnson de la Ley Pública 89-732 denominada "Ley para Ajustar el Status de los Refugiados Cubanos a la de Residentes Permanentes Legales de Estados Unidos" y que popularmente se conoce como **"Ley de Ajuste Cubano"**.

Bajo esta ley el ciudadano cubano, que al llegar a Estados Unidos fuese inspeccionado y admitido o puesto bajo palabra (*parole*) y que hubiese permanecido en el país durante un año, tendría derecho a obtener la residencia permanente. Utilizando esta ley, a todos los cubanos que llegaron a partir de 1959 y a los que llegarían a partir de la fecha de firma de la misma, se les otorgaría la residencia permanente y en menos de 3 años podrían optar a la nacionalidad norteamericana.

2.3. El éxodo del Mariel en 1980

De mayo a septiembre de 1980 cerca de 125.000 cubanos llegaron a Miami, como consecuencia de un conjunto de sucesos que condujeron a que el 20 de Abril de 1980, Fidel Castro planteara que todos los que quisieran emigrar a Estados Unidos podían hacerlo a través del puerto de Mariel. De esos cubanos, el 60 % eran hombres, un 22 % eran mujeres y el resto niños.

En 1980, el 35.5 % de la población de Miami en su áera metropolitana, eran personas nacidas en el extranjero, de las cuales el 56 % habían nacido en Cuba.

> *Ustedes conocen los hechos, y si no se tratara de la presencia de periodistas extranjeros no haría falta hablar mucho de los antecedentes. Pero la cuestión se desencadena a partir de las provocaciones en las embajadas de Perú y Venezuela. Todo el mundo sabe que el imperialismo quería afectar las relaciones entre Cuba y Venezuela y Cuba y Perú, desde hace mucho tiempo viene con esa idea maquinando cosas.... Sencillamente le retiramos la custodia a la embajada y nosotros sabíamos lo que iba a pasar, porque no se puede estar estimulando durante tanto tiempo al lumpen por parte del imperialismo y por parte de los lacayos del imperialismo ofreciéndole villas y castillas, ofreciéndole el paraíso,*

> *ofreciéndole todo, llenándolos de ilusiones; mientras, por otro lado, les cierran la entrada a sus países. Cosa curiosa: los alientan a penetrar ilegalmente por la fuerza, los alientan a salir ilegalmente; pero no les dan entrada si lo solicitan normal y pacíficamente.* ***Discurso de Fidel Castro, 1 de Mayo de 1980***.

El éxodo del Mariel representó el primero diferenciado en las características raciales y sociales de los éxodos precedentes, al ser un exilio de las clases más bajas de la sociedad cubana, donde el 20 % no era de raza blanca lo que de provocó profundos cambios en la composición del exilio cubano en Miami caracterizado hasta la fecha por su estereotipo de exilio de éxito (Skop, 2001).

Según Card (1990) la brecha salarial entre los cubanos llegados desde el Mariel y el resto de cubanos asentados en Miami en esos años era del 34 % motivado fundamentalmente por el bajo nivel educacional de los emigrados. Las características de estos inmigrantes ha sido objeto de controversia desde que se ha dicho que muchos fueron sacados de los hospitales psiquiátricos y las cárceles.

Antes de producirse el éxodo de Mariel, el gobierno del Presidente James Carter aprobó el 17 de marzo de 1980 la denominada **Ley para Refugiados**, con la cual por primera vez los cubanos que arribaban a tierras norteamericanas, tenían que someterse al mismo procedimiento que el resto de inmigrantes en aras de conseguir la residencia. A estos inmigrantes no se les dio la categoría de exiliados o refugiados.

En julio de 1984 delegaciones de ambos países se reunieron en Nueva York, para celebrar conversaciones conjuntas en torno al tema migratorio; llegándose a la firma de un acuerdo de Normalización de Relaciones Migratorias entre ambos países. La delegación de Estados Unidos planteó la deportación de 5.000 cubanos con antecedentes criminales y la posibilidad de que un grupo reducido de cubanos que había salido de Cuba por el puerto de Mariel regresara a Cuba de forma voluntaria según su deseo.

No es hasta diciembre de ese año en que se llega a un acuerdo, donde el gobierno cubano acepta recibir a 2.746 cubanos de los 5.000 previamente sugeridos y el gobierno norteamericano se compromete a otorgar 20.000 visas anuales, además de aceptar una cuota adicional de entre 6 mil y 7 mil visas para ex-presos políticos y sus familiares.

Estos acuerdos fueron interrumpidos por la parte cubana entre 1985 y 1987 debido a la denuncia por la parte cubana de las trasmisiones radiales de la emisora Radio Martí y reanudadas en 1987 con la firma de un nuevo acuerdo migratorio en el que quedaría incluido el tema de las trasmisiones radiales hacia Cuba.

Figura 2.3: *Balseros cubanos en 1994.*

2.4. El éxodo de 1994

La emigración en los años 90 tuvo un marcado carácter "económico" en combinación con la reunificación familiar y la pérdida de confianza en el proyecto social de la Revolución cubana para salir de la crisis (Aja Díaz, 2010).

En 1993 los Servicios de Guardacostas de Estados Unidos interceptaron a 2.882 cubanos en el mar y en 1994 a 38.560, llegando a ser el número máximo de cubanos interceptados en un día de 3.253 (US Custom and Border Protection, 2016). Los cubanos se iban de Cuba en embarcaciones improvisadas, construidas de forma artesanal en las mismas costas cubanas con materiales endebles, básicamente con neumáticos de coches.

De ahí que a este nuevo éxodo se le conozca como **La Crisis de los Balseros**. Se asemeja al de 1980 en que ambos surgieron producto de situaciones de explosión ciudadana ante los problemas económicos, esta vez agravados por la desparición del sistema socialista en la Europa del Este.

- *"Hoy 23 de agosto de 1994 me decido a irme para los Estados Unidos o para cualquier país, menos quedarme aquí porque esto no hay quien lo resista...."*
- *"He intentado irme con esta cuatro veces; pero mira, no aguanto, prefiero hundirme en el mar"*

Después del pronunciamiento de Fidel Castro, el 5 de agosto de 1994, en que declaraba que no custodiaría más las costas de Estados Unidos, el Presidente de Estados Unidos, Bill Clinton, anunció la existencia de un plan de contingencia para evitar otro éxodo como el de Mariel en 1980.

El proyecto *Distant Shore* contemplaba un bloqueo naval, la detención y procesamiento fuera de la Florida de cualquier refugiado que tratara de entrar a Estados Unidos a través del Estrecho de la Florida y su reubicación en otros estados de la Unión.

El 18 de agosto de 1994 Clinton se reunió con sus principales asesores y decidieron que era necesario poner un alto a la política vigente por 28 años de dar la bienvenida incondicional a todos los refugiados cubanos y anunció que todos los cubanos rescatados en el mar serían trasladados a la base naval de Guantánamo. Al mismo tiempo advertía que Estados Unidos detendría, investigaría y si fuera necesario procesaría judicialmente a los estadounidenses que se hicieran a la mar con el propósito de recoger a los cubanos.

A pesar de los anuncios de no dejar entrar a los cubanos en territorio estadounidense, las personas que decideron emigrar de Cuba por esa vía, no desistieron.

- *"Aunque me envíen a Guantánamo yo me voy a sentir mucho mejor"*
- *"Es preferible estar en la Base Naval de Guantánamo que estar aquí en Cuba"*

El 28 de agosto, sin un fin de la crisis a la vista, Estados Unidos manifestó su disposición a negociar. El gobierno cubano trató de imponer como condición para una solución a esta nueva crisis migratoria, el levantamiento del embargo y el cierre de Radio y TV Martí, concluyendo con el anuncio de un nuevo acuerdo migratorio y la disposición por ambas partes de sentarse nuevamente a conversar.

El 9 de septiembre se suscribió un comunicado conjunto con el objetivo de normalizar y regular la migración de una forma ordenada, legal y segura y se ratificaba por parte de Estados Unidos el otorgamiento de un mínimo de 20.000 visas anuales, el establecimiento del sorteo especial de visas también para los cubanos y el compromiso de llevar a los emigrantes ilegales interceptados en el mar a instalaciones fuera de Estados Unidos. Por su parte La Habana se comprometía a evitar nuevas salidas de balseros.

2.5. La emigración más reciente

En mayo de 1995 los gobiernos de Cuba y Estados Unidos vuelven a reunirse por temas migratorios y se firma la denominada Declaración Conjunta como acuerdo complementario al acuerdo de septiembre de 1994.

En ese acuerdo se admite la admisión paulatina de un máximo de 5 mil cubanos dentro de las 20 mil visas anuales, de los cubanos que estaban en la Base Naval de Guantánamo y su traslado a territorio de Estados Unidos. Además, se pacta la devolución a territorio cubano de los emigrantes interceptados en el mar a partir de la firma del acuerdo, con el compromiso por parte del gobierno cubano, de que no serán represaliados cuando retornen a sus domicilios.

Así mismo, se acepta por parte de Cuba la deportación de aquellos ciudadanos provenientes de la base naval de Guantánamo que se consideren inadmisibles para entrar a Estados Unidos y se reafirma por ambas partes el compromiso de aunar esfuerzos para disuadir e impedir las salidas ilegales de Cuba.

Fruto de estas conversaciones surge lo que se denomina **"política de pies secos, pies mojados"**, como complemento a la Ley de Ajuste Cubano y lo que hace esta política es garantizar a todos los cubanos que logren pisar suelo norteamericano (pies secos) sin una visa de inmigrante válida, obtener la residencia después de un año.

Dentro de esta categoría se incluyen a los que arriban por mar y no son interceptados por el Servicio de Guardacostas, a los que llegan por las fronteras de México y Canadá y a los que llegan a un aeropuerto de Estados Unidos procedentes de cualquier país con un visado de turista. En cambio, las personas interceptadas en el mar (pies mojados) serían repatriadas a Cuba.

En cada una de las conversaciones, se reafirma el compromiso de otorgar 20 mil visas anuales debido a que entre 1985 y 1994 de las 100 mil visas que debieron otorgarse, sólo fueron tramitadas 11.222. A partir de la firma de los acuerdos de 1994 y 1995 se realizan conversaciones bianuales entre ambos gobiernos, alternando entre La Habana y Nueva York como sedes de las conversaciones. A partir de 1995 se ha ido cumpliendo con el acuedo de otorgar las 20 mil visas anuales.

Aunque los cubanos pueden solicitar una visa de inmigrante, la cual es concedida a un ciudadano extranjero que pretenda residir de manera permanente en Estados Unidos y tiene que ser respaldado por un familiar que posea ciudadanía estadounidense o por un

empleador potencial en los Estados Unidos; los cubanos se acogen al Programa de Parole Cubano.

El *parole* es un tipo especial de admisión a los Estados Unidos, se ejecuta a discreción de los Servicios de Inmigración y Ciudadanía de los Estados Unidos (USCIS) y a diferencia de las visas de inmigrante, los beneficiarios de *parole* no entran a los Estados Unidos con estatus de Residente Legal Permanente (LPR). Sin embargo, la Ley de Ajuste Cubano (CAA) permite que los cubanos beneficiarios de *parole* soliciten ajustar su estatus a Residente Legal Permanente a partir del año de entrar a los Estados Unidos.

Como parte del Programa de Parole Cubano, en el año 2007 el Presidente George Busch crea el Programa de Parole Cubano de Reunificación Familiar (CFRP) el cual permite a los ciudadanos y residentes legales permanentes de los Estados Unidos solicitar *parole* para sus familiares en Cuba.

Otra parte del Programa de Parole Cubano es el Parole Familiar Cubano (CP3) el cual permite que los ciudadanos cubanos mayores de 21 años a los que se les ha emitido visa o se les ha otorgado *parole* soliciten *parole* familiar para algunos miembros de su familia.

Este procedimiento de reagrupación familiar tiene sus costes. El familiar en Estados Unidos debe realizar la petición de visa de inmigrante para su familiar residente en Cuba (modelo I-130, a un costo de 420 dólares) y a partir de febrero de 2015 también se debe rellenar la solicitud de documento de viaje (modelo I-131) ante el Departamento de Ciudadanía e Inmigración (USCIS) con un coste de 360 dólares.

Los solicitantes pueden pedir una exención de pago mediante el modelo I-912, que favorece a reclamantes de bajos recursos. Todas las solicitudes de parole son revisadas por los Servicios de Ciudadanía e Inmigración (USCIS) del Departamento de Seguridad Interna de los Estados Unidos (DHS) y la decisión final de los casos de *parole* depende únicamente del DHS.

Además de las visas de inmigrante, existen varias categorías de visas de no inmigrante entre las que se incluyen visas de turismo/visita, de estudiante y de negocios.

Tipos de visas de no inmigrantes actuales:

- Turismo, vacaciones, visitas familiares: B-2
- Negocios: B-1
- Diplomáticos y funcionarios oficiales: A

- Tratamientos Médicos: B-2
- Profesores, maestros, académicos: J
- Estudiantes: F, M
- Atletas o grupos de espectáculos: P
- Personas con trabajo religioso: R
- Personalidades destacadas: O

El número de cubanos que ha entrado a Estados Unidos por la frontera con México y por el mar ha sufrido un dramático ascenso ante el anuncio de la normalización de relaciones entre Cuba y Estados Unidos en diciembre de 2014.

Entre enero y marzo de 2015, 9.900 cubanos llegaron a Estados Unidos por algún puerto de entrada, más del doble de los 4.746 que lo hicieron en el mismo período del año anterior llegando a 17.057 personas entre octubre y diciembre de 2015. Estos cubanos llegan por vía marítima en balsas o a través de mafias que se dedican al tráfico de personas. Por tierra, los cubanos hacen un largo recorrido, que comienza en países como Ecuador, Guyana, Trinidad Tobago, Colombia o Panamá y recorren Centroamérica hasta llegar a la frontera de Estados Unidos con México.

De esta manera, la mayoría de cubanos que llegan por tierra a Estados Unidos lo han hecho a través de la frontera en el sector de Laredo, México. En 2015, 28.371 cubanos entraron por esta vía, lo que representa un incremento del 82 % con respecto al mismo período del año fiscal anterior.

En el año 2015 entraron por algún aeropuerto de Florida 4.709 cubanos, provenientes principalmente de países europeos con el objetivo de acogerse a la Ley de Ajuste Cubano, cifra que alcanzó los 10.992 en 2016. En el año fiscal 2016, el Servicio de Guardacostas de Estados Unidos aprehendió en el mar a 5.263 cubanos, los cuales fueron deportados a Cuba. Esta cifra supera la cifra de 2015 que alcanzó las 3.505 personas.

En resumen, si en el año 2014 llegaron a Estados Unidos por todas las vías posibles, 24.278 cubanos con el objetivo de residir de forma permanente; en 2015 la cifra alcanzó las 43.159 personas y en 2016 las 46.635 (US Custom and Border Protection, 2016).

El 12 de enero de 2017 nuevamente el gobierno de Cuba y de Estados Unidos alcanzan un acuerdo para dar paso a la normalización de las relaciones migratorias en el interés de garantizar una migración regular, segura y ordenada. Barack Obama anuncia el fin

de la **"política de pies secos, pies mojados"** así como del Programa de admisión de profesionales médicos cubanos en terceros países.

A partir de este momento todos los ciudadanos cubanos que lleguen a pisar suelo norteamericano de manera ilegal ya sea por mar o por tierra serán deportados a Cuba, lo que significa que se les aplicará el mismo procedimiento que se aplica a ciudadanos del resto del mundo, quedándose miles de cubanos varados en los diferentes países que forman parte de la ruta utilizada en años recientes para llegar a Estados Unidos.

Comienza entonces una nueva etapa en la historia de la emigración cubana a Estados Unidos, pues se ha puesto a los cubanos en el mismo lugar que el resto de personas que quieren emigrar a Estados Unidos.

Parte II

Algunos modelos de regresión con respecto a la media

CAPÍTULO 3

De los modelos lineales clásicos a los modelos lineales generalizados

Este capítulo comienza una breve reseña a las diferentes formas en las que se han especificado los modelos de regresión usando la media condicional de la variable en estudio. Referencia que se extiende al Capítulo 7, para concatenar con los modelos cuantílicos aditivos estructurados.

Usar la media como medida de tendencia central ha sido una aproximación que ha permanecido de manera dominante en la investigación y el análisis. Estos modelos tienen la ventaja de que los cálculos son más asequibles así como la interpretación de los resultados pero tienen la desventaja importante de que no siempre los resultados obtenidos pueden ser generalizados a otras partes de la distribución de probabilidades de la variable en estudio que no sean las centrales y donde podría estar enfocado el interés del estudio en cuestión.

También nos podemos encontrar con situaciones en las que existen muchos valores outliers o colas con gran peso dentro de la distribución de la variable de interés y ante esto, la media responde de manera sensible. Existen además situaciones en que los supuestos básicos se incumplen y específicamente que los errores no son normales lo que nos lleva a obtener conclusiones falsas referentes a la significatividad, y por ende, a la importancia de las variables.

Podemos encontrar numerosa literatura relacionada con estos modelos dentro de los que se destacan los libros de Seber and Lee (2012) y Draper and Smith (2014) para los modelos lineales, Fahrmeir et al. (2013) donde se estudian los modelos lineales, los modelos lineales generalizados y los modelos aditivos estructurados, Faraway (2014) y

Faraway (2016) para especificar y estimar modelos lineales y generalizados utilizando el lenguaje de programación R y los clásicos de Greene (2003) y Wooldridge (2015). Para los modelos lineales generalizados (GLM) además de los clásicos Nelder and Wedderburn (1972) y McCullagh and Nelder (1989), es muy útil el libro de Kleiber and Zeileis (2008) con diferentes aplicaciones en R y en los GLM de locación, escala y forma (GLMSS) insustituible el libro de Rigby and Stasinopoulos (2005).

3.1. El modelo de regresión lineal clásico

El término de regresión fue utilizado por primera vez por Francis Galton (1886) quien relacionó la estatura de los hijos con la de sus progenitores y observó que ambas variables tenían una fuerte relación lineal introduciendo así el concepto de **correlación**. A medida que aumentaba el valor de la estatura de los padres, aumentaba el valor de la de los hijos.

Sin embargo, a padres de estatura muy elevada correspondían hijos de estatura elevada pero no tanto como la de sus padres y a padres muy bajos en estatura, correspondían hijos no tan bajos. De ahí el término **regresión**, al exponer que la estatura de los hijos "regresaba" a la estatura media de la población.

Supongamos que estamos interesados en estudiar la relación entre una variable de interés $\boldsymbol{y}$ y un conjunto de variables que expliquen esa variable de interés $x_1, \ldots, x_n$, siendo esta relación no exacta lo que nos conduce a tener en cuenta una perturbación $\boldsymbol{\varepsilon}$. El modelo clásico de regresión lineal se sustenta en la siguiente relación

$$y_i = \beta_0 + \beta_1 x_{1i} + \ldots + \beta_p x_{pi} + \varepsilon_i \qquad i = 1, \ldots, n \tag{3.1}$$

donde y_i son los valores de la que se denomina variable respuesta, dependiente, explicada o regresando; $\boldsymbol{x} = (x_1, x_2 \ldots x_p)$ representa el vector de las n observaciones en el conjunto de datos para las p variables denominadas explicativas, predictores, regresores o covariables; $\beta_0, \beta_1, \ldots, \beta_p$ son los parámetros desconocidos que deben ser estimados siendo β_0 el intercepto y ε_i es denominado el término de error o perturbación en la relación entre las covariables y la variable respuesta. En el término de error se recoge todo lo que no pueden explicar las covariables de y_i, errores de medida, etc.

El modelo clásico de regresión lineal se centra en el valor esperado de la variable respuesta $\boldsymbol{y}$, condicional a los valores de las covariables $\boldsymbol{x}$, lo que se expresa como $E(\boldsymbol{y}|\boldsymbol{x})$ y ha sido ampliamente utilizado por su propiedad de proveer una descripción parsimoniosa

y completa de la relación entre las covariables y la variable respuesta, siempre que se cumplan los supuestos básicos, a saber:

1. *Linealidad.*

   Este supuesto se basa en la forma genérica del modelo descrita en la Ecuación 3.1. La linealidad hace referencia a la forma en que los parámetros y la perturbación entran a formar parte de la ecuación de regresión y no necesariamente a la relación entre las variables y nos viene a decir que las covariables y la perturbación serán modeladas como una combinación lineal. Este supuesto parece muy restrictivo pero podemos tener relaciones no lineales entre las variables que pueden ser llevadas a una estructura lineal. De ahí que ecuaciones tales como $\boldsymbol{y} = \alpha + \boldsymbol{\beta}\cos(\boldsymbol{x}) + \boldsymbol{\varepsilon}$, $\ln \boldsymbol{y} = \alpha + \boldsymbol{\beta} \ln \boldsymbol{x} + \boldsymbol{\varepsilon}$ sean lineales en alguna función de $\boldsymbol{x}$.

2. *Identificabilidad de los parámetros del modelo.*

   Este supuesto sifgnifica que no existe relación lineal exacta entre las variables incluidas como regresores lo que se traduce en que $\boldsymbol{x}$ será una matriz de rango completo, o sea $\text{rank}(\boldsymbol{x}) = p + 1$[1]. El supuesto de independencia entre las columnas de la matriz $\boldsymbol{x}$ es necesario para garantizar la obtención de una única estimación de los parámetros desconocidos $\boldsymbol{\beta}$.

3. *Supuestos sobre el término de error.*

   El valor esperado de la perturbación aleatoria condicionando en x es igual a cero, o sea $E(\boldsymbol{\varepsilon}|\boldsymbol{x}) = 0$. El supuesto de media cero de las perturbaciones implica que $E(\boldsymbol{y}|\boldsymbol{x}) = \boldsymbol{x}'\boldsymbol{\beta}$.

   Se asume además que la varianza es constante a través de las distintas observaciones

   $$Var[\varepsilon_i|\boldsymbol{x}] = \sigma^2 \qquad \forall \quad i = 1, \ldots, n$$

   Los errores son llamados heterocedásticos cuando la varianza de los mismos varía entre las observaciones, $Var(\varepsilon_i) = \sigma_i^2$. En adición al supuesto de homocedasticidad, se supone que los errores no están correlacionados $Cov[\varepsilon_i, \varepsilon_j|\boldsymbol{x}] = 0 \qquad \forall \quad i \neq j$.

4. *La variable respuesta $\boldsymbol{y}$ y las covariables $\boldsymbol{x}$ son estocásticos.*

   Este supuesto significa que los valores de los regresores y la variable respuesta son realizaciones de variables aleatorias.

[1] Esto implica que el número de observaciones debe ser al menos igual o mayor al número p de parámetros desconocidos.

5. *La perturbaciones del modelo siguen una distribución normal con media cero y varianza constante.*

   De cara a la inferencia, se supone en la mayoría de las situaciones que los errores siguen una distribución normal. Este supuesto junto con el supuesto (3) permite llegar al resultado de que $\varepsilon_i \sim NID(0, \sigma^2)$.

Si definimos los vectores

$$\boldsymbol{Y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} \quad \text{y} \quad \boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

así como la matrix $\boldsymbol{X}$

$$\boldsymbol{X} = \begin{pmatrix} \mathbf{1} & x_{11} & x_{12} & \dots & x_{1p} \\ \mathbf{1} & x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & \vdots & \vdots & & \vdots \\ \mathbf{1} & x_{n1} & x_{n2} & \dots & x_{np} \end{pmatrix}$$

Entonces el modelo de la Ecuación 3.1 puede ser escrito en forma matricial

$$\boldsymbol{Y} = \boldsymbol{X}'\boldsymbol{\beta} + \boldsymbol{\varepsilon} \qquad \boldsymbol{\varepsilon} \sim N\left(0, \sigma^2 \boldsymbol{I}_n\right) \tag{3.2}$$

donde $\boldsymbol{Y}$ y $\boldsymbol{\varepsilon}$ son vectores $n \times 1$, $\boldsymbol{X}$ es una matriz conocida de orden $n \times (p+1)$ y $\boldsymbol{\beta}$ es el vector de orden $(p+1) \times 1$ de parámetros desconocidos. $\boldsymbol{I}_n$ es la matriz identidad $n \times n$.

Tomando valor esperado en la Ecuación 3.2 y sabiendo que cualquier transformación o función lineal de una v.a normal es también una v.a normal, tenemos

$$\boldsymbol{Y} \sim N\left(\mu, \sigma^2 \boldsymbol{I}_n\right) \text{ donde } E[\boldsymbol{Y}|\boldsymbol{X}] = \boldsymbol{X}'\boldsymbol{\beta} \tag{3.3}$$

Para estimar el vector de parámetros desconocidos $\boldsymbol{\beta}$ el método de los mínimos cuadrados ordinarios (MCO)[2] ha sido el más empleado, el cual minimiza la suma de los cuadrados de los residuos [3] teniendo en cuenta el cumplimiento de los supuestos básicos del modelo de

[2] Carl Friedrich Gauss (1977-1855) fue el primero en utilizar este método para predecir el recorrido del asteroide Ceres en 1801.

[3] Llamamos residuos a la desviación entre el valor observado de la variable respuesta y su valor estimado es $\hat{\varepsilon}_i = y_i - \boldsymbol{x}_i'\hat{\boldsymbol{\beta}}$

regresión lineal. De la solución al problema planteado en la Equación (3.4) (para lo que es necesario que la matriz $\boldsymbol{X}$ sea de rango completo y que $\boldsymbol{X'X}$ sea invertible) se llega a la estimación del vector de parámetros desconocidos $\boldsymbol{\beta}$

$$\begin{aligned}\hat{\beta} &= \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \left\{ \sum_{i=1}^{n} \left(y_i - \boldsymbol{x}_i'\boldsymbol{\beta}\right)^2 \right\} \\ \hat{\beta} &= \underset{\beta}{\operatorname{argmin}} \left(\boldsymbol{Y} - \boldsymbol{X\beta}\right)' \left(\boldsymbol{Y} - \boldsymbol{X\beta}\right) \\ \hat{\beta} &= \underset{\beta}{\operatorname{argmin}} \left(\boldsymbol{Y'Y} - 2\boldsymbol{Y'X\beta} + \boldsymbol{\beta'X'X}\beta\right)\end{aligned} \tag{3.4}$$

La función de verosimiltud queda planteada de la siguiente forma

$$L\left(\beta, \sigma^2\right) = \left(2\pi\sigma^2\right)^{-\frac{n}{2}} exp\left\{ -\frac{1}{2\sigma^2} \left(\boldsymbol{Y} - \boldsymbol{X}\beta\right)' \left(\boldsymbol{Y} - \boldsymbol{X}\beta\right) \right\} \tag{3.5}$$

Si se aplica logaritmo a la función de verosimiltud

$$\ln\left[L\left(\beta, \sigma^2\right)\right] = -\frac{n}{2} \ln\left(2\pi\sigma^2\right) - \frac{1}{2\sigma^2} \left(Y - X\beta\right)' \left(Y - X\beta\right) \tag{3.6}$$

Maximizando el logaritmo de la función de verosimilitud en la Ecuación 3.6 para $\boldsymbol{\beta}$, obtenemos el mismo resultado que si se minimiza la cantidad mínimo cuadrática en la Ecuación 3.4, por lo que el estimador máximo verosímil (MLE) y el estimador mínimo cuadrático (MCO) para $\boldsymbol{\beta}$ son iguales

$$\hat{\beta} = \left(\boldsymbol{X'X}\right)^{-1} \left(\boldsymbol{X'Y}\right) \tag{3.7}$$

La inferencia en estos modelos queda supeditada al supuesto de que la variable respuesta sigue una distribución normal. El modelo de regresión lineal es, en muchas situaciones, de gran utilidad puesto que:

- Provee una descripción simple de los datos.
- Si el modelo es lineal en las variables, se puede resumir la contribución de cada predictor sobre la variable respuesta a través de un simple coeficiente.
- Proporciona un método sencillo para la predicción de nuevas observaciones para la variable respuesta.

Diferentes generalizaciones han sido implementadas para permitir que los errores no sean esféricos pero aún todos los esfuerzos realizados para emplearlo en distintos ámbitos del análisis econométrico, el mayor inconveniente del modelo clásico de regresión lineal es que sólo se fija en un valor de toda la distribución condicional de $\boldsymbol{y}$: la media.

Los modelos lineales generalizados (GLM) de Nelder and Wedderburn (1972), los modelos aditivos generalizados (GAM) de Hastie and Tibshirani (1990), los modelos parcialmente lineales (PLM) de Engle et al. (1986) y de Härdle and Liang (2007) son exponentes de lo que se ha realizado para encontrar estimaciones cada vez mejores. En este sentido, el uso de modelos no paramétricos y semiparamétricos ha demostrado reducir el error de especificación al compararlos con los modelos paramétricos.

3.2. Los Modelos Lineales Generalizados (GLM)

Los modelos lineales generalizados (GLM) (Nelder and Wedderburn, 1972) son una extensión de los modelos de regresión lineal clásicos, permitiendo que la distribución de la variable respuesta sea diferente a la distribución normal y que se relajen los supuestos de errores incorrelados y varianza constante.

McCullagh and Nelder (1989) plantean que un GLM consta de un componente aleatorio el cual identifica la variable respuesta $\boldsymbol{y}$ y su función de distribución de probabilidad; un componente sistemático que especifica los regresores $\boldsymbol{x} = [x_1, x_2, \dots x_p]$ los cuales producen un predictor lineal η_i dado por:

$$\eta_i = \alpha + \sum_{j=1}^{p} x_j \beta_j \tag{3.8}$$

y una función de enlace (*link function*) (g), conocida, que une la parte sistemática con la aleatoria:

$$\eta_i = g\left(\mu_i\right) \tag{3.9}$$

La diferencia que hay entre usar la función *link* y usar una transformación, es que la función *link* transforma el valor esperado de la variable respuesta $E[\boldsymbol{y}]$ y no la variable respuesta $\boldsymbol{y}$.

En estos modelos se asumen los siguientes supuestos en relación a la influencia de las covariables $\boldsymbol{x_i}$ sobre la variable respuesta $\boldsymbol{y}$ (Fahrmeir et al., 2013):

(i) ***Distribución de la variable respuesta y*:**

Condicional a las covariables $\boldsymbol{x}$, la variable respuesta $\boldsymbol{y}$ sigue una distribución de la familia exponencial, dentro de la cual se incluyen distribuciones como la Poisson, Binomial, Gamma y la Normal [4]; siendo la función de densidad de probabilidad de la forma :

$$f\left(y_i \mid \theta_i, \phi, w_i\right) = exp\left\{\frac{y_i'\theta_i - b\left(\theta_i\right)}{\phi} w_i + c\left(y_i, \phi, w_i\right)\right\} \qquad i = 1, \ldots, n \tag{3.10}$$

donde $b\left(\cdot\right), c\left(\cdot\right)$ son funciones conocidas y determinan qué distribución de la familia exponencial se va a tener en cuenta; θ es denominado el parámetro natural o canónico de la familia exponencial que representa la locación[5], w_i son los pesos y ϕ es el parámetro de escala o de dispersión independiente de i, por lo que es común para todas las observaciones.

(ii) ***Supuesto estructural*:**

La dependencia de la media condicional $E\left(y_i|x_i\right) = \mu_i$ sobre las covariables x_i es especificada a través de :

$$g\left(\mu_i\right) = x_i'\beta \tag{3.11}$$

Este vínculo entre el valor esperado y el predictor lineal η_i es posible por el nexo existente entre la parte sistemática y la parte aleatoria:

$$\begin{aligned} \mu_i &= h\left(\eta_i\right) \\ \eta_i &= g\left(\mu_i\right) \end{aligned} \tag{3.12}$$

donde h es denominada la función de respuesta natural y es una función suave y biyectiva, g es la inversa de h $(g = h^{-1})$ y es considerada la *link function* y $\boldsymbol{\beta}$ es el vector de parámetros desconocidos o vector de coeficientes de regresión que usualmente se estima por máxima verosimilitud (ML) usando el método de mínimos cuadrados ponderados iterativo (IWLS) (Zeileis et al., 2008).

[4]Las distribuciones más importantes de esta familia se recogen en la Tabla 3.1

[5]El parámetro natural puede ser expresado como una función de la media, o sea $\theta_i = \theta\left(\mu_i\right)$, de donde podemos obtener la media μ a través de $\mu_i = b'\left(\theta_i\right) = \frac{\partial b(\theta_i)}{\partial \theta}$. En la Tabla 3.1 aparecen las medias de las diferentes distribuciones de la familia exponencial, en función de los parámetros naturales.

En los GLM la varianza de y_i también depende del predictor lineal:

$$Var\left(y_i|x_i\right) = \frac{\phi V\left(\mu_i\right)}{w_i} \tag{3.13}$$

donde $V\left(\mu_i\right) = b^{''}\left(\theta_i\right)$ es la función de varianza de la familia exponencial subyacente.

La estructura básica de un modelo lineal generalizado puede, entonces, definirse como:

$$g\left(\mu_i\right) = \eta_i = \alpha + \sum_{j=1}^{p} x_j \beta_j \tag{3.14}$$

La *link function* $g\left(\cdot\right)$ relaciona el valor esperado de $\boldsymbol{y}$ con el predictor lineal $\boldsymbol{x'\beta}$:

$$\begin{aligned} \boldsymbol{\mu} &= E\left(\boldsymbol{y}\right) \\ \boldsymbol{\eta} &= g\left(\boldsymbol{\mu}\right) = \boldsymbol{x'\beta} \\ E(\boldsymbol{y}) &= g^{-1}\left(\boldsymbol{\eta}\right) \end{aligned} \tag{3.15}$$

Por ejemplo, si consideramos que la variable respuesta sigue una distribución de Poisson, sólo tendríamos un parámetro a tener en cuenta desde que en la distribución de Poisson la media y la varianza son iguales a λ:

$$\begin{aligned} y &\sim \mathcal{P}\left(\mu\right) \\ g\left(\mu\right) &= \beta_0 + \beta_1 x_1 + \ldots, \beta x_p \\ \mu &= e^{\beta_0 + \beta_1 x_1 + \cdots + \beta x_p} \\ g\left(\mu\right) &= \ln\left(\mu\right) \end{aligned} \tag{3.16}$$

Para la distribución de Poisson, la *link function* es, por tanto, el logaritmo neperiano de μ, relacionando éste con el predictor lineal.

Función	Notación	Rango de y	$b(\theta)$	$\mu(\theta)$	Canonical link $\theta(\mu)$	Variance $V(\mu)$	ϕ
Bernoulli	$B(1,\mu)$	$\{0,1\}$	$log\left(1+e^{\theta}\right)$	$e^{\theta}/\left(1+e^{\theta}\right)$	logit	$\mu(1-\mu)$	1
Binomial	$B(k,\mu)$	$\{0,1,\ldots,k\}$	$klog\left(1+e^{\theta}\right)$	$ke^{\theta}/\left(1+e^{\theta}\right)$	logit	$\mu\left(1-\frac{\mu}{k}\right)$	1
Poisson	$P(\mu)$	$\{0,1,2,\ldots\}$	$exp(\theta)$	$exp(\theta)$	log	μ	1
Geométrica	$GE(\mu)$	$\{0,1,2,\ldots\}$	$-log\left(1-e^{\theta}\right)$	$e^{\theta}/\left(1-e^{\theta}\right)$	$log\frac{\mu}{1+\mu}$	$\mu+\mu^2$	1
Binomial Negativa	$NB(\mu,k)$	$\{0,1,2,\ldots\}$	$-klog\left(1-e^{\theta}\right)$	$ke^{\theta}/\left(1-e^{\theta}\right)$	$log\frac{\mu}{k+\mu}$	$\mu+\frac{\mu^2}{k}$	1
Normal	$N\left(\mu,\sigma^2\right)$	$(-\infty,\infty)$	$\theta^2/2$	θ	identidad	1	σ^2
Exponencial	$Exp(\mu)$	$(0,\infty)$	$-log(-\theta)$	$-1/\theta$	recíproca	μ^2	1
Gamma	$G(\mu,\nu)$	$(0,\infty)$	$-log(-\theta)$	$-1/\theta$	recíproca	μ^2	$1/\nu$
Gausiana Inversa	$IG\left(\mu,\sigma^2\right)$	$(0,\infty)$	$-(-2\theta)^{1/2}$	$-(-2\theta)^{1/2}$	recíproca al cuadrado	μ^3	σ^2

Tabla 3.1: *Características de algunas distribuciones de la familia exponencial.*

La función de enlace puede ser una función monótona diferenciable, siendo un caso particular la denominada función natural (o canónica) que se obtiene como:

$$\theta_i = \theta\left(\mu_i\right) = \eta_i \tag{3.17}$$

Esto nos lleva a concluir que un modelo GLM específico es completamente determinado por la elección de un tipo de distribución de la familia exponencial, la elección de la función de enlace y la definición y selección de las covariables (Fahrmeir et al., 2013).

Los supuestos de linealidad e independencia de las observaciones que se asumen en los modelos lineales generalizados (GLM) han hecho que este método de estimación se vuelva restrictivo en determinadas situaciones:

- Puede existir una relación entre predictores y variable respuesta que no sea lineal y que además esa relación no lineal sea desconocida.
- Las observaciones pueden estar correladas en tiempo y espacio.
- La interacción entre predictores podría no ser lineal.
- No resulta estar bien recogida la heterogeneidad de los individuos o grupos de ellos.

Dentro de los modelos lineales generalizados (GLM) cabe destacar la distinción existente entre aquellas especificaciones que cuentan con una variable dependiente continua y los que tienen como variable dependiente una variable discreta.

3.2.1. Modelos para variable respuesta continua

Distribución Normal

El modelo de regresión lineal clásico puede considerarse dentro de los GLM, en este caso $\boldsymbol{y}$ sólo puede seguir una distribución normal y la *link function* es la función identidad, lo que significa que $g\left(\mu\right) = \mu$. En la Tabla 3.1 puede observarse que la varianza es constante y el parámetro de escala es igual a la varianza del término de error del modelo de regresión lineal.

Distribución Gamma

La función de respuesta natural es dada por $h(\eta) = -\eta^{-1} = \mu$. En la práctica se utilizan como *link function* la función logarítmica y como función respuesta se utiliza la función exponencial para garantizar la no negatividad de la función respuesta:

$$\begin{aligned} g(\mu) = log(\mu) &= \eta \\ h(\eta) = exp(\eta) &= \mu \end{aligned} \tag{3.18}$$

3.2.2. Modelos con variable respuesta discreta

Modelos con variable respuesta discreta: datos de recuento

Los modelos cuya variable respuesta es del tipo recuento (*count data*) son modelos con variable respuesta discreta entera positiva y que, por tanto, puede tomar los valores $0, 1, 2, \ldots, \infty$. La manera clásica de modelar este tipo de datos, sobre todo si el campo de variación de la variable discreta es pequeño es utilizando la distribución de Poisson.

La función de respuesta natural es exponencial: $h(\eta) = exp(\eta) = \mu$ y la *link function* es el logaritmo natural $g(\mu) = log(\mu) = \eta$. Estos modelos son conocidos como los modelos log-lineales y a diferencia de los modelos con variable respuesta normal o Gamma, el parámetro de escala es $\phi = 1$. Lo que sucede es que este tipo de datos exhiben sobre dispersión, lo que es un problema dentro de los GLM así como exceso en el número de ceros.

Ante estos problemas, se suelen utilizar los denominados modelos cuasi-Poisson empleando un estimador robusto para la covarianza (*sandwich covariance*) introducido por White (1980) basado en los trabajos previos de Eicker et al. (1963) y Huber (1967) o estimando con un parámetro adicional para la dispersión. Otra forma de estimar estos modelos puede ser utilizando un modelo de regresión binomial negativo (NB)(McCullagh and Nelder, 1989). En el caso de la existencia de exceso de ceros pueden emplearse los modelos de Greene (1994) y Mullahy (1986).

Modelos con variable respuesta discreta: binaria y binomial

En los modelos con variable respuesta binaria $y_i \in \{0, 1\}$ el valor esperado es dado por la probabilidad $\pi = P(y = 1) \in [0, 1]$ lo que requiere una apropiada función respuesta que

garantice esto. En cualquier caso, el parámetro de escala es establecido $\phi = 1$.

Todos los modelos de regresión binaria combinan la probabilidad π_i con el predictor lineal η_i a través de una relación de la forma:

$$\pi_i = h(\eta_i) = h(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik}) \tag{3.19}$$

donde h es una función de distribución acumulada monótana creciente lo que garantiza que $\pi = P(y = 1) \in [0, 1]$ y hace que la relación entre π_i y η_i pueda expresarse como:

$$\eta_i = g(\pi_i) \tag{3.20}$$

Dentro de estos modelos debemos citar los modelos Logit, los modelos Probit y los modelos log-log complementarios (*complementary log-log model*).

En los modelos Logit la función respuesta es la función de distribución logística:

$$\pi = h(\eta) = \frac{exp(\eta)}{1 - exp(\eta)} = \pi \tag{3.21}$$

o equivalentemente, la *link function* puede escribirse como

$$g(\pi) = log\left(\frac{\pi}{1 - \pi}\right) \tag{3.22}$$

En los modelos Probit la función logística es sustituida por la función de distribución normal estandarizada. De esta forma

$$\pi = \Phi(\eta) = \Phi(x'\beta) \tag{3.23}$$

Los modelos log-log utilizan como función respuesta

$$h(\eta) = 1 - exp(-exp(\eta)) \tag{3.24}$$

cuya inversa es la *link function*

$$g(\pi) = log(-log(1 - \pi)) \tag{3.25}$$

Los modelos log-log son utilizados en aplicaciones más específicas como los modelos de duración discreta.

CAPÍTULO 4

Un modelo logit en el análisis de la autoselección de los cubanos que emigran a Estados Unidos

Este capítulo toma como referencia el artículo: **Aleida Cobas-Valdés y Ana Fernández-Sainz (2014). Cuban Migration to the United States and the Educational Self-Selection Problem. *International Journal of Cuban Studies*, Vol. 6, Nº 1 , pp. 41-54** y en el mismo se explica el papel de diferentes variables y en específico el nivel de educación en la probabilidad de emigrar de los cubanos. Se ha seleccionado el modelo logit en aras de dar una visión diferente al problema de autoselección de los individuos según su nivel de educación, que ha sido tratado en la literatura principalmente a través de los modelos gravitacionales.

4.1. Cuban Migration to the United States and the Educational Self-Selection Problem

Until the early twentieth century Cuba was considered a country of immigrants. Cuban people have been shaped by three major migration flows: European (mainly Spanish), African and Chinese, the most important being the Spanish power, involving around one and a half million people. In the second half of the nineteenth century one third of the Cuban population was born outside the island.

From 1850 to 1899, 900 thousand inmigrants entered Cuba, primarily Spanish immigrants, representing 90 % of European immigration. Mainly men working in the sugar and tobacco industry. In 1899, 10.97 % of the Cuban population was born abroad, 81 % of which were male. From 1902-1932, 1.25 million immigrants entered Cuba, of which 800 thousand

were Spanish (Pérez de La Riva, 2000). After 1926, immigration declined until becoming insignificant in 1930. The global crisis of 1929 and the subsequent collapse of the sugar industry in the early years of the 1930s resulted in the loss of the immigration country status that characterized Cuba up until then. In 1953, the proportion of people born abroad dropped to 3.95 % (Aja Díaz, 2006).

Cuba's external migration rate, defined as the ratio of the difference in the number of immigrants and emigrants with respect to average population, per 1.000 population has been negative for several decades. In the last 30 years it reached its lowest level in 1980 and 1994. In 1980 the figure reached 14.6 per thousand and in 1994, 4.4 per thousand, coinciding with two major waves of migration from Cuba to the United States, the first known as the *Puente Marítimo del Mariel* and the second as the *Crisis de los Balseros* (ONE, 2010).

United States has been for Cuba, and for other Latin American countries, the main destination of their migration. The U.S. Census for 2010 revealed that 50.5 million people (16.36 % of the entire population) in the United States are Hispanic, and this number rose from 35.3 million in 2000 to 50.5 million in 2010 (US Census Bureau, 2015). Of these Hispanics, 1.12 million were born in Cuba, representing 3 % of foreigners living in the U.S. (Motel and Patten, 2012).

Based on this data, the aim of this article is to analyse the characteristics, mainly educational, of Cubans who have emigrated to the United States and compare them with those of Cubans who have remained in Cuba. In this way we intend to address the self-selection problem among Cuban emigrants to the United States in terms of educational levels and analyse the importance of educational levels on the probability of Cuban migration.

The self-selection problem means that rational agents optimize their decision to participate in different markets, work, education, migration, etc and therefore the migrants choose markets that offer more attractive expectations. Roy (1951) was the first to address this problem, analysing how individuals optimize their decision to belong to one group or another in a given market, depending on their skills.

Self-selection not only depends on the unobservable characteristics of an individual such as ability, motivation, relatives or friends in the United States (Borjas, 1987) or access to financial resources (Chiquiar and Hanson, 2005) but mainly depends on observable characteristics such as education.

If there is a positive relationship between migration and education, i.e. more educated

persons migrate, we could be talking about the existence of human capital flight. If the migration of highly skilled labor is permanent, this process results in an increase in the growth potential of the receiving country of migrants and may represent a loss to the country of origin (Albo and Díaz, 2011).

The self-selection problem of Cuban migrants has not been addressed since Borjas (1991) in the context of migration to the United States from different countries, including Cuba.

4.1.1. Methodology

Consider perfectly rational individuals, so they will always choose the alternative that is in their best interests. The individual decision will be based on comparing the utility of living in their origin country with the expected utility in the destination country including the disutility of moving to that destination (Sjaastad, 1962).

Let U^e be the utility that reports Cuban migration to the United States and U^{no} the no-migration utility, so that:

$$\begin{aligned} U^e &= \beta_e' X + \varepsilon_e \\ U^{no} &= \beta_{no}' X + \varepsilon_{no} \end{aligned} \tag{4.1}$$

The X vector consists of a set of individual, observable characteristics, such as education, age, gender, professional category, etc.

The parameters vector β reflects the impact that covariate X has on the individual utility, ε_e and ε_{no} are disturbances or error terms and are considered independent of vector X and it is assumed that they follow a logistic distribution. Error terms, ε_e and ε_{no} may be related to each other with correlation coefficient ρ.

An individual will migrate if the utility from migrating is higher than the utility from not migrating, ie, if:

$$\begin{aligned} U^e &> U^{no} \\ \beta_e' X + \varepsilon_e &> \beta_{no}' X + \varepsilon_{no} \end{aligned} \tag{4.2}$$

Taking account that the utility (to migrate or not migrate) is unobservable, what we observe is the decision taken by the individual. We assume $Y = 1$ when the individual

selects the alternative to emigrate and $Y = 0$ when the individual selects the alternative of not migrating, so that:

$$\begin{aligned} Prob[Y = 1|X] &= Prob[U^e > U^{no}|X] \\ &= Prob[\beta'_e X + \varepsilon_e - \beta'_{no} X - \varepsilon_{no} > 0|X] \\ &= Prob[(\varepsilon_e - \varepsilon_{no}) > -(\beta_e - \beta'_{no} X)|X] \end{aligned} \tag{4.3}$$

where $[(\varepsilon_e - \varepsilon_{no}) > -(\beta_e - \beta'_{no} X)|X]$ is the self-selection condition.

Emigrating implies costs associated with migration formalities (both in the origin country in case of legal migration and in the destination country) and transport. Most illegal immigrants have to advance money to enter the destination country and therefore they must also consider costs related to the decision to emigrate (Orrenius and Zavodny, 2005).

Let π the migration cost in "time-equivalent" units (number of working hours required to migrate). Borjas (1987), (Borjas, 1991) assumes that π is constant, whereby implying that all individuals require the same number of working hours to cover the migration costs. Without loss of generality, this article assumes that π is constant. Borjas (1991) demonstrated that considering π as a random variable does not lead to particularly different results compared to those obtained when it is considered to be constant.

Now for the probability that an individual migrates the vector conditional on observable characteristics that are taken into account, will be raised as follows:

$$\begin{aligned} Prob[Y = 1|X] &= Prob[U^e - U^{no} - \pi > 0|X] \\ &= Prob[\beta'_e X + \varepsilon_e - \beta'_{no} X - \varepsilon_{no} - \pi > 0|X] \\ &= Prob[(\varepsilon_e - \varepsilon_{no}) > \pi - (\beta_e - \beta'_{no} X)|X] \\ &= F(\beta' X) \\ &= \Lambda(\beta' X) \end{aligned} \tag{4.4}$$

where $\Lambda(\beta' X)$ is the Cumulative Logistics Distribution Function and β is the parameter vector.

Emigration of those with a high level of education will be more likely if the education has a greater performance in the United States than in Cuba, that is, if the difference between

the parameter associated with educational level in the United States and the parameter associated with educational level in Cuba in the regression equation is positive. Hence, the most qualified individuals find incentives to migrate. This implies positive selection of individuals depending on their educational level.

Emigration of Cubans with low education levels will be more likely if the difference between the parameter associated to the educational level in the United States and the parameter associated to educational level in Cuba in the regression equation is negative, which implies that individuals with a high educational level will have little incentive to migrate and therefore results in negative selection (Borjas, 1991).

4.1.2. Data

The data used in this paper come from the random sample of 1 % of the 2010 U.S. Population and Housing Census, provided by Ruggles et al. (2011a). This sample includes only individuals who entered the United States at the age of 17 or over. This approach intends to avoid people who have completed their training in the United States (Lowell et al., 2008). In addition, we have considered only individuals under the age of 50, since the group between 16 and 49 are those who most migrate for economic reasons (Bertoli et al., 2013).

The sample above is completed with sample of Cubans living in Cuba in 2002 provided by Ruggles et al. (2011b) which corresponds to a 10 % random sample of the Population and Housing Census of Cuba. As in the sample of Cubans in U.S, we considered only individuals between 17 and 49 years old.

In both samples we have only considered working individuals. Thus, we have a total of 12.176 observations for Cubans in the United States and for Cubans in Cuba 81.641 observations. Table 4.1 describes the characteristics of the samples used.

In the Cubans in Cuba sample, the most commonly observed age group are individuals between 33 and 40 years old (34.7 %), while for the Cubans in the United States sample the most observed group are individuals between 41 and 49 years old (43.2 %).

The proportion of Cubans in Cuba between 17 and 24 years old is 3 times higher than that of Cubans in the United States in this age group. Figure 4.1 shows that 50 percent of Cubans in Cuba is made up of those aged 35 years and above and that 50 % of the Cubans in the United States sample are 39 years old or more. The mean age for Cubans

	Cubans in Cuba		Cubans in USA	
	Absolute Frequencies	Relative Frequencies	Absolute Frequencies	Relative Frequencies
Educational Level				
0-8 years	7892	0.097	479	0.039
9-12 years	61643	0.755	6690	0.549
13 or more years	12106	0.148	5007	0.411
Observations	81641	1.000	12176	1.000
Age				
17-24 years	9415	0.115	477	0.039
25-32 years	22843	0.280	2028	0.167
33-40 years	28309	0.347	4417	0.363
41-49 years	21074	0.258	5254	0.432
Observations	81641	1.000	12176	1.000
Occupational Category				
Category 1	16274	0.199	3202	0.263
Category 2	12335	0.151	774	0.064
Category 3	53032	0.650	8200	0.673
Observations	81641	1.000	12176	1.000

Tabla 4.1: *Sample Description.*

in the United States is 38 years while in the case of those who have not migrated is 35 years. It is therefore younger Cubans who not emigrate.

In the case of Cubans in United States sample, the percentage of individuals with more than 12 years of education is higher (41 %) whereas in the case of Cubans in Cuba sample, this group represents 15 % of individuals, whereby the proportion of Cubans with higher education in the U.S. is almost 3 times higher than the proportion of people with the same educational level in Cuba. This fact suggests that those who have received more education decided to migrate.

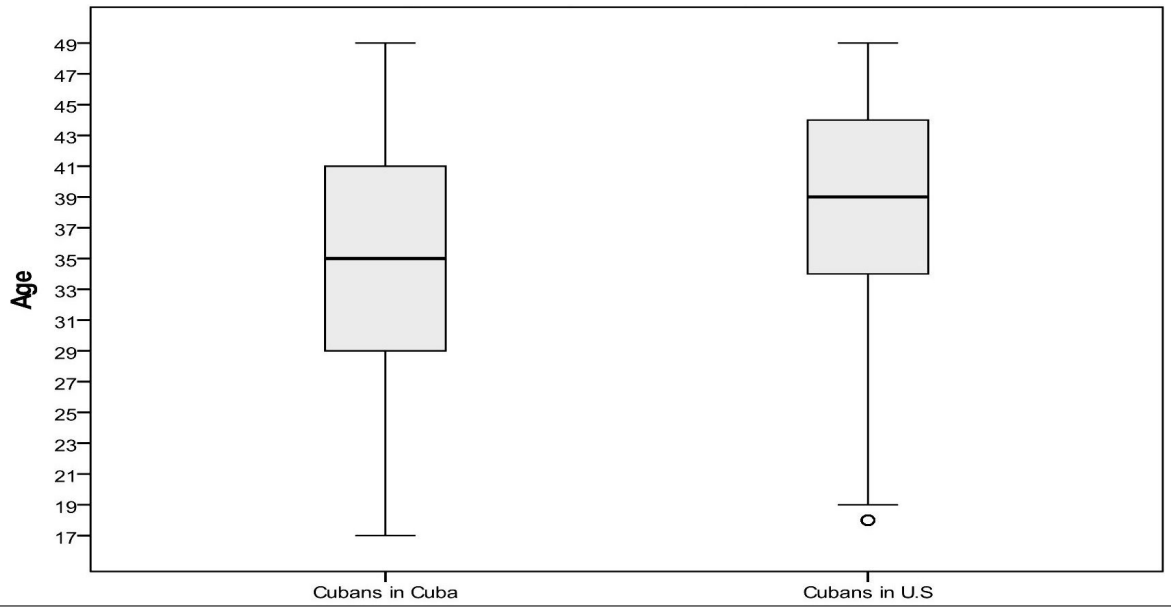


Figura 4.1: *Boxplot of Age.*

Only 4 % of the Cubans in the United States have less than 9 years of education. This would indicate that these individuals do not have incentives to emigrate to the United States since, possibly, they are under-valued in the U.S. labor market.

In both samples we are more likely to find individuals with between 9 and 12 years of education, 76 % of Cubans in Cuba have reached this level of education and 55 % in the case of the Cubans in the U.S sample. In the sample of Cubans in the United States the mean education years figure is 12.91, in the case of Cubans living in Cuba this mean is 11.15 years. In both samples about 50 percent of the individuals observed have 12 or more years of education.

These results are consistent with the data provided by US Census Bureau (2010), indicating that 20.8 % of the Cuban population in the U.S. have studied between 9 and 12 years and 30.2 % have received 13 or more years of education.

Regarding professional category, in both samples the highest percentage of individuals fall in what we have called Category 3 (Skilled Workers). In Cuba they represent 65 % of the observations while in the United States they represent 67 % of the sample. For Category 2 (Technical Level) the proportion of individuals in Cuba that work in this category is 2.35 times higher than the proportion of individuals in this category in the United States sample. For Category 1 (Academics-Managers-Executives) the percentage of individuals in the U.S. sample is 26 % while in the Cuba sample they represent 20 %.

4.1.3. Empirical Results

To study the self-selection problem in terms of levels of education of Cubans who emigrate to the United States, we estimate a cumulative logit model (Section 3.2.2 to calculate the migration probability using the sample described in the previous section. Our interest is to analyse the impact of variables such as age, level of education and professional category on the migration probability. In particular we have calculated the impact of education, so as to conclude what kind of self-selection problem we are facing.

The variables used are described in Table 4.2 with the main estimation results figuring in Table 4.3. According to the obtained results, individuals with higher educational levels are more likely to migrate. Keeping all other variables constant, the migration probability of an individual with 9 to 12 years of education would be multiplied by 0.165, while the probability of an individual with less than 9 years of education would be multiplied by 0.067.

A decrease in the probability of migration is higher if the individual's years of education is less. Thus, if we only take level of education into consideration in the model, the opportunity to emigrate for an individual with 13 or more years of education is 14.93 times higher than for an individual with 8 years or less of education and 6.06 times than for an individual having 9 to 12 years of education.

Figure 4.2 shows the behavior of the migration probability according to the years of education of the individual, keeping all other variables constant. It peaks when the individual has 13 or more years of education. This group contains individuals whose level of education is above the mean of years of studies in Cuba (11.15 years, as shown in the previous section). Hence, we can conclude that according to the results obtained with the maximum likelihood estimation, Cubans show positive self-selection, in terms of their educational level to immigrate to the United States.

Variable	Description
Age 1	Dummy variable = 1 if age is between 17 and 24 years old.
Age 2	Dummy variable = 1 if age is between 25 and 32 years old.
Age 3	Dummy variable = 1 if age is between 33 and 40 years old.
Age 4	Dummy variable = 1 if age is between 41 and 49 years old. Reference group for age.
Level 1	Dummy variable = 1 if the individual has between 0 and 8 years of education.
Level 2	Dummy variable = 1 if the individual has between 9 and 12 years of education.
Level 3	Dummy variable = 1 if the individual has 13 or more years of education. Reference group for educational level.
OCCAT 1	Dummy variable = 1 if the individual belongs to the professional status of Academics, Managers and Executives.
OCCAT 2	Dummy variable = 1 if the individual belongs to the professional status of Technical Professional.
OCCAT 3	Dummy variable = 1 if the individual belongs to the professional status of Qualified Workers. Reference group for professional category.

Tabla 4.2: *Model Variables Description.*

It must be taken into account that the individual's choice affects both the origin country and the host country. Migration of better educated individuals has a negative impact on countries of birth since the educational investment of these people is unrecoverable (Moraga, 2011). However, some authors claim that the emigration of skilled individuals has a positive impact on the origin country through remittances or investments made by migrants (Durand et al., 2001).

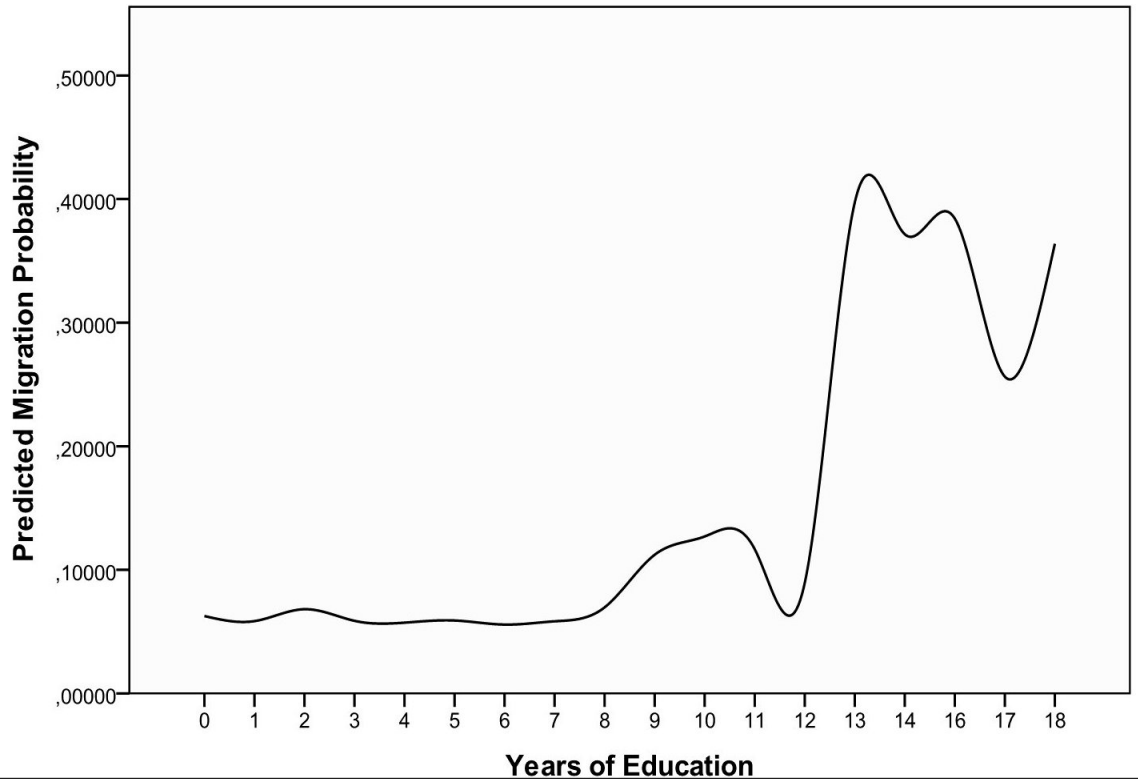


Figura 4.2: *Predicted Migration Probability depending on Years of Study.*

In terms of benefits to the host country, migration increases its production and technological capacity and it implies, from economic point of view, no significant cost in terms of social assistance, which would be required if the majority of those who migrate were people with less training (Cuecuecha, 2005).

In terms of age, if we only consider this variable in the model, the opportunity to migrate for an individual aged between 41 and 49 years is 1.59 times higher than that of an individual aged between 33 and 40 years, 2.73 times higher than that of an individual aged between 25 and 32 years and 3.94 times higher than that of an individual under 25 years. This result is consistent with the sample descriptive analysis we performed: the Cubans that migrated to the United States are older than Cubans who remained in Cuba.

Individuals with professional category 3 (Skilled Workers) have a migration opportunity of 2.95 times that of individuals in category 1 (Academics-Managers-Executives) and 3.04 times that of individuals with category 2 (Technical Professionals), if the model only considers this variable.

		Estimation (B)	Typical Error Mean	Wald Statistic	Degrees of Freedom	Significance Level	Exp(B)	I.C. 95% para EXP(B)	
								Inferior	Superior
Step 1[a]	LEVEL			4494.432	2	.000			
	LEVEL (1)	-1.919	.050	1475.023	1	.000	.147	.133	.162
	LEVEL (2)	-1.338	.021	3995.165	1	.000	.262	.252	.274
	Constant	-.883	.017	2760.832	1	.000	.414		
Step 2[b]	LEVEL			3824.577	2	.000			
	LEVEL (1)	-1.931	.050	1466.998	1	.000	.145	.131	.160
	LEVEL (2)	-1.231	.022	3259.769	1	.000	.292	.280	.304
	AGE			1677.654	3	.000			
	AGE(1)	-1.336	.050	711.606	1	.000	.263	.238	.290
	AGE(2)	-1.003	.028	1237.456	1	.000	.367	.347	.388
	AGE(3)	-.473	.023	420.939	1	.000	.623	.595	.652
	Constant	-.462	.021	486.706	1	.000	.630		
Step 3[c]	LEVEL			4525.688	2	.000			
	LEVEL (1)	-2.697	.055	2434.497	1	.000	.067	.061	.075
	LEVEL (2)	-1.803	.029	3930.423	1	.000	.165	.156	.174
	AGE			1662.855	3	.000			
	AGE(1)	-1.369	.050	737.059	1	.000	.254	.231	.281
	AGE(2)	-1.004	.029	1213.903	1	.000	.367	.346	.388
	AGE(3)	-.463	.023	390.160	1	.000	.629	.601	.659
	OCCAT			1632.515	2	.000			
	OCCAT(1)	-1.082	.031	1187.121	1	.000	.339	.319	.360
	OCCAT(2)	-1.112	.041	749.559	1	.000	.329	.304	.356
	Constant	.335	.030	123.197	1	.000	1.398		

Figura 4.3: *Estimation results: Cumulative Logit Model.*

4.1.4. Conclusions

In this paper we have analysed the self-selection problem of Cuban emigrants to the United States in terms of individual and observable characteristics: age, professional category and educational level. We have used the Population and Housing Census of the United States (2010) and the Census of Population and Housing for Cuba (2002). In both samples we have only considered workers aged between 17 and 49 years.

For the analysis, we have proposed a cumulative logit model that explains the choice of the individual, at the time of emigration, depending on the variables considered. The main conclusion obtained of the maximum likelihood estimation of the model is that Cubans positively self-select in their migration decision to move to the United States, in terms of educational level, that is, those with more years of study than the mean of the distribution of years of study in Cuba.

In terms of age, the most who migrate are people aged between 41 and 49 years. The Qualified Workers category also contributes to the migration probability, which is contrary to the fact that most migrants are the most qualified people. However, we will not consider this result to be conclusive given the fact that we have not considered whether individuals in both Cuba and the United States have a professional category in line with their educational level.

The positive educational self-selection problem has negative consequences on Cuba not only in terms of the non-recoverable educational investment but also, and more importantly, in terms of an important loss of human capital.

Parte III

La regresión cuantílica lineal

CAPÍTULO 5

Los modelos de regresión cuantílica lineal

Sería difícil resumir la amplia variedad de trabajos existentes en diferentes campos del conocimiento donde se estima la media condicional o una función cuantílica condicional. Supongamos que tenemos una muestra aleatoria $y_i, \boldsymbol{x_i}$ desde $i = 1, \ldots, n$ y queremos estimar la media condicional $\mu_y(x) = E(y|x)$ o la función cuantílica condicional $Q_\tau(x)$ de tal forma que $P(y \leq Q_\tau(x)|x) = \tau$ para algún $0 < \tau < 1$.

Si asumimos que $\mu_y(x)$ o $Q_\tau(x)$ son lineales, entonces se asumiría que $\mu_y(x) = \beta_0 + \beta_1' x$ y que $Q_\tau(x) = \beta_0 + \beta_1' x$ y esos parámetros desconocidos β se obtendrían a través de métodos ampliamente utilizados, como es el método de mínimos cuadrados ordinarios en el caso de la media condicional o el método de desviación absoluta mínima en el caso de la mediana condicional.

La regresión cuantílica lineal emergió como una alternativa para superar los problemas de los modelos de regresión lineal y para estudiar en detalle la influencia de las covariables en todas las locaciones de la distribución de la variable respuesta.

Alrededor del año 1760, Rudjer Joseph Boscovich (1711-1787), físico y matemático croata radicado en Italia propone un método de estimación consistente en minimizar la suma absoluta de las desviaciones, considerado mucho más antiguo que el método de estimación de mínimos cuadrados ordinarios y que actualmente se conoce como la regresión con respecto a la mediana (*median regression*), desde que la solución a ese problema de minimización nos proporciona un estimador para la mediana de la variable respuesta condionada a las covariables (Farebrother, 1999).

La regresión cuantílica se debe al trabajo de Koenker and Bassett (1978) y es ampliamente abordada en Koenker (2005) siendo el modelo de regresión cuantílica una extensión natural

del modelo de regresión lineal ya que el modelo de regresión lineal especifica el cambio en la media condicional de la variable respuesta asociado a los cambios de las covariables mientras el modelo de regresión cuantílica especifica los cambios en los cuantiles condicionados.

La regresión por cuantiles ha sido muy utilizada en campos como la medicina, la educación, la ecología y la economía siendo vital en los estudios donde se quiere analizar la relación entre la variable respuesta y las covariables describiendo puntos no centrales de la distribución.

El investigador podría estar interesado en no todos los cuantiles, sino en aquellos que son imprescindibles en su investigación, por ejemplo para estudiar la malnutrición infantil en los países pobres (ver Fenske et al., 2012 y Fenske et al., 2013a); analizar la obesidad infantil (ver Mitchell et al., 2013) o el valor en riesgo (VaR) (ver Gaglianone et al., 2012 y Xiao et al., 2015) o conocer el estado de las barreras de coral (ver Martínez-Silva et al., 2013).

Comparada con los modelos para la media condicionada, la regresión cuantílica ofrece un análisis más flexible de la relación estocástica entre variables aleatorias. Cuando la forma de la distribución de la variable respuesta depende de las covariables, cuando el término de error no es *iid* o cuando la variable respuesta no sigue una distribución de probabilidades conocida o existen muchos outliers en los datos y por ende la distribución no es simétrica, es muy útil utilizar la regresión cuantílica.

5.1. Los cuantiles y la función cuantílica.

> *We say that a student scores at the τ quantile of a standardized exam if he performs better than the proportion τ of the reference group of students and worse than the proportion $(1-\tau)$. Thus, half of students perform better than the median student and half perform worse. Similarly, the quartiles divide the population into four segments with equal proportions of the reference population in each segment. The quintiles divide the population into five parts; the deciles into ten parts. The quantiles, or percentiles, or occasionally fractiles, refer to the general case.* ***Koenker and Hallock (2001)***.

Siendo $\boldsymbol{y}$ la variable respuesta o dependiente y $\boldsymbol{x}$ la variable predictor o covariable, es objetivo de la Econometría y la Estadística inferir la relación existente entre $\boldsymbol{y}$ y $\boldsymbol{x}$. Los ***cuantiles*** pueden ser definidos como locaciones particulares de una distribución de tal forma que el τ-th cuantil es el valor q_τ tal que $P(\boldsymbol{y} \leq q_\tau) = \tau$. Otra forma de entender los

cuantiles consiste en plantear que

$$P(\boldsymbol{y} \leq q_\tau) \geq \tau \quad \text{y} \quad P(\boldsymbol{y} \geq q_\tau) \geq 1 - \tau \tag{5.1}$$

Lo que se interpreta como que la probabilidad de observar un valor igual o inferior a q_τ es al menos τ mientras que la probabilidad de observar un valor igual o superior a q_τ es al menos $1 - \tau$.

Si y es una variable continua con función de distribución acumulada $F(y)$ y función de densidad $f(y)$ tal que

$$F_{\boldsymbol{y}}(y) = F(y) = P(\boldsymbol{y} \leq y) \tag{5.2}$$

La función $Q(\tau)$ es denominada la función cuantílica de la distribución de y y puede ser definida como la inversa de la función de distribución acumulada (CDF)

$$Q_{\boldsymbol{y}}(\tau) = Q(\tau) = F_{\boldsymbol{y}}^{-1}(\tau) = inf\{y : F(y) \geq \tau\} \tag{5.3}$$

Para cada $-\infty < y < \infty$ y $\tau \in [0, 1]$ se cumple que $F(y) \geq \tau$ si y solo si $Q(\tau) \leq y$.

De esta forma si existe un y tal que $F(y) = \tau$ entonces $F(Q(\tau)) = \tau$ y $Q(\tau)$ es el valor más pequeño de y que satisface $F(y) = \tau$. Además si $F(y)$ es continua y estrictamente creciente, entonces $F^{-1}(\tau)$ es igual al número real y que será único tal que $F(y) = \tau$.

Algunas de las propiedades [1] más importantes de las funciones cuantílicas son las siguientes

1. A partir de la definición de $Q(\tau)$ como una función de distribución general.
   - $Q(\tau)$ es una función no decreciente en $(0, 1)$ con $Q(F(y)) \leq y$ para todo $-\infty < y < \infty$ para el cual $0 < F(y) < 1$.
   - $F(Q(\tau)) \geq \tau$ para cualquier $0 < \tau < 1$.
   - $Q(\tau)$ es continua desde la izquierda, o lo que es lo mismo $Q(\tau-) = Q(\tau)$.
   - $Q(\tau+) = \inf\{y : F(y) > \tau\}$ de manera que $Q(\tau)$ está acotada por arriba.
   - Puntos de discontinuidad en $F(y)$ son puntos planos en $Q(\tau)$ y puntos planos en $F(\tau)$ son puntos de discontinuidad en $Q(\tau)$.

[1] Para un estudio detallado de las funciones cuantílicas, es recomendable el libro de Nair et al. (2013).

2. Sea U es una v.a uniforme en el intervalo $[0, 1]$, entonces $y = Q(U)$ tiene función de distribución $F(y)$

$$P(Q(U) \leq y) = P(U \leq F(y)) = F(y)$$

3. Si $T(y)$ es una función no decreciente de y, entonces $T(Q(\tau))$ es una función cuantílica. de la misma forma que si $T(y)$ es una función no creciente de y, entonces $T(Q(1-\tau))$ es también una función cuantílica.

4. Si $Q(\tau)$ es una función cuantílica de y con función de distribución continua $F(y)$ y $T(\tau)$ es una función no decreciente tal que $T(0) = 0$ y $T(1) = 1$, entonces $Q(T(\tau))$ es una función cuantílica de una variable aleatoria con el mismo campo de variación de y.

5. La suma de dos funciones cuantílicas es también una función cuantílica.

6. El producto de dos funciones cuantílicas positivas es también una función cuantílica.

7. Si y tiene como función cuantílica a $Q(\tau)$ entonces $\dfrac{1}{y}$ tiene como función cuantílica a $\frac{1}{Q(1-\tau)}$

La mediana es el cuantil más conocido y divide la distribución de una variable en dos partes iguales, de manera que el 50 % de las observaciones tienen valores por encima de la mediana y el otro 50 % por debajo y se prefiere a la media por ser más robusta a la presencia de valores extremos.

5.2. Los cuantiles como centro de la distribución

Cuando nos referimos a la media condicional estamos hablando de la media condicional de una variable respuesta $\boldsymbol{y}$ dado uno o más predictores $\boldsymbol{x}$, pudiéndose entonces modelar la influencia de los predictores sobre la variable respuesta de una forma aditiva paramétrica: $\hat{\mu} = E(\boldsymbol{y}/X = \boldsymbol{x}) = \boldsymbol{x}'\beta$ ó de una forma aditiva no paramétrica: $\hat{\mu} = \sum_{j=1}^{p} f_j(\boldsymbol{x}_j)$.

Los cuantiles ofrecen una información más completa acerca de la distribución de $\boldsymbol{y}$ como una función de la covariable $\boldsymbol{x}$ que lo que puede ofrecer la media condicional. De ahí que si en lugar de la media, se utiliza la mediana como centro c de la distribución, ésta minimizaría el valor absoluto de la suma de las desviaciones de la variable $\boldsymbol{y}$ con respecto a su mediana, de tal forma que

$$Me = \arg\min_{c} E|\boldsymbol{y} - c| \tag{5.4}$$

El τ-th cuantil es tal que

$$q_\tau = \arg\min_{c} E\left[\rho_\tau\left(\boldsymbol{y} - c\right)\right] \tag{5.5}$$

donde $\rho_\tau\left(\cdot\right)$ denota la siguiente función de pérdida

$$\begin{aligned} \rho_\tau\left(y\right) &= \left[\tau - I\left(y < 0\right)\right]y \\ &= \left[\left(1-\tau\right)I\left(y \leq 0\right) + \tau I\left(y > 0\right)\right]|y| \end{aligned} \tag{5.6}$$

A partir de los valores muestrales, podemos entonces obtener estimaciones para la mediana $\hat{Me}$ como centro de la distribución.

La función de pérdida es una función asimétrica absoluta de ahí que sea una suma ponderada de desviaciones absolutas donde $(1-\tau)$ es la ponderación asignada a las desviaciones negativas y τ es la ponderación asignada a las desviaciones positivas.

Si definimos a $\boldsymbol{y}$ como una variable aleatoria discreta con función de cuantía $P\left(y\right) = P\left(\boldsymbol{y} = y\right)$, el problema de minimización planteado en la Ecuación (5.5) puede verse como

$$\begin{aligned} q_\tau &= \arg\min_{c} E[\rho_\tau\left(\boldsymbol{y} - c\right)] \\ &= \arg\min_{c}\left\{\left(1-\tau\right)\sum_{y\leq c}|y-c|f\left(y\right) + \tau\sum_{y>c}|y-c|f\left(y\right)\right\} \end{aligned} \tag{5.7}$$

Si la variable aleatoria es continua, se sustituyen los sumatorios por integrales, de manera que

$$\begin{aligned} q_\tau &= \arg\min_{c} E\left[\rho_\tau\left(\boldsymbol{y} - c\right)\right] \\ &= \arg\min_{c}\left\{\left(1-\tau\right)\int_{-\infty}^{c}|y-c|f\left(y\right)d\left(y\right) + \tau\int_{c}^{+\infty}|y-c|f\left(y\right)d\left(y\right)\right\} \end{aligned} \tag{5.8}$$

$f\left(y\right)$ es la función de densidad de probabilidad de la variable aleatoria $\boldsymbol{y}$. Utilizando la muestra que tengamos de la población en estudio, obtendremos el estimador muestral $\hat{q}_\tau$ para $\tau \in [0,1]$. Para $\tau = 0.5$ tendremos como solución la mediana planteada en la Ecuación 5.4.

5.3. Los cuantiles como solución del problema de minimización

Si asumimos, sin pérdida de generalidad que $\boldsymbol{y}$ es una variable aleatoria continua, el valor esperado del valor absoluto de la suma de desviaciones con respecto al valor central c se define como

$$\begin{aligned} E|\boldsymbol{y}-c| = \int_{y\in\Re}|y-c|f\left(y\right)dy &= \int_{y<c}|y-c|f\left(y\right)dy+\int_{y>c}|y-c|f\left(y\right)dy \\ &= \int_{y<c}\left(c-y\right)f\left(y\right)dy+\int_{y>c}\left(y-c\right)f\left(y\right)dy \end{aligned} \tag{5.9}$$

Desde que $E|\boldsymbol{y}-c|$ es una función convexa, diferenciamos $E|y-c|$ con respecto a c, e igualamos la derivada parcial a cero para obtener la solución al problema de minimización

$$\frac{\delta}{\delta c}E|\boldsymbol{y}-c|=0 \tag{5.10}$$

Derivando e integrando por partes, tenemos

$$\begin{aligned} &\left\{\left(c-y\right)f\left(y\right)\Big|_{-\infty}^{c}+\int_{y<c}\frac{\delta}{\delta c}\left(c-y\right)f\left(y\right)dy\right\}+ \\ &\left\{\left(y-c\right)f\left(y\right)\Big|_{c}^{+\infty}+\int_{y>c}\frac{\delta}{\delta c}\left(y-c\right)f\left(y\right)dy\right\}=0 \end{aligned} \tag{5.11}$$

y teniendo en cuenta que

$$\lim_{x\to-\infty}=\lim_{x\to+\infty}=0 \tag{5.12}$$

Para una función de densidad de probabilidad bien definida, tenemos

$$\left\{\underbrace{\left(c-y\right)f\left(y\right)\Big|_{y=c}}_{=0\text{ cuando y=c}}+\int_{y<c}f\left(y\right)dy\right\}+\left\{\underbrace{\left(y-c\right)f\left(y\right)\Big|_{y=c}}_{=0\text{ cuando y=c}}-\int_{y>c}f\left(y\right)dy\right\} \tag{5.13}$$

Usando la definición de la función de distribución acumulada (CDF) planteada en la Ecuación 5.2, este problema se reduce a

$$\begin{aligned} F(c) - [1 - F(c)] &= 0 \\ 2F(c) - 1 = 0 \Rightarrow F(c) &= \frac{1}{2} \Rightarrow c = Me \end{aligned} \tag{5.14}$$

siendo la mediana la solución a este problema de minimización.

La solución obtenida anteriormente no cambia si se multiplican las dos partes en las que se descompone $E|\boldsymbol{y} - c|$ por una constante: τ y $(1 - \tau)$ respectivamente, lo que permitiría formular el mismo problema para el cuantil τ (Davino et al., 2014). Utilizando el mismo procedimiento planteado en la Ecuación 5.5 obtenemos

$$\begin{aligned} \frac{\delta}{\delta c} E[\rho_\tau(\boldsymbol{y} - c)] &= \frac{\delta}{\delta c}\left\{(1-\tau)\int_{-\infty}^{c} |y - c| f(y)\, d(y) + \qquad \tau \int_{c}^{+\infty} |y - c| f(y)\, d(y)\right\} \\ &= (1-\tau) F(c) - \tau(1 - F(c)) \\ &= 0 \end{aligned} \tag{5.15}$$

Siendo por tanto q_τ la solución al problema de minimización

$$F(c) - \tau F(c) - \tau + \tau F(c) = 0 \quad \Rightarrow \quad F(c) = \tau \quad \Rightarrow \quad c = q_\tau \tag{5.16}$$

5.4. La regresión cuantílica lineal

La regresión cuantílica tiene como objetivo modelar la función cuantílica condicional de una variable continua $\boldsymbol{y}$ que depende de un conjunto de covariables $\boldsymbol{x}$. El modelo de regresión cuantílica simple se puede escribir como

$$y_i = \boldsymbol{x}_i^{'}\boldsymbol{\beta}_\tau + \varepsilon_{\tau i}, \qquad \varepsilon_{\tau i} \sim H_{\varepsilon_{\tau i}} \quad \text{sujeta a } H_{\varepsilon_{\tau i}}(0) = \tau \tag{5.17}$$

donde y_i y $\boldsymbol{x_i} = (1, x_{i1}, \ldots, x_{ik})$ denotan, respectivamente la variable respuesta y el vector de covariables, incluyendo el término independiente, para el individuo i. El efecto lineal específico de la variable $\boldsymbol{x}$ en el cuantil τ de la variable respuesta y_i, es dado por $\boldsymbol{\beta}_\tau = (\beta_{\tau o}, \beta_{\tau 1}, \ldots, \beta_{\tau k})$ con $\tau \in (0, 1)$.

Sobre el término de error $\varepsilon_{\tau i}$ con función de distribución acumulada $H_{\varepsilon_{\tau i}}$ se asume la restricción impuesta en la Ecuación 5.17, aparte de asumirse independencia, lo que implica que $H^{-1}_{\varepsilon_{\tau i}}(\tau) = 0$. Lo que se hace es sustituir el supuesto de media cero utilizado en los

modelos de regresión sobre la media por el supuesto de cero-cuantiles para los errores. No se tienen que establecer supuestos sobre la forma de la distribución de la variable respuesta.

Como consecuencia del supuesto de cero-cuantiles para el término de error, el predictor $\eta_{\tau i} = x_i'\beta$ es el τ-cuantil para la variable respuesta y_i desde que

$$\tau = H_{\varepsilon_{\tau i}}(0) = P\left(\varepsilon_{\tau i} \leq 0\right) = P\left(y_i - \eta_{\tau i} \leq 0\right) = P\left(y_i \leq \eta_{\tau i}\right) = H_{y_i}\left(\eta_{\tau i}\right) \tag{5.18}$$

Este supuesto sobre el término de error permite que se pueda asumir que los mismos sean no-normales, lo que es útil en muchas aplicaciones y permite que la regresión cuantílica sea aplicada hasta en situaciones de heterocedasticidad para el término de error (Reich et al., 2010).

En la Figura 5.1 se muestra el comportamiento de las curvas de regresión cuantílica ante el cumplimiento o no del supuesto de independencia en el término de error.

La función $Q_{\boldsymbol{y}_i}\left(\tau/x_i\right)$ sobre la variable respuesta continua y_i condicionada al vector de regresores $\boldsymbol{x}_i$ en el cuantil τ se puede escribir como

$$Q_{y_i}\left(\tau/x_i\right) = \boldsymbol{x}_i'\boldsymbol{\beta}_\tau = \eta_{\tau i} \tag{5.19}$$

donde los parámetros $\boldsymbol{\beta_\tau}$ cuantificarán la relación lineal entre las covariables y la función cuantílica de la variable respuesta para un determinado valor de τ.

Las funciones cuantílicas condicionadas son entonces estimadas, siguiendo a Koenker and Bassett (1978), minimizando la suma ponderada de desviaciones absolutas o lo que se denomina el $L_1 - norm$, lo cual es una extensión de la regresión lineal clásica donde las funciones de la media condicional son estimadas minimizando la suma de cuadrados de desviaciones o $L_2 - norm$.

$$\hat{\beta}_\tau = argmin_{\beta_\tau} \sum_{i=1}^{n} \rho_\tau\left(y_i - \eta_{i\tau}\right) \tag{5.20}$$

siendo ρ la función de pérdida o "*check function*"

$$\rho_\tau\left(u\right) = \begin{cases} u\tau, & u \geq 0 \\ u(1-\tau) & u < 0 \end{cases} \tag{5.21}$$

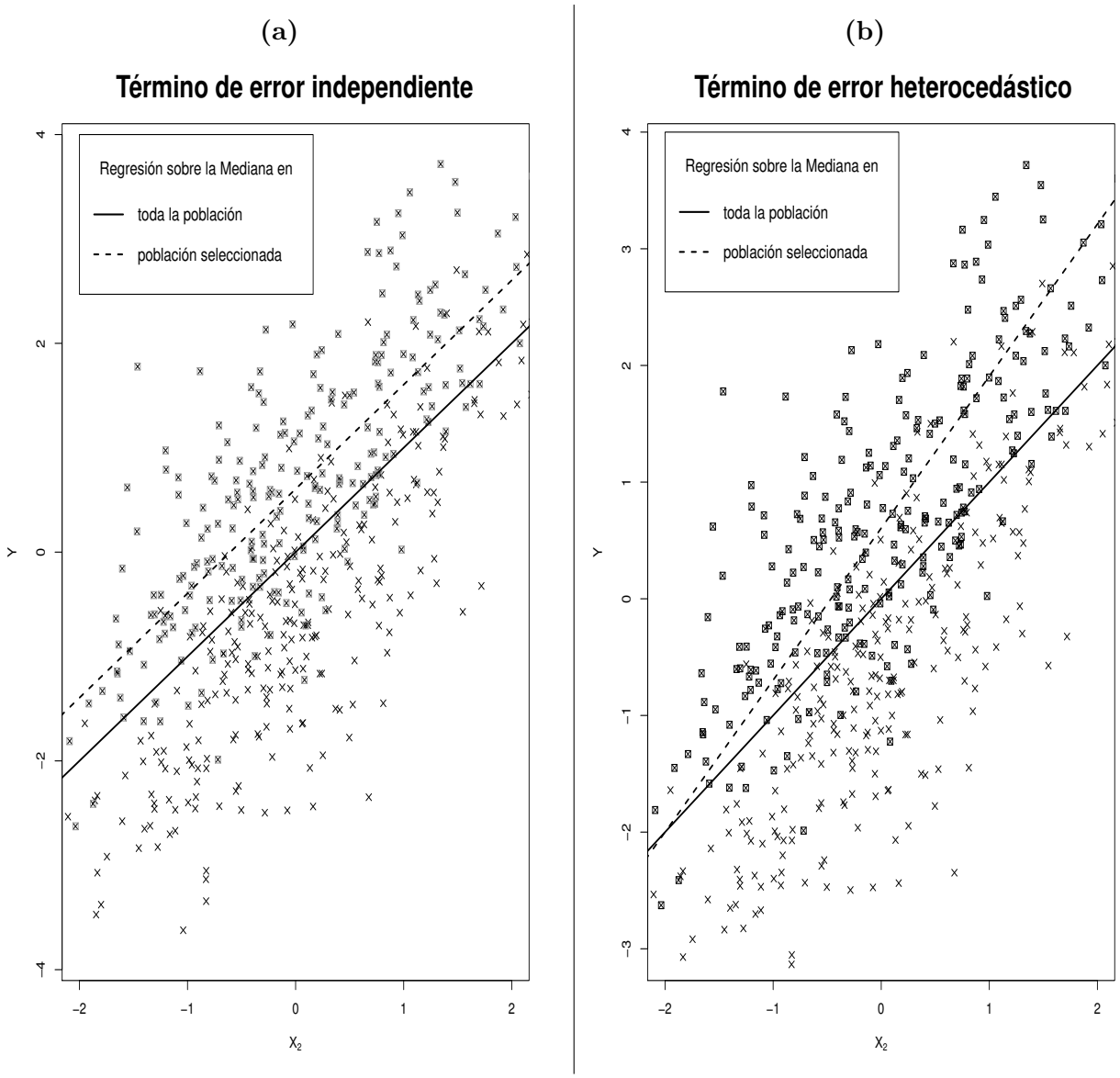


Figura 5.1: *Término de error en los modelos de regresión cuantílica.*

En la literatura también podemos encontrar la Ecuación 5.20 planteada como

$$\hat{\beta}_\tau = argmin_{\beta_\tau} \left\{ \tau \sum_{y_i \geq x_i' \beta_\tau} |y_i - x_i' \beta_\tau| + (1-\tau) \sum_{y_i < x_i' \beta_\tau} |y_i - x_i' \beta_\tau| \right\} \tag{5.22}$$

donde

$$\rho_\tau \left(y_i - \boldsymbol{x_i'} \boldsymbol{\beta_\tau} \right) = \begin{cases} \tau |\boldsymbol{x_i'} \boldsymbol{\beta_\tau}|, & y_i \geq \boldsymbol{x_i'} \boldsymbol{\beta_\tau} \\ (\tau - 1) \, |\boldsymbol{x_i'} \boldsymbol{\beta_\tau}| & y_i < \boldsymbol{x_i'} \boldsymbol{\beta_\tau} \end{cases} \tag{5.23}$$

Para la regresión cuantílica, la programación lineal es la forma estándar de estimar en una especificación paramétrica que permite una rápida optimización de la función de pérdida o la función asimétrica $L_1 - norm$ de pérdida (para detalles, Koenker (2005)).

Los coeficientes estimados se pueden interpretar como en los modelos clásicos de regresión lineal, es decir como tasas de rendimiento de las covariables sobre la variable respuesta en los diferentes puntos de la distribución condicional de la variable respuesta (Machado and Mata, 2005).

La regresión cuantílica permite que se pueda analizar toda la distribución de $\boldsymbol{y}$ condicional a $\boldsymbol{x}$ en cada una de las locaciones de la distribución y con ello obtendremos una fotografía más global de la interrelación entre las variables $\boldsymbol{y}$ y $\boldsymbol{x}$.

La principal ventaja de la regresión cuantílica lineal consiste en que, salvo la independencia, no es necesario ningún supuesto sobre la distribución del término de error lo que permite tener en cuenta diferentes tipos de distribuciones y además permite que no se tenga que asumir homocedasticidad. Así mismo, modelar la mediana en lugar de la media es mucho más robusto ante la presencia de valores extremos (Koenker, 2005).

Una desventaja de la regresión cuantílica según Kneib (2013) es que la función de distribución acumulada estimada para la variable respuesta es una función escalonada, cuando se supone continua. Este problema no es demasiado importante si el conjunto de datos es grande porque los escalones serían cada vez más pequeños.

Otro problema que surge en la regresión cuantílica es que la estimación $\hat{\beta}_{\tau 1}, \ldots, \hat{\beta}_{\tau q}$ de los cuantiles es obtenida de forma separada para cada cuantil $\tau_1 < \cdots < \tau_q$, lo que provoca que en muestras finitas, algunas curvas para los cuantiles puedan cruzarse obteniendo como resultado que la curva condicional al vector de covariables $\boldsymbol{x}$ no sea una función

monótona creciente de τ y esto invalida la distribución para la variable respuesta. Por ejemplo, puede ocurrir que estemos estimando una determinada variable respuesta y dado un vector de covariables, el percentil 95 estimado resulte más pequeño que el percentil 90, lo cual es imposible.

En este sentido se han realizado diferentes propuestas para evitar el problema del cruce de las curvas de regresión cuantílica. Entre otros, Koenker (2005) discute la potencial utilidad de la comonoticidad[2] de un grupo de variables aleatorias en el contexto de la regresión cuantílica; Neocleous and Portnoy (2008) proponen la interpolación de las curvas de regresión para garantizar asintóticamente que la probabilidad de cruzarse tienda a cero; Chernozhukov et al. (2009) proponen modificar la estimación de la función de distribución acumulada condicional de la variable respuesta; Bondell et al. (2010) proponen estimar los cuantiles simultáneamente imponiendo restricciones a la regresión cuantílica que fuercen el no cruce.

Los trabajos más recientes en este sentido son el artículo de Schnabel and Eilers (2013) donde se introduce una superficie bidimensional para las covariables $\boldsymbol{x}$ y para la probabilidad τ obteniendo la curva para un determinado cuantil en el sitio donde se corta la superficie a esa probabilidad τ a la que denominan *quantile sheet* y que se construye como una suma de productos de B-splines y el artículo de Cai and Jiang (2015) donde se desarrolla un método quasi-Bayesiano para estimar una secuencia de curvas de cuantiles condicionales. En ambos trabajos las curvas para los diferentes cuantiles se estiman de forma simultánea.

5.5. Inferencia en la regresión cuantílica lineal

5.5.1. Bondad del Ajuste

En el contexto de Mínimos Cuadrados Ordinarios (OLS) la variación de la variable respuesta se define en términos de desviaciones respecto de su media $(y_i - \bar{y}_i)$. Así la variación total de y es la suma de las desviaciones al cuadrado y se define como Suma

[2]En la teoría de la probabilidad, la comonoticidad es entendida como la dependencia positiva perfecta entre variables aleatorias y esencialmente dice que se pueden representar a través de funciones crecientes de una variable aleatoria. Por ejemplo los coeficientes de Kendall y Spearman son medidas bivariantes, invariantes ante transformaciones estrictamente monótonas, de la dependencia de variables aleatorias continuas llamadas a ser comonótonas (cuando la relación es lineal positiva) o countermonótonas (cuando la relación es lineal negativa).

Total de los Cuadrados (SCT)

$$SCT = \sum_{i=1}^{n} (y_i - \bar{y}_i)^2 \tag{5.24}$$

La variación muestral de los residuos $\hat{u}_i$ se medirá a través de la suma de los cuadrados de los residuos

$$SCR = \sum_{i=1}^{n} \hat{u}_i{}^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \tag{5.25}$$

La Bondad del Ajuste R^2 es entendida como el complemento a uno del cociente entre la Suma de Cuadrados de los Residuos (SCR) y la Suma Total de los Cuadrados (SCT) y es interpretada como la proporción de la variación de la variable respuesta que es explicada por las covariables incluidas en el modelo. Es un valor que se mueve entre 0 y 1 y un valor cercano a 1 indica que el modelo es un buen ajuste.

En el contexto de la regresión cuantílica, Koenker and Machado (1999) introdujeron el término de bondad de ajuste de forma análoga al utilizado en OLS sugiriendo medir la bondad del ajuste realizado para cada cuantil a través de la comparación de la suma de las desviaciones ponderadas del modelo de interés con la suma de las desviaciones ponderadas de un modelo donde sólo se tenga en cuenta el intercepto.

La función objetivo para cualquier τ en el modelo de regresión cuantílica es dada por

$$V(\boldsymbol{\beta}(\tau)) = \sum_{y_i \geq x_i'\beta} \tau |y_i - x_i'\beta| + \sum_{y_i < x_i'\beta} (1-\tau) |y_i - x_i'\beta| = \sum \rho(\varepsilon_i) \tag{5.26}$$

donde $\rho(\varepsilon_i) = \varepsilon_i (\tau - I(\varepsilon < 0))$ siguiendo a Koenker and Bassett (1978). Cuando $\tau = 0.5$ la función objetivo queda planteada como

$$V(\boldsymbol{\beta}(0.5)) = \sum |y_i - x_i'\beta| \tag{5.27}$$

El equivalente a la suma al cuadrado de los residuos en OLS sería la suma de los valores absolutos de los residuos ponderados en el modelo sin restringir o modelo de interés

$$SAR_\tau = \sum_{y_i \geq \times_i'\hat{\beta}} \tau |y_i - x_i'\hat{\beta}| + \sum_{y_i < x_i'\hat{\beta}} (1-\tau) |y_i - x_i'\hat{\beta}| \tag{5.28}$$

mientras que el equivalente a la suma total de cuadrados de la variable respuesta en OLS sería la suma total de los valores absolutos de las desviaciones ponderadas de la variable respuesta observada con respecto al cuantil τ estimado en un modelo donde sólo aparezca el intercepto

$$SAT_\tau = \sum_{y_i \geq \tau} \tau \left| y_i - \hat{\tau} \right| + \sum_{y_i < \tau} (1 - \tau) \left| y_i - \hat{\tau} \right| \tag{5.29}$$

El pseudo R^2 puede entonces obtenerse como

$$pseudoR^2_\tau = 1 - \frac{SAR_\tau}{SAT_\tau} \tag{5.30}$$

Como SAR_τ es siempre menor a SAT_τ, el $pseudoR^2$ tomará valores entre 0 y 1 y no puede ser considerado una medida de la bondad del ajuste del modelo completo porque hay que calcularlo cuantil a cuantil. Una extensión a esta forma de ver la bondad del ajuste puede ser incluir en el modelo restringido no sólo el intercepto sino un número inferior de covariables a las tenidas en cuenta en el modelo sin restringir (Hao and Naiman, 2007).

5.5.2. Errores Estándar e Intervalos de Confianza

Los errores estándar para la regresión cuantílica son sencillos de obtener si se supone que los errores son *iid*. En este caso, la matriz asintótica de covarianzas para $\hat{\beta}(\tau)$ es de la forma

$$V = \frac{\tau(1-\tau)}{n} \cdot \frac{1}{\left[f_{\varepsilon(\tau)}(0) \right]^2} (\boldsymbol{x}'\boldsymbol{x})^{-1} \tag{5.31}$$

donde $f_{\varepsilon(\tau)}(0)$ es la función de densidad del término de error evaluada en el τ-cuantil de la distribución de errores. Como en la regresión lineal condicionada a la media, la varianza asintótica es una escalar que multiplica a la matriz $(\boldsymbol{x}'\boldsymbol{x})$. Sin embargo aquí ese escalar $\frac{\tau(1-\tau)}{n} \cdot \frac{1}{\left[f_{\varepsilon(\tau)}(0) \right]^2}$ es la varianza asintótica del cuantil muestral, basado en la muestra $\varepsilon_1(\tau), \ldots, \varepsilon_n(\tau)$.

La función de densidad $f_{\varepsilon(\tau)}(0)$ denota la densidad de $\varepsilon_i(\tau)$ en cero. En el caso en que se consideren que los errores no son *iid*, los términos de error ε_i no tendrán ya una distribución común aunque sí se mantiene el supuesto de que la función de distribución de los errores

evaluada en cero es igual a τ[3].

La regresión cuantílica puede ser utilizada para construir intervalos de predicción para la variable respuesta (Meinshausen, 2006). Así si queremos obtener el $(1-\tau)\cdot 100\,\%$ intervalo de confianza para la variable respuesta, tenemos que regresar para los cuantiles $\tau = \alpha/2$ y $\tau = 1-\alpha/2$ y el intervalo para la nueva variable $\boldsymbol{y}$ estaría dado por

$$I_{1-\alpha}(x) = \left[\hat{Q}_{\boldsymbol{y}}\left(\frac{\alpha}{2}|X=x\right), \hat{Q}_{\boldsymbol{y}}\left(1-\frac{\alpha}{2}|X=x\right)\right] \tag{5.32}$$

Por ejemplo, el 95 % intervalo de confianza para el valor $\boldsymbol{y}$ es dado por

$$I(x)_{0.95} = \left[\hat{Q}_{0.025}(x), \hat{Q}_{0.975}(x)\right] \tag{5.33}$$

5.5.3. Contrastes de hipótesis

No existen muchos trabajos sobre tests de especificación en la regresión cuantílica. Koenker and Bassett (1982) implementaron un simple test para probar la igualdad entre los parámetros de las pendientes a través de los cuantiles, comparando el rango intercuantílico de dos muestras con lo que el test es utilizado para probar la existencia de heterocedasticidad.

Koenker and Machado (1999) definen un test de razón de verosimilitudes (LR) para verificar la significatividad conjunta de un subconjunto de covariables en el modelo. Considerando el modelo de regresión lineal $y_i = x_{1i}\beta_1 + x_{2i}\beta_2 + \varepsilon_i$ para el cuantil τ, siendo β_1 un vector $k-d$ y β_2 un vector con d coeficientes. La hipótesis nula sería $H_0 : \beta_2 = 0$.

Machado and Silva (2000) proponen un test para comprobar la existencia de heterocedasticidad. El estadístico de dicho test es n veces el R^2 de una regresión auxiliar $\rho_\tau(\hat{\varepsilon}_i(\tau))$ sobre una constante y una apropiada función de x. Este estadístico sigue una distribución $\chi^2_{(J-1)}$ siendo J el número de parámetros en la regresión auxiliar.

Furno (2007) implementa un test de razón de verosilimitudes para probar el cambio estructural en la regresión cuantílica lo que es de gran utilidad porque los cambios no tienen por qué tener el mismo impacto en cada uno de los puntos de la distribución y con este test se describe en qué nivel de la variable respuesta el cambio es más efectivo.

[3]Una descripción detallada de los métodos analíticos utilizados para aproximar los errores estándar es encontrada en Koenker (2004).

5.5.4. Distribución asintótica del estimador QR

Bajo determinadas condiciones de regularidad, incluyendo que los errores $\varepsilon_i(\tau)$ son *iid*

$$\sqrt{n}\left(\hat{\beta}(\tau) - \beta(\tau)\right) \xrightarrow{d} \mathcal{N}(0, V) \tag{5.34}$$

La matriz V ha sido definida en la Ecuación 5.31. La distribución asintótica de los estimadores en la regresión cuantílica bajo condiciones muy generales fue considerada en Koenker and Bassett (1982), Powell (1984), Chamberlain (1994) y Kim and White (2003), quienes plantearon que cuando los errores son independientes pero no idénticamente distribuidos se llega a un problema de mis-especificación del modelo

$$\sqrt{n}\left(\hat{\beta}(\tau) - \beta(\tau)\right) \xrightarrow{d} \mathcal{N}\left(0, D^{-1}AD^{-1}\right) \tag{5.35}$$

donde

$$D = E\left[f_{\varepsilon(\tau)}(0|x_i)\, x_i' x_i\right], \qquad A = E\left[\left(\tau - I\left(y_i < x_i'\beta(\tau)\right)\right)^2 x_i x_i'\right] \tag{5.36}$$

Si el modelo está correctamente especificado y los errores son *iid* entonces $D^{-1}AD^{-1} = V$.

Las matrices A y D pueden ser estimadas consistentemente lo que conllevaría a obtener una estimación consistente de la matriz de varianzas y covarianzas asintótica del estimador en la regresión cuantílica (Machado and Silva, 2013).

$$\hat{A} = \frac{1}{n}\sum_{i=1}^{n}\left(\tau - I\left(\hat{\varepsilon}_i(\tau) < 0\right)\right)^2 x_i' x_i \tag{5.37}$$

y

$$\hat{D} = \frac{1}{2n\delta_n}\sum_{i=1}^{n} I\left(-\delta_n \leq \hat{\varepsilon}_i(\tau) \leq \delta_n\right) x_i' x_i \tag{5.38}$$

siendo δ un parámetro de suavizamiento.

Buchinsky (1995) demuestra que si los datos son sospechosos de ser heterocedásticos, el mejor método para estimar la matriz asintótica de varianzas y covarianzas es el *Design Matrix Bootstrap Estimator (DMB)*. En cambio si los errores son independientes de los regresores es más conveniente estimarla a través del *Order Statistic Estimator (OS)* al ser

computacionalmente más sencillo.

5.6. Modelos similares a los Modelos de Regresión Cuantílica Lineal

Existen modelos que aún siendo considerados extensiones de los modelos GAM o GML tienen en común con la regresión cuantílica que la distribución de la variable respuesta puede obtenerse a partir de las covariables, lo que hace que puedan ser utilizados en situaciones similares en las que un modelo de regresión cuantílica sería el más recomendado. Dentro de estos modelos están los modelos aditivos generalizados estructurados (STAR) que serán analizados en la Sección 7.2, los modelos GAMLSS y la regresión por expectiles.

5.6.1. Los Modelos GAMLSS (Generalized Additive Models for Location, Scale and Shape)

Estos modelos surgieron por la necesidad de tratar algunas limitaciones de los modelos GLM y los modelos GAM. Introducidos por Rigby, R.A and Stasinopoulos, D.M (2001) y Rigby and Stasinopoulos (2005) son modelos que requieren supuestos para la distribución de la variable respuesta en el sentido paramétrico y en el sentido semiparamétrico a la hora de modelar los parámetros de la distribución como funciones de las covariables.

En estos modelos, el supuesto de distribución de la familia exponencial para la variable respuesta es relajado y sustituido por una familia de distribución general incluyendo distribuciones con gran simetría, teniendo curtosis y distribuciones discretas.

Dos supuestos son necesarios en este tipo de modelos: (i) las observaciones de la variable respuesta y_s son independientes entre sí y (ii) la variable respuesta sigue una distribución de probabilidades con función de densidad $f\left(y_s|\boldsymbol{\theta}\right)$ condicional al vector de parámetros $(\boldsymbol{\theta}_s = \theta_{s1}, \theta_{s2}, \theta_{s3}, \theta_{s4})'$ el cual puede contener hasta cuatro parámetros, a saber: la locación $(\theta_1 = \mu)$, la escala $(\theta_2 = \sigma)$, la simetría $(\theta_3 = \nu)$ y la curtosis $(\theta_4 = \varphi)$. Cada uno de esos parámetros θ_k con $k = 1, \ldots, 4$ es modelado por un predictor aditivo estructurado:

$$g_1\left(\mu_s\right) = \eta_s^{(\mu)} \quad g_2\left(\sigma_s\right) = \eta_s^{(\sigma)} \quad g_3\left(\nu_s\right) = \eta_s^{(\nu)} \quad g_4\left(\varphi_s\right) = \eta_s^{(\varphi)}$$

donde $g_1(\cdot), \ldots, g_4(\cdot)$ denotan funciones de enlace (*link function*) apropiadamente monótonas y los predictores aditivos estructurados $\eta^{(\mu)}, \eta^{(\sigma)}, \eta^{(\nu)}, \eta^{(\varphi)}$ pueden incluir diferentes covariables que pueden estar definidas como predictores en el marco de un modelo cuantílico aditivo estructurado (STAQ), que serán estudiados en la Sección 7.3.

Si se asume una distribución *Box-Cox Power Exponential* (BCPE) para la variable respuesta, la función cuantílica condicional puede ser expresada como

$$Q_{y_s}\left(\tau|\eta_s^{(\mu)}, \eta_s^{(\sigma)}, \eta_s^{(\nu)}, \eta^{s(\varphi)}\right) = \begin{cases} \mu_s\left(1+\sigma_s\nu_s\tilde{q_\tau}\right)^{1/\nu_s} & \text{si } \nu_s \neq 0 \\ \mu_s \exp\left(\sigma_s\tilde{q_\tau}\right) & \text{si } \nu_s = 0 \end{cases} \tag{5.39}$$

donde los parámetros $\tilde{q_\tau}$ dependen de la función cuantílica de una variable aleatoria que se distribuye como una Gamma (Rigby and Stasinopoulos, 2006).

5.6.2. La Regresión por Expectiles

Newey and Powell (1987) propusieron la regresión por expectiles como una alternativa a la regresión por cuantiles en momentos donde la forma del predictor aditivo afecte la diferenciabilidad del criterio de desviación absoluta en el que se fundamenta la regresión por cuantiles y por ello se necesiten criterios de optimización más específicos dentro de los que cabe el Boosting.

Es por esa razón que estos autores sustituyen la minimización de la suma asimétrica ponderada de las desviaciones absolutas planteada en la Ecuación 5.22 por la suma ponderada de las desviaciones al cuadrado dando lugar al problema de optimización

$$\hat{\beta_{i\tau}} = argmin_{\beta_{i\tau}} \sum_{i=1}^{n} w_{i\tau}\left(y_i - \eta_{i\tau}\right)^2 \tag{5.40}$$

donde $w_{i\tau}$ estaba definido en la regresión cuantílica lineal como ρ_τ. Por lo que se calcula como

$$w_{i\tau} = \begin{cases} \tau, & y_i \geq \eta_{i\tau} \\ 1-\tau & y_i < \eta_{i\tau} \end{cases} \tag{5.41}$$

Mientras los cuantiles son una generalización de la regresión con respecto a la mediana, los

expectiles son una generalización de la regresión con respecto a la media (Waltrup et al., 2015).

Así, como hemos visto en la Sección 5.4 la base de la regresión cuantílica consiste en minimizar la suma ponderada L_1, que no es más que la suma de las desviaciones absolutas entre la variable respuesta y el predictor.

En el caso de los expectiles, lo que se hace es minimizar la suma al cuadrado de las desviaciones $L_2 - norm$ asumiendo que el α-expectil del término de error es igual a cero. En realidad los expectiles tienen una estrecha vinculación con las estimación por mínimos cuadrados ordinarios, siendo la media un caso especial.

Yao and Tong (1996) muestran que los expectiles son en realidad cuantiles de una función de distribución únicamente relacionada con la distribución de $\boldsymbol{y}$ existiendo una única función biyectiva $h : (0,1) \to (0,1)$ tal que $q_\tau = m_{h(\tau)}$ siendo $h(\cdot)$ definida como

$$h(\tau) = \frac{-\tau q_\tau + G(q_\tau)}{-m_{0.5} + 2G(q_\tau) + (1-2\tau) q_\tau} \tag{5.42}$$

donde $G(q) = \int_{-\infty}^{q} y dF(y)$ es la función parcial de momentos y $F(y)$ es la función de distribución acumulada de y, cumpliéndose que $m_{0.5} = E(y) = G(\infty)$.

La regresión por expectiles está ganando adeptos en los últimos años, para ello cabe citar los trabajos de Schnabel and Eilers (2009), Sobotka and Kneib (2012), Waltrup et al. (2015) y Waltrup and Kauermann (2015) aunque presenta el inconveniente de que los coeficientes estimados no tienen una interpretación intuitiva como en el caso de los cuantiles.

5.7. Extensiones de la Regresión Cuantílica Lineal

En la literatura sobre especificación y estimación de los modelos de regresión cuantílica se distingue entre las propuestas que no establecen una distribución para el término del error, nombradas *Distribution-free approaches* y las propuestas que sí lo hacen y que se denominan *Distribution-based approaches*.

Las *Distribution-free approaches* están evidentemente relacionadas con el modelo de regresión cuantílica lineal que debemos a Koenker and Bassett (1978) y con extensiones del mismo dentro de las que es imprescindible citar el trabajo de Koenker et al. (1994) que constituye uno de los primeros intentos de estimar funciones no lineales suaves en modelos cuantílicos y de Koenker and Mizera (2004); los trabajo de Bollaerts et al. (2006), Wang

et al. (2009) y de Galvao and Montes-Rojas (2010), quienes utilizan los B-splines y los P-splines. En estos planteamientos se minimiza la función de pérdida a partir de los métodos de programación lineal así como con las más recientes técnicas del *Machine Learning* aplicado a la Estadística dentro de lo que se incluye la técnica Boosting, desarrollada en esta tesis en el Capítulo 7.

Dentro de los modelos incluidos en la propuesta *Distribution-based approaches* están aquellos en los que se asume una específica distribución para el término de error y la estimación se basa en la verosimilitud o en los métodos bayesianos. Aquí se incluyen modelos como los modelos quantílicos bayesianos cuyos primeros trabajos los encontramos en Yu and Moyeed (2001), Kottas and Gelfand (2001) y B Dunson et al. (2003).

En esta sección intentaremos hacer un breve repaso de algunas de las extensiones de la regresión cuantílica clásica.

5.7.1. Regresión Cuantílica por Splines

Hendricks and Koenker (1992) y Koenker et al. (1994) utilizan la regresión cuantílica con splines proponiendo los *Quantile Smoothing Splines* como minimización al siguiente problema:

$$min_{f \in C^1} \sum_{i=1}^{n} \rho_\tau \left[y_i - f\left(x_i\right) \right] + \lambda \int_0^1 \left(\left| f''\left(x\right) \right|^p dx \right)^{1/p} \tag{5.43}$$

donde $p \geq 1$ y C debe ser apropiadamente seleccionada.

Una forma típica de la utilización de los splines de suavizado es la introducción de un término de penalización $\lambda > 0$ en la clásica función objetivo de regresión (Fox 2000a):

$$min_{f \in C} \sum_{i=1}^{n} \left[y_i - f\left(x_i\right) \right]^2 + \lambda \int_{x_{min}}^{x_{max}} \left[f''\left(x\right) \right]^2 dx \tag{5.44}$$

Craig and Ng (2001) distinguen los dos términos que componen la ecuación anterior en fidelidad (*fidelity*) y penalización (*penalty*) donde el primer término mide la bondad de ajuste de la función estimada y el segundo representa el grado de suavidad del ajuste estimado. El parámetro λ juega el mismo papel que los anchos de bandas en los kernel o en los polinomios de regresión locales.

5.7.2. Modelos de Parámetros Cambiantes

Para introducir flexibilidad en el modelo de regresión, una de las especificaciones que se han utilizado consiste en los modelos de coeficientes cambiantes. Estos modelos a diferencia de los modelos de regresión lineal clásicos permiten que los coeficientes dejen de ser constantes y sean funciones que varían dependiendo de otra variable, que puede ser por ejemplo, el tiempo.

Andriyana et al. (2016) aproximaron cada función de los coeficientes a través de la utilización de los P-splines. Este método, enmarcado en el campo de la Estadística no paramétrica, fue introducido por Hastie and Tibshirani (1993) como una propuesta para tratar el problema de la dimensionalidad del vector de covariables.

En datos de sección cruzada, ha sido utilizado en artículos como los de Zhang et al. (2002) y Fan and Zhang (2008). En el tratamiento de datos longitudinales caben destacar los trabajos de Huang et al. (2004), Lin et al. (2007) y Ma and Song (2015). Concercientes a la regresión cuantílica podemos citar los artículos de Honda (2004), Kim (2007), Wang et al. (2009) y Andriyana et al. (2016).

De manera general, el planteamiento sería el siguiente para un vector $(T, X, Y) \in [0,1] \times R^{p+1} \times R$

$$q_\tau(t,x) = \beta_0(t)\, x_0 + \beta_1(t)\, x_1 + \cdots + \beta_p(t)\, x_p \tag{5.45}$$

donde $q_\tau(t,x)$ denota el τ cuantil condicional de Y, dado que $(T,X) = (t,x)$ y $\beta_j(t)$ son funciones suaves desconocidas de t para $j = 0, 1, \ldots, p$

Una muestra aleatoria es modelada por

$$Y_i = q_\tau(T_i, X_i) + \varepsilon_i, \qquad 1 \leq i \leq n \tag{5.46}$$

siendo ε_i una variable aleatoria con τ cuantil igual a 0 e independiente de (T_i, X_i).

5.7.3. Lasso Adaptativo

La selección de variables a través de la verosimilitud penalizada (*penalized likelihood*) se ha utilizado en muchos trabajos recientes y dentro de éstos en la regresión cuantílica. En

esta metodología se destaca Lasso (Tibshirani, 1996), SCAD (Fan and Li, 2001), LARS (Efron et al., 2004) y Adaptative Lasso (Zou, 2006). El primer uso de la regularización en la regresión cuantílica la encontramos en el artículo de Koenker and Mizera (2004) y posteriormente ha sido utilizado por autores como Wang et al. (2007), Li et al. (2010) y Alhamzawi et al. (2012).

5.7.4. Regresión Cuantílica Bayesiana

Yu and Moyeed (2001) introdujeron la regresión cuantílica bayesiana para datos independientes. En la regresión cuantílica bayesiana es necesario definir una específica función de distribución para el término de error para que la estructura bayesiana funcione debidamente.

La principal dificultad de los modelos cuantílicos bayesianos es que en la regresión cuantílica no se especifica una función de verosimilitud, lo cual es indispensable en la inferencia bayesiana.

Para resolver ese problema se han propuesto distintos métodos dentro de los que destacan: Yu and Moyeed (2001) y Benoit and Van den Poel (2012), quienes emplean la distribución asimétrica de Laplace (ALD), para implementar una pseudo función de verosimilitud; B Dunson et al. (2003) quienes construyen una función de verosimilitud utilizando un vector de cuantiles.

Yu and Moyeed (2001) fueron los primeros en utilizar una función de distribución Laplace para el término de error. En este caso, si se tiene el modelo

$$y_i = \boldsymbol{x}_i'\boldsymbol{\beta}_\tau + \varepsilon_{i\tau}, \qquad i = 1, \dots n \tag{5.47}$$

Asumiendo errores independientes e idénticamente distribuidos como una distribución de Laplace $\varepsilon_{i\tau} \sim ALD(0, \sigma^2, \tau)$ con densidad

$$f(\varepsilon_{i\tau}|\sigma^2) = \frac{\tau\,(1-\tau)}{\sigma^2} exp\left(-\rho_\tau\left(\varepsilon_{i\tau}, 0\right)\frac{|\varepsilon_{i\tau}|}{\sigma^2}\right) \tag{5.48}$$

La distribución asumida para el término de error induce a que la variable respuesta $y_i|\boldsymbol{\beta}_\tau$, $\sigma^2 \sim ALD\left(\boldsymbol{x}_i'\boldsymbol{\beta}_\tau, \sigma^2, \tau\right)$ siendo su función de densidad

$$f(y_i|\beta_\tau, \sigma^2) = \frac{\tau\,(1-\tau)}{\sigma^2} exp\left(-\rho_\tau\left(y_i, \boldsymbol{x}_i'\boldsymbol{\beta}_\tau\right)\frac{|y_i - \boldsymbol{x}_i'\boldsymbol{\beta}_\tau|}{\sigma^2}\right) \tag{5.49}$$

Entonces, maximizar la función de verosimilitud

$$L\left(\boldsymbol{\beta}_\tau|\boldsymbol{y},\sigma_2\right) \propto \prod_{i=1}^{n} f(y_i|\beta_\tau,\sigma^2) \propto exp\left(-\sum_{i=1}^{n}\rho_\tau\left(y_i,\boldsymbol{x}_i'\boldsymbol{\beta}\right)\frac{|y_i-\boldsymbol{x}_i'\boldsymbol{\beta}_\tau|}{\sigma_2}\right) \tag{5.50}$$

con respecto a $\boldsymbol{\beta}_\tau$ para $\sigma_2 = 1$ conllevaría a que este problema de maximización sea equivalente al problema de minimización para la regresión cuantílica lineal planteado en la Ecuación 5.20.

Kottas and Gelfand (2001) utilizan el *Dirichlet Process Mixture Models* para estimar la distribución de probabilidades de los errores en un modelo de regresión semiparamétrica bayesiano para estimar la media y luego extendiendo este procedimiento a la regresión cuantílica en Kottas, Athanasios and Krnjajic, Milovan (2009) y Yang et al. (2012) se considera la verosimilitud empírica bayesiana. Waldmann et al. (2013) extienden el modelo geoaditivo sobre-dimensionado a la regresión cuantílica para tener una visión más detallada de la distribución cuantílica condicional.

Tokdar et al. (2012) introducen la estructura semi paramétrica bayesiana para el análisis simultáneo en modelos de regresión cuantílica lineal; Wang et al. (2016) introduce la regresión cuantílica bayesiana en los modelos de ecuaciones estructurales (SEM)[4], anteriormente B Dunson et al. (2003) y Burgette and Reiter (2012) utilizaron la regresión cuantílica en los modelos con respuesta latente bayesiana. Das and Ghosal (2017) utilizan los métodos bayesianos para la regresión cuantílica simultánea utilizando bases B-splines.

5.7.5. Regresión Cuantílica a través de Variables Instrumentales

La estimación utilizando variables instrumentales para modelar y estimar *Quantile Treatment Structural Effects* (QTE) en presencia de endogeneidad [5] fue desarrollado inicialmente por Chernozhukov and Hansen (2006, 2008).

Desde el ámbito no paramétrico se pueden mencionar los trabajos de Chernozhukov et al. (2007), Horowitz and Lee (2007) y Gagliardini and Scaillet (2012). Alternativas al método de variables instrumentales en la regresión cuantílica aparecen en los trabajos de Abadie et al. (2002) donde se utilizan variables binarias endógenas con instrumentos binarios; Imbens and Newey (2009) investigan la identificación y la estimación de modelos

[4] Aquí los modelos de ecuaciones estructurales no se entienden como los conocidos modelos de ecuaciones simultáneas sino como un modelo con variables latentes.

[5] La endogeneidad ocurre cuando una o algunas covariables está correlada con el término de error.

de ecuaciones simultáneas triangulares, no-paramétricos y no-separables; Chesher (2003, 2005, 2007) provee resultados de identificación importantes para modelos no-separables utilizando *Conditional Quantile Restrictions* (CQR).

Ma and Koenker (2006) y Lee (2007) aplican los resultados de identificación de Chesher (2003) a modelos paramétricos no-separables y desarrollan una función de control para modelos paramétricos, Lee (2007) extienden el enfoque de la función de control para modelos de regresión cuantílica de una manera semiparamétrica.

CAPÍTULO 6

La regresión cuantílica lineal y el estudio del salario de los inmigrantes cubanos en Estados Unidos

El presente Capítulo, en el que se utiliza la regresión cuantílica lineal, es fruto de dos artículos:

- El primer artículo: **Aleida Cobas-Valdés, Javier Fernández-Macho y Ana Fernández-Sainz (2016).** ***What determines the earnings distribution of Cuban immigrants in the United States? A conditional quantile regression analysis***, se encuentra en proceso de revisión en la revista ***Applied Economics Letters*** y en el mismo se utiliza la regresión cuantílica lineal con la novedad de recuperar el uso del centercepto (Wainer, 2000).

- El segundo artículo: **Aleida Cobas-Valdés, Javier Fernández-Macho y Ana Fernández-Sainz (2016).** ***Earnings Distribution of Cuban Immigrants in the USA: Evidence from Quantile Regression with Sample Selection***, publicado online en la revista ***Applied Economics***, **pp.1-16**; donde se estima un modelo de regresión cuantílica lineal teniendo en cuenta el problema de la selección muestral (Buchinsky, 1998a, 2002).

6.1. What determines the earnings distribution of Cuban immigrants in the United States? A conditional quantile regression analysis

A considerable number of studies have examined the economic assimilation of immigrants in host countries, considering a range of different socioeconomic characteristics related to the country specific human capital (Borjas (2015a, 1985), Hunt (2012), Elliott and Lindley (2008), Chiswick et al. (2008) and Cortes (2004)). In particular, studies have analyzed the effect of language skills on earnings (Adsera and Pytlikova (2015), Di Paolo and Tansel (2015), and Chiswick and Miller (2015, 2002), among others), and the returns to education on earnings (Brunner and Pate (2016), Li and Sweetman (2014), and Khan (1997), among others). Some of the aforementioned studies have found a positive correlation between the socioeconomic characteristics analyzed and earnings, while others have found the opposite.

In the USA, Cuban immigrants have officially been classified as refugees and therefore given a special status. The Cuban Adjustment Act (CAA) of 1996 provides for the granting of permanent residence to them and their spouses and children and other privileges to ease their adaptation (Eckstein, 2009). On the other hand, Cuban immigrants have positively self-selected in their migration decision to move the USA in terms of educational level, that is, they are people with more years of education than the mean among people who have decided not to leave Cuba (Cobas and Fernández, 2014). On average, Cuban immigrants aged 25-64 years have 13 years of education, while Mexicans and Salvadorans, the largest Hispanic groups in the USA, have 9 years of education (US Census Bureau, 2015).

The number of Cubans who have arrived in the USA has suffered a significant rise since Presidents Barack Obama and Raul Castro announced the renewal of diplomatic relations. According to US Custom and Border Protection (2016) 46.635 Cubans have entered the USA via ports of entry during the first 10 months of fiscal year 2016. In full fiscal year 2015, 43.159 Cubans arrived in the USA in contrast with year 2011 when 7.759 Cubans entered.

With all these premises, it is undoubtedly of interest to analyze the whole distribution of earnings of Cuban immigrants in the USA and quantify the effect of socioeconomic variables at different points on the earnings distribution. We use quantile regression (QR) in an effort to go beyond traditional ordinary least squares (OLS) analysis to describe the assimilation process of Cubans in the USA.

OLS enables the effect of covariates on the mean of the response variable to be calculated, while QR allows us to describe the conditional distribution of earnings on the covariates at different points of the distribution, and hence offers a more comprehensive overview of the link between the response variable and the covariates. In our study, the covariates considered are: years of education, age at time of migration, potential job experience, gender, marital status, ethnicity, citizenship status and proficiency in English.

With this work, we have made some notable findings. We show that the return to education on earnings for Cuban immigrants is less than would be expected and there are differences in the impact of the variables included between higher and lower income workers.

The paper is structured as follows: Section 2 describes the methodology used; Section 3 describes the data and presents the results of our estimation; and Section 4 summarizes our conclusions.

6.1.1. Methodology

We want to model the relationship between a response variable y and a set of regressors or covariates $[\boldsymbol{x}] = \{x_1, \ldots, x_k\}$ through a linear regression model

$$w_i = \boldsymbol{x}_i^{'}\boldsymbol{\beta} + \varepsilon_i, \qquad i = 1, \ldots, n \tag{6.1}$$

where w_i is the logarithm of gross hourly earnings for individual i, using total pre-tax wage and salary income (expressed in contemporary dollars), i.e., money received as an employee for the previous year as the measure of earnings; $\boldsymbol{x_i}$ is a covariate vector of socioeconomic characteristics of Cuban immigrants in the USA including an intercept; ε_i is the error term; and $\boldsymbol{\beta} \in \mathbb{R}^p$ is a vector of unknown parameters.

In a classical linear model, we usually assume that the error term is independent and has zero mean and constant variance. As we use repeated cross-sectional data in this work, there is likely to be heteroskedasticity. Given this, we assume that error term ε_i is independent but potentially heteroskedastic and we use robust OLS estimation on Equation 6.1 with the covariance matrix (HC4) estimator proposed by Cribari-Neto (2004), recommended mainly when there are influential observations in the data (Zeileis, 2004).

Next, we estimate the model proposed in Equation 6.1 under the conditional QR estimation procedure (Koenker and Bassett, 1978). For any $\tau \in (0, 1)$, a linear QR model can be written as

$$w_i = \boldsymbol{x}_i^{'}\boldsymbol{\beta}_\tau + \varepsilon_{\tau i} \tag{6.2}$$

There are no additional assumptions (apart from independence) that we need to make for the error term, and therefore, QR can be applied in situations of heteroskedasticity. The quantile function $Q_{w_i}(\tau|\boldsymbol{x_i})$ of the response variable w_i conditional on covariate vector $\boldsymbol{x_i}$ at a given quantile parameter τ is given by

$$Q_{w_i}(\tau|\boldsymbol{x_i}) = \boldsymbol{x_i}'\boldsymbol{\beta_\tau} \tag{6.3}$$

Davino et al. (2014) said that "QR is to classical regression what quantiles are to mean in terms of describing locations of a distribution".

Using the centercept concept (Wainer, 2000), all continuous covariates are centered at their median value, which has a more meaningful economic interpretation than the usual intercept because it provides the value of the (log) hourly earnings for the median individual.

6.1.2. Data and Empirical Results

We use repeated cross-section between 2000-2007 from the random sample of 1 % of the 2011 Population and Housing Census in the USA provided by Integrated Public Use Microdata Series (IPUMS) (Ruggles et al., 2011a). This sample includes 19079 individuals aged 25-50 years old at the time of the census and who were 17-49 years old at time of migration.

Table 6.4a summarizes the variables considered in the models. We can observe that 43 % of the people in the sample are women, 53 % are proficient in English[1] and 3 % are black. Notably, according to this data, half of Cubans in the USA have 12 or more years of education and they have 12.46 years of education on average. Figure 6.1 shows a histogram and box plot of (log) hourly earnings for Cuban immigrants. The mean of (log) hourly earnings is 2.6219 ($13.762), higher than the federal minimum wage ($7.25 per hour in 2011)[2].

[1]Individuals were classified as having proficiency in English if they indicated that they spoke English well or very well on the ACS questionnaire

[2]United States Department of Labor, 2011

Variable	*Description*	*Mean*	*Standard Deviation*
Woman	Dummy variable = 1 if woman	0.4264	0.4946
Black	Dummy variable = 1 if black	0.0309	0.1730
Married	Dummy variable = 1 if married	0.6283	0.4833
American citizen	Dummy variable = 1 if an American citizen	0.4461	0.4971
English proficiency	Dummy variable = 1 if proficiency in English	0.5314	0.4990
Age at migration	Age of the individual in years at time of migration	29.8100	8.2119
Years of education	Years of education of the individual	12.4600	2.9781
Experience	Potential experience of the individual	27.4300	11.5426
(log) Hourly earnings	Response variable	2.6219	0.7724

Note: Potential Experience = Age-Years of Education-6.

Tabla 6.1: *Descriptive statistics.*

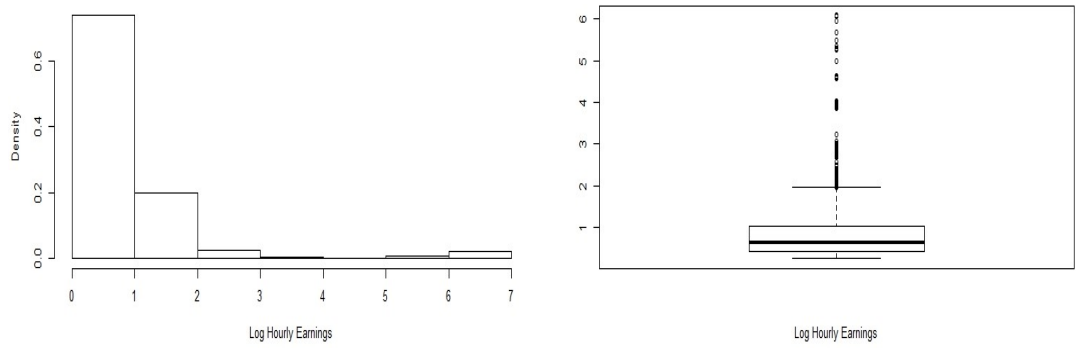


Figura 6.1: *Histogram and box plot of (log) hourly earnings..*

There are significant correlations between Proficiency in English and citizenship status ($\rho = 0.3387$), that is, Cubans with greater Proficiency in English are more likely to be American citizens. Age on arrival in the USA and Proficiency in English are negatively correlated ($\rho = -0.3611$); this implies that younger individuals are more predisposed to learn English. Those with higher levels of education appear to have less professional experience ($\rho = -0.2925$) but to be more willing to learn English ($\rho = 0.2916$).

Table 6.2 reports OLS estimation results with robust standard errors and QR estimation for five values of quantiles (τ): 0.10, 0.25, 0.50, 0.75 and 0.90. As can be seen in the first

column of this table, in the OLS estimation, only one covariate (Black) has a non-significant effect on hourly earnings, the other covariates all being statistically significant.

The estimate of the return to years of education on the mean earnings is 6.23 %, which is lower than expected. One additional year of potential job experience is shown to increase mean earnings by 0.46 %. Being a woman decreases mean earnings by 28.39 %, this variable having the strongest negative impact on hourly earnings. Overall, the model explains approximately 15 % of the variance in log hourly earnings.

In particular, note that (log) hourly earnings is positively related to being married, being a naturalized American, speaking English well or very well, years of education and potential job experience, while the response variable is negatively associated with being a woman and age on entry to the USA.

Figure 6.2 shows the QR results and Table 6.2 presents the QR estimations. These results indicate that the influence of the different socioeconomic characteristics studied on earnings tends to vary across the earnings distribution[3]. The exception is being married which produces a homogeneous effect across the different quantiles of the distribution.

In contrast, being black is not a relevant variable in explaining the variability in (log) hourly earnings from the center to the upper half of the distribution; however, for people who earn less, it is a significant predictor and is associated with lower earnings. Being a woman is associated with lower wages in all quantiles of the distribution, and the greatest negative impact is observed from the 25th to the 75th percentile. Being an American citizen has relatively little effect on hourly earnings.

[3]We have tested whether these coefficients differ significantly from one point in the distribution to another and the null hypothesis of equality of coefficients is rejected at the 5 % significance level.

Dependent Variable: (log) Hourly Earnings							
Variable	OLS	10 %	25 %	50 %	75 %	90 %
Centercept	2.1266	1.3763	1.6915	2.0707	2.4683	2.9478
	(0.0134)***	(0.0180)***	(0.0143)***	(0.0142)***	(0.0184)***	(0.0298)***
Woman	−0.2839	−0.2626	−0.3128	−0.3256	−0.2889	−0.2355
	(0.0108)***	(0.0141)***	(0.0109)***	(0.0109)***	(0.0144)***	(0.0230)***
Black	−0.0377	−0.0646	−0.0992	−0.0344	−0.0310	0.0946
	(0.0323)	(0.0266)**	(0.0344)**	(0.0354)	(0.0326)	(0.0824)
Married	0.0654	0.0709	0.0710	0.0683	0.0688	0.0547
	(0.0108)***	(0.0143)***	(0.0114)***	(0.0109)***	(0.0145)***	(0.0226)**
American citizen	0.0884	0.1029	0.1111	0.0898	0.0871	0.0024
	(0.0133)***	(0.0194)***	(0.0141)***	(0.0135)***	(0.0171)***	(0.0288)
Proficiency in English	0.1564	0.0992	0.1514	0.1577	0.1432	0.1824
	(0.0118)***	(0.0162)***	(0.0123)***	(0.0121)***	(0.0157)***	(0.0260)***
Age at migration	−0.0110	−0.0087	−0.0113	−0.0121	−0.0114	−0.0017
	(0.0008)***	(0.0011)***	(0.0008)***	(0.0008)***	(0.0010)***	(0.0017)***
Years of education	0.0623	0.0425	0.0525	0.0653	0.0727	0.0751
	(0.0023)***	(0.0030)***	(0.0023)***	(0.0024)***	(0.0029)***	(0.0049)***
Experience	0.0046	0.0009	0.0038	0.0053	0.0054	0.0075
	(0.0006)***	(0.0009)	(0.0006)***	(0.0006)***	(0.0008)***	(0.0013)***
Years of education squared	0.0051	0.0029	0.0041	0.0054	0.0063	0.0064
	(0.0003)***	(0.0004)***	(0.0003)***	(0.0004)***	(0.0004)***	(0.0008)***
Experience squared	−0.0003	−0.0003	−0.0003	−0.0003	−0.0003	−0.0002
	(0.00004)***	(0.00005)***	(0.00004)***	(0.00004)***	(0.00006)***	(0.00009)**

Note: Standard errors are given in parentheses."*", "**", and "***" denote significance at the 10 %, 5 % and 1 % levels, respectively.

Tabla 6.2: *OLS and Linear Quantile Regression.*

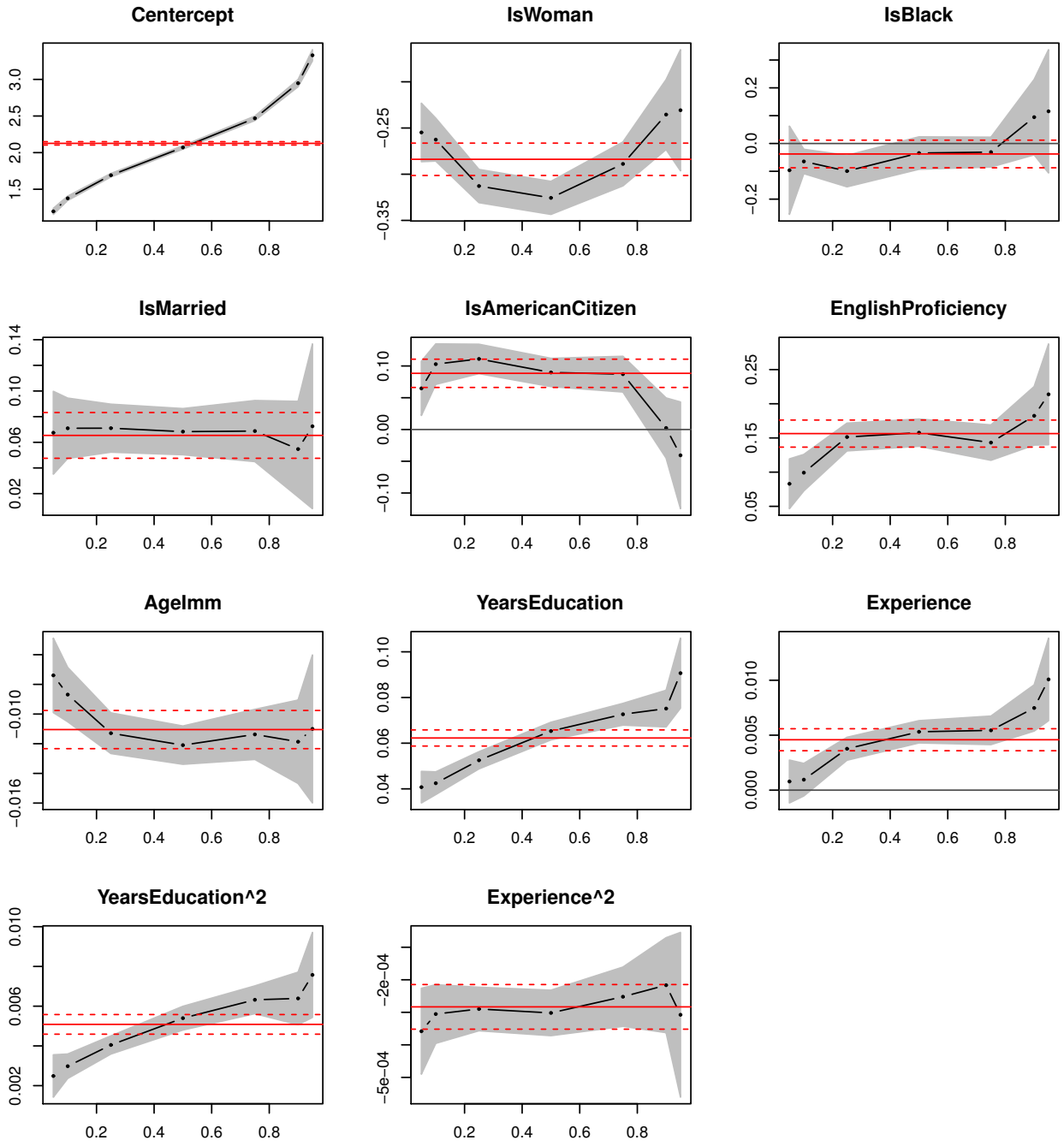


Figura 6.2: *Quantile Regression Estimation..*

The return to proficiency in English is higher for people who earn more: for the 90th percentile, speaking English well or very well is associated with about 18 % higher earnings compared to around 10 % higher for the 10th percentile. Borjas (2015a) has shown that

more recent immigrants to the USA are improving their English language skills more slowly than earlier immigrants. Regarding age at time of migration, being younger is more important in the upper quantiles of the distribution of earnings. The effect is negative for all quantiles and highest (in absolute terms) for the upper half of the distribution.

The age at which an individual migrates to the USA is a potentially important determinant of how that immigrant will eventually do in the labor market (Friedberg, 1992). The return to potential experience is almost zero across all the distribution of earnings of workers, but it is greater for people who earn more.

The number of years of education has a smaller impact on earnings in the lower quantiles (4.25 % for 10th percentile and 5.25 % for 25th percentile) than in the upper quantiles (7.27 % for the 75th percentile and 7.51 % for the 90th percentile). With respect to the median value of the earnings distribution, one additional year of education is associated with hourly earnings 6.53 % higher than the median value of the earnings distribution. That is, our data indicate that the effects of years of schooling are quite different across the different levels of hourly earnings. Specifically, the returns to education are higher at the top of the conditional earnings distribution but lower than expected.

Martins and Pereira (2004) found that returns to education for male workers in the USA in 1995 were 3.9 % for the 10th percentile and 7.9 % for the 90th percentile. This means that the returns to education for Cuban workers in the 2000s are similar to those for workers in the USA 16 years ago. One possible explanation for this situation is related to overeducation. Workers are overeducated if the skills that they can bring to their jobs exceed the skills needed (Groot and Van Den Brink, 2000). Overeducation may have become a problem if the types of job available to immigrants in the USA, particularly to skilled workers, are not be as good as the jobs available in their own country, even though the pay may be better. Another explanation would be that Cuban workers are not as highly valued in the market in the USA now as they were some years ago, due to increases in the emigration of highly skilled people. Machado and Mata (2005) suggest that when skilled workers become relatively more abundant, their relative wages decrease.

6.1.3. Conclusions

In this paper, we analyze the distribution of earnings of Cuban immigrants in the USA in terms of certain observable characteristics. For the analysis, we used Ordinary Least Squares (OLS) and Quantile Regression (QR), a technique which allows to characterize the whole distribution of earnings of Cuban immigrants in the USA.

When we use OLS estimation, all explanatory variables except the fact of being black prove to be significant at the 5 % significance level. All other variables being constant, speaking English well or very well is the variable that produces the biggest increase in mean hourly earnings at about 15.64 %; being a woman produces a decrease in mean hourly earnings of 28.39 % and being married produces an increase of about 6.54 %; being a U.S. citizen produces an increase of 8.84 %; one more year of study increases expected hourly earnings by 6.23 % and one year of potential experience increases expected hourly earnings by 0.46 %.

However, the application of Quantile Regressions shows how the influence of the different variables considered on hourly earnings varies across the earnings distribution. With this type of estimation method, differences can be detected between highly-skilled and low-skilled Cuban immigrants in the labor market.

The main conclusions of this article are the following: being a woman decreases hourly earnings at all points of the distribution, with the decrease being greater for individuals in the central part of the earnings distribution. The return to proficiency in English is greater for those people who earn more and the returns to education has a smaller impact on earnings in the lower quantiles and a greater impact at the top of the conditional earnings distribution but even then it is lower than expected.

6.2. Earnings Distribution of Cuban Immigrants in the USA: Evidence from Quantile Regression with Sample Selection

In terms of educational level Cuban immigrants have positively self-selected in their migration decision to move to the USA in the sense that people with the highest levels of education tend to migrate (Cobas and Fernández, 2014). In countries with low returns to skill and low wage dispersion there will be positive selection of immigrants. Those with above-average skill levels, and specifically more years of education than the workers' average in the source country, will have the greatest incentive to migrate (Borjas, 1987).

Moreover, Cuban immigrants in the USA have been under the protection of the Cuban Adjustment Act (CAA) that became law on November 2, 1966, entitling all Cuban immigrants to resident status after one year on US soil and to citizenship after five years of USA resident status.

Since it is the most highly skilled people who migrate, it is undoubtedly of interest to describe the whole distribution of earnings of people in the host country and to explore differences between different periods comparing the role of socioeconomic characteristics across all quantiles of this earnings distribution (Cattaneo, 2007; Chiswick, 2000).

The USA has been the main destination for migrants from Cuba and other Latin American countries over the last century. The US Census reveals that 54 million people in the USA, 17 % of the entire population, are of Hispanic origin (US Census Bureau, 2015), and it is estimated that the Hispanic population will reach nearly 130 million in 2060, constituting approximately 31 % of the population (Colby and Ortman, 2015). Among Hispanics, Mexicans ranked as the largest group in 2013 at 64.1 %, followed by Puerto Ricans (9.5 %), Cubans (3.7 %) and Salvadorans (3.7 %) (Brown and López, 2013). Besides, Cubans in the USA represent almost 18 % of the total Cuban population in mid-2015.

Along time, the socioeconomic characteristics of Cuban immigrants have been evolving. Until the 1980s, Cubans included a large proportion of middle-aged and elderly persons with a relatively high socioeconomic status (Pérez, 1986). However, for those who entered between 1980 and 1990, the median age is 28 years. This decade is marked by the immigrants third big wave in importance from Cuba in 1980–1981, known as the Mariel boat-lift period when some 125,000 Cubans migrated, including those from lower socioeconomic classes. Moreover, the 1990 Census reports that nearly 93 % of people who arrived to the

USA this way have less than 16 years of education, while in the 1990s, 23 % of Cubans had a Bachelor's degree or higher compared to 28 % of Non-Hispanic and white people in the USA. In 2000s, 25 % of Cubans held a bachelor's degree compared with 18 % of Central and South Americans, the median age of Cubans was 40 years, more than ten years older than the next eldest who were Central and South Americans, averaging 30 years of age (Wasem, 2009). In 2014, 21 % of Cuban immigrants of 25 years of age and over had a Bachelor's degree or higher, compared to 34 % of the total immigrant population and 38 % of native-born persons (US Bureau of Labor Statistics, 2015).

It is also an important aspect to compare the influence of socioeconomic characteristics on earnings of Cuban immigrants with those of two other important groups in the USA such as Indian and Chinese immigrants. India followed closely by China are the two largest immigrant groups of non-Hispanic origin, accounting for about 5 % each. They also have the highest percentage of adults with a bachelor's degree or higher since 1980s. In 2015, a majority of Asians of 25 years of age and older had at least a bachelor's degree (54 %) (Ryan and Bauman, 2016).

To analyze the distribution of earnings of Cuban immigrants working in the USA and to quantify the effect of socioeconomic variables, such as years of education, potential work experience, etc., at different points of the earnings distribution, we use quantile regression. This type of regression analysis (Koenker and Bassett, 1978) enables the conditional distribution of a response variable to be described as a function of the explanatory variables at different points in the distribution, and thus offers an overview of the link between the response variable and the explanatory variables.

In the analysis of earnings, however, we can only observe wages when individuals are at work. This problem is known as the sample selection problem, that is, the variables of interest are only observed for a non-random subsample of the population. That means that individuals may have selected themselves into the sample based on observed and unobserved characteristics. That is, there are unobservable skills that affect both the decision to work and the earnings.

With respect to immigrants, we have a double selection problem into (i) the people who decided to migrate to the USA and into (ii) the people who, once in the USA, decided to work and hence have observable earnings. In general, the people who decided to migrate were people that anticipated that their wages in the host country would exceed their wages in the source country plus all the costs associated with the migration process (Cattaneo, 2007).

In this article, we address only the second selection problem, because we do not have

data on people's earnings in Cuba, and hence we estimate a quantile regression model based on the sample selection model of Buchinsky (1998a, 2002) for Cuban individuals who decided to migrate to the USA. The explanatory variables considered are some of the most important in the relevant literature: years of education, age on arrival into the USA, potential work experience, gender, marital status, ethnicity, citizenship status and proficiency in English.

Upon arrival in the USA, immigrants can be expected to be at an earnings disadvantage relative to natives because they lack certain skills and information possessed by natives (Friedberg, 1992). Over time, they may increase their income when they improve their English and adapt to the specific sets of labor skills needed in the host country.

Some recent research into the earnings distribution of Latino immigrants in the USA reveal considerable disadvantages with respect to native-born people. It has been found that immigrants from non-English speaking countries have average hourly earnings of around 12 % less than those of native born workers in the USA (Chiswick et al., 2008). A similar pattern of non-English-speaking immigrants earning less than their native counterparts has been found in the UK (Hunt, 2012). Other authors have concluded that differences in human capital endowment and socioeconomic characteristics account for part of the lack of income assimilation (Elliott and Lindley, 2008, Brunner and Pate, 2016).

The contribution of this article is twofold. First, we use sample selection correction for quantile regression proposed by Buchinsky (1998a, 2002) to estimate the effects of different socioeconomic characteristics on the conditional probability distribution of the earnings of Cuban immigrants in the USA. Second, we test the conditional independence assumption necessary to obtain consistent estimates of those effects using the recent tests proposed by Huber and Melly (2015). To our knowledge, this is the first time that quantile regression has been used in sample selection migration research.

The paper is structured as follows: Section 6.2.1 outlines the methodology used; Section 6.2.2 describes the estimated model; Sections 6.2.3 and 6.2.4 describe the data and presents the results of our estimation, and Section 6.2.5 sets out the main conclusions.

6.2.1. Methodology

Many of the issues that social researchers are currently analyzing are related to the values of variables of interest located at the tails of the distribution. For example, in studies of inequality in wages, income, health and social skills, an important part of current research

cannot be statistically treated in an appropriate way using the classical ordinary least squares (OLS) regression which refers only to the analysis of the conditional mean as pointed out in Martins and Pereira (2004) and Haupt et al. (2014).

Moreover, in data sets describing economic phenomena, the assumptions of no covariance and constant variance are sometimes unreasonable. In relation to this, if the assumptions fail to hold, OLS estimators are not efficient relative to other estimators. On the other hand, OLS estimators may be seriously deficient in linear models with non-normal errors. The extreme sensitivity of OLS estimators to modest amounts of outlier contamination makes them very poor estimators in many non-Gaussian, especially long-tailed, situations (Koenker and Bassett, 1978).

As mentioned in Huber and Melly (2015), paraphrasing Mincer (1974)'s seminal paper on human capital earnings, the significant heterogeneity found in labor economics applications can be dealt with a quantile regression approach. Quantile regression allows more detailed analysis and it is possible to consider the lower and the upper tails of the distribution and to identify, at each quantile, the most appropriate policy to achieve the chosen target (Buchinsky, 1998b). Several parametric estimation of the conditional quantile functions have been used since Koenker and Bassett (1978) in absence of selection. Buchinsky (1998a, 2002) was the first author to consider the problem of estimating quantile regression in the presence of sample selection, extending the series estimator of Newey (1991) for the mean to the quantiles. Buchinsky assumes that wages are a separable and additive function of observable and unobservable characteristics and assume independence between the error term and the covariates conditional on the selection probability.

The Linear Quantile Regression Model

Although most analyses of social sciences have used OLS regression methods it should be recognized that the resulting estimates of various effects, on the conditional mean, are not necessarily indicative of the size and nature of those effects on the lower or upper extremes of the distribution.

A linear quantile regression model (Koenker and Bassett, 1978) links the conditional quantiles of the dependent variable to the covariates linearly. Quantile regression enables the whole conditional distribution of the response variable to be estimated. With this method, it is possible to study the conditional distribution of the response variable over the covariates at different points and thus provide an overview of the links between the response variable and the covariates selected for the study.

The τ-quantiles of a distribution are a set of descriptors alternative to moments. A τ-quantile is defined as the data value such that a proportion τ of the data is less than or equal to this data value while $(1-\tau)$ is greater than it, that is, it is the value y_τ such that $P(Y \leq y_\tau) = \tau$. Therefore, for a continuous random variable Y, quantiles are based on the cumulative distribution function (CDF):

$$\tau = F_Y\big(y_\tau\big) = P\big(Y \leq y_\tau\big) = \int_{-\infty}^{y_\tau} f\big(u\big)du, \qquad \text{for} \quad \tau \in \big(0,1\big). \tag{6.4}$$

For any such $\tau \in (0,1)$ a linear quantile regression model can be written as

$$y_i = \boldsymbol{x_i}'\boldsymbol{\beta_\tau} + \varepsilon_{\tau i}, \tag{6.5}$$

where y_i, $\boldsymbol{x_i} = \big(1,\ x_{i1}, x_{i2}, \ldots x_{ip}\big)'$ denote respectively the observation of the response variable and a vector of given covariates corresponding to individual i, and $\boldsymbol{\beta_\tau} = \big(\beta_{\tau 0}, \beta_{\tau 1}, \ldots \beta_{\tau p}\big)'$ are quantile-specific linear effects.

No specific assumptions for the error term ε_τ are made, except that $\varepsilon_{\tau i}$ and $\varepsilon_{\tau j}$ are independent for $i \neq j$, and that the distribution function at 0 is equal to τ:

$$\int_{-\infty}^{0} f_{\varepsilon_{\tau i}}\big(\varepsilon_{\tau i}\big)d\varepsilon_{\tau i} = F_{\varepsilon_{\tau i}}\big(0\big) = \tau. \tag{6.6}$$

This assumption implies that $F^{-1}_{\varepsilon_{\tau i}}\big(\tau\big) = 0$ and hence the quantile function $Q_Y\big(\tau|\boldsymbol{x_i}\big)$ of the response variable conditional on the covariate vector $\boldsymbol{x_i}$ at a given quantile parameter τ is given by

$$Q_Y\big(\tau|\boldsymbol{x_i}\big) = \boldsymbol{x_i}'\boldsymbol{\beta_\tau}. \tag{6.7}$$

The linear conditional quantile function in Eq. (6.7) can be estimated as in Koenker and Bassett (1978) by solving

$$\hat{\beta}_\tau = \underset{\beta_\tau \in \mathbb{R}^{p+1}}{\operatorname{argmin}} \sum_{i=1}^{n} \rho_\tau\big(y_i - \boldsymbol{x_i'}\boldsymbol{\beta_\tau}\big), \tag{6.8}$$

where $\rho_\tau\big(\,\cdot\,\big)$ denotes the "check function"

$$\rho_\tau\big(u\big) = \begin{cases} u \cdot \tau & \text{if} \quad u \geq 0, \\ u \cdot \big(1-\tau\big) & \text{if} \quad u < 0. \end{cases} \tag{6.9}$$

Since the derivative of the check function does not exist at its minimum, two main computational strategies could be used: linear programming techniques, formulating the minimization problem in Eq. (6.8) as a set of linear constraints (Koenker, 2005) or smoothing the cusp of the check function to allow the use of computational techniques that rely on differentiability (Nychka et al., 1995; Horowitz, 1998; Oh et al., 2011).

The estimation of regression quantile coefficients β_τ can vary across τ and hence the marginal effect of a particular explanatory variable may not be homogeneous across different quantiles. The $\hat{\beta}_\tau$ coefficients are rates of change conditional on adjusting for the effects of the other variables in the model, but are now defined for a specified quantile (Koenker, 2005).

In general, the parameter estimates in linear quantile regression models have the same interpretation as those in any other linear regression model, that is $\hat{\beta}_\tau$ can be interpreted as the change in the conditional quantile function when the corresponding covariate changes, assuming all other covariates remain constant. For instance, Machado and Mata (2005) interpreted the estimated quantile regression coefficients as rates of return on labor market skills at different points of the conditional wage distribution.

The Sample Selection Quantile Regression Model

A sample selection problem arises whenever the outcome of interest is only observable for some sample that is non-randomly selected even if it is conditional on observed factors. Gronau (1974) and Heckman (1974) were the first to analyze this problem. The last of these authors used a parametric selection model and, in particular, he introduced a system of simultaneous equations where the error terms of the selection and outcome equations are assumed to follow a bivariate normal distribution and the non-zero correlation indicates non-random sample selection (Heckman, 1974, 1976, 1979).

Several authors have analyzed the self-selection of immigrants. From a theoretical point of view Chiswick (2000) and Cattaneo (2007) conclude that those who migrate are the most motivated, skilled and ambitious. Moreover, Borjas (1987, 1991) obtains that self-selection not only depends on unobserved individual characteristics, such as motivation, skills or to have access to financial resources, but also on observed characteristics such as education. Using an alternative theoretical framework, Grogger and Hanson (2011) analyze the selection of migrants where individuals maximize linear utility and migration is driven by absolute earnings differences between high and low-skilled workers.

This type of model has been applied in several different contexts. For example, in a

historical context there are studies such as Abramitzky and Braggion (2006), Abramitzky et al. (2012) and Biavaschi and Elsner (2013). For more recent periods, the selection of migrants across developed countries appears in works such as Pirttilä (2004), Kleven et al. (2014) and Borjas et al. (2015). On the other hand, several authors have analyzed the immigrant self-selection focusing mainly on migration flows from poor countries (Yashiv, 2008; Borjas et al., 2015) while, in particular, for the typical migration of Mexicans to the USA there is an extensive literature using self-selection such as Durand et al. (2001), Chiquiar and Hanson (2005), Ibarraran and Lubotsky (2007), McKenzie and Rapoport (2010) and Kaestner and Malamud (2014).

Various different generalizations of the classical sample selection model have been considered in the literature but, in general, authors have considered parametric sample selection models where the model is estimated using the two-step procedure proposed by Heckman (1979).

A semi-parametric sample selection process where only the linear predictor is parametrically pre-specified has been considered (Ahn and Powell, 1993; Newey, 2009). Blundell et al. (2007) used the method suggested by Manski and Sims (1994) to correct for sample selection using restrictions that impose positive selection into work where certain variables are assumed to affect the labor force participation equation but not the distribution of wages.

Das et al. (2003) proposed a more flexible treatment of this problem in a non-parametric context where the linear regression with the Heckman's selection correction is replaced with series expansions. Non-parametric methods have also been used by other authors (Lee, 2009; Chen and Zhou, 2010; Biavaschi and Elsner, 2013).

Bayesian methods are proposed by Chib et al. (2009) and Wiesenfarth and Kneib (2010) using a simultaneous equation system that incorporates a Bayesian version of penalized smoothing splines. Marra and Radice (2013) used a similar model to Wiesenfarth and Kneib (2010) based on the penalized maximum likelihood estimation framework. Many authors have used specific copula functions in the outcome and selection equations (*e.g.* Hasebe and Vijverberg, 2012; Schwiebert, 2013; Pigini, 2015).

In the quantile regression approach, Buchinsky (1998a, 2002) was the first author to consider the problem of estimating quantile regression in the presence of sample selection, extending the estimator of Newey (1991) for the mean to the quantiles. In a first step, he estimated the single (linear) index selection equation using the semi-parametric procedure of Ichimura (1993), and in a second step, incorporated a non-parametric method to correct sample selection in quantile regression using a similar idea to that of Heckman (1979) and

Newey (1991) for mean regression. He then used a semi-parametric model similar to that employed for the mean by other authors (Ahn and Powell, 1993; Newey, 2009).

Chernozhukov and Hansen (2006) described the instrumental variable quantile regression to evaluate the impact of endogenous variables on the entire distribution of economic outcomes when the variables are self-selected or selected in relation to potential outcomes. Albrecht et al. (2009) extend the decomposition developed by Machado and Mata (2005) to account for selection analyzing the gender log wage gap for full-time workers.

Arellano and Bonhomme (2017) proposed to correct sample selection in quantile regression via the cumulative distribution function or copula of the percentile error term in the outcome equation and in the selection equation; in this way, they readjusted to correct the percentiles for selection. Picchio and Mussida (2011) adapt the hazard function estimator proposed by Donald et al. (2000) to panel data framework with sample selection correction.

The paper of Buchinsky (1998a) was a great contribution to sample selection studies, but he assumed conditional independence between the error terms and the regressors given the selection probability and this implies that all quantile regression curves are parallel and the quantile slope coefficients are identical to the mean slope coefficients as mentioned by Huber and Melly (2015). These last authors implemented an independence test based on the sample selection correction proposed by Buchinsky (1998a, 2002) for quantile regression.

6.2.2. The estimated model

We first estimate the traditional outcome equation (Mincer, 1974) for Cuban immigrants in the USA considering the sample selection quantile regression model proposed by Buchinsky (1998a, 2002). After that, we check the assumption of independence using the test proposed by Huber and Melly (2015).

Lemieux (2006) mentioned that Mincer's canonical earnings model implies that wage dispersion is higher for people with more work experience and for those with more years of education. This can be analyzed with a quantile regression approach correctly specified and not forgetting the sample selection problem due to its ability to capture heterogeneous effects.

Buchinsky (1998a) pointed out that in models where the sample bias is due to sample selection into the labor force, the bias is of an unknown form, and therefore, it cannot be corrected using the traditional parametric correction as in Heckman (1974, 1976, 1979).

Let y_i be the (log) gross hourly earnings for individual i using the total pretax wage and salary income (expressed in contemporary dollars) and $\boldsymbol{x_{2i}}$ a vector of socioeconomic characteristics. Then, we may define the wage outcome equation in terms of a latent variable y_i^* which depends linearly on the vector of characteristics $\boldsymbol{x_{2i}}$:

$$y_i^* = \boldsymbol{x_{2i}'\beta_0} + u_i, \tag{6.10}$$

where $\boldsymbol{\beta_0}$ denotes the vector of slope coefficients and the intercept is incorporated in the error term as it is not relevant for testing later the conditional independence (Huber and Melly, 2015). However, instead of the latent y_i^*, we observe y_i for people who are actually working, that is

$$y_i = D \cdot y_i^* = D \cdot \big(\boldsymbol{x_{2i}'\beta_0} + u_i\big), \tag{6.11a}$$

where D is determined by a latent single-index crossing model such that

$$D \equiv I\big(\boldsymbol{x_{1i}'\alpha_0} + v_i \geq 0\big) = \begin{cases} 1 & \text{if} \quad \boldsymbol{x_{1i}'\alpha_0} + v_i \geq 0, \\ 0 & \text{if} \quad \boldsymbol{x_{1i}'\alpha_0} + v_i < 0, \end{cases} \tag{6.11b}$$

with $I(\cdot)$ as the usual indicator function and $\big(\boldsymbol{x_{1i}'\alpha_0} + v_i\big)$ as the selection rule where $\boldsymbol{\alpha_0}$ will be estimated using the semi-parametric procedure suggested by Klein and Spady (1993). In order to allow identification of the model parameters, we assume that $\boldsymbol{x_{1i}}$ is a superset of $\boldsymbol{x_{2i}}$ including at least one continuous variable which is not included in $\boldsymbol{x_{2i}}$ and is a significant variable in the selection equation.

The outcome and the selection equations are assumed to be linear in the regressors as in Heckman (1979) for mean regression. The difference with respect to his model is that we consider the joint distribution of errors to be unknown. In both equations (selection and outcome) the errors are assumed to be independent of the covariates, conditional on the selection probability.

It is also assumed that the joint distribution of the error terms $\boldsymbol{(u, v)}$ is absolutely continuous with respect to the Lebesgue measure and independent of $\boldsymbol{x_1}$ conditional on the latent selection index

$$f_{\boldsymbol{u,v}}\big(\cdot\,|\boldsymbol{x_1}\big) = f_{\boldsymbol{u,v}}\big(\cdot\,|\boldsymbol{x_1}'\alpha_0\big). \tag{6.12}$$

Equation (6.12) implies the assumption of conditional independence between $\boldsymbol{x_1}$ and $\big(\boldsymbol{u, v}\big)$ in the absence of sample selection and is equivalent to

$$f_{\boldsymbol{u,v}}\big(\cdot\,|\boldsymbol{x_1}\big) = f_{\boldsymbol{u,v}}\Big(\cdot\,|Pr\big(D = 1|\boldsymbol{x_1}\big)\Big). \tag{6.13}$$

Rewriting Eq. (6.10) considering quantile regression, we have

$$y_i^* = \boldsymbol{x'_{2i}}\boldsymbol{\beta_\tau} + u_{\tau i}, \qquad 0 \leq \tau \leq 1, \tag{6.14}$$

and then we have to estimate the following quantile regression model

$$Q_y\big(\tau \mid \boldsymbol{x_1}, D = 1\big) \equiv \boldsymbol{x'_2}\boldsymbol{\beta_0} + h_\tau\big(\boldsymbol{x'_1}\boldsymbol{\alpha_0}\big) \tag{6.15}$$

where $h_\tau\big(\boldsymbol{x'_1}\boldsymbol{\alpha_0}\big) \equiv Q_u\big(\tau \mid \boldsymbol{x'_1}\boldsymbol{\alpha_0},\, D = 1\big)$.

These results imply that $\boldsymbol{\beta_0}$ can be estimated by the τth quantile regression of y on $\boldsymbol{x_2}$ and a nonparametric function of $\boldsymbol{x'_1}\hat{\boldsymbol{\alpha}}$ in the selected sample (Buchinsky, 1998a).

Huber and Melly (2015) propose testing the assumption of conditional independence between the error terms and the vector of covariates. The testing approach is based on the Kolmogorov-Smirnov and Cramér-von-Mises statistics applied to the whole conditional quantile process after addressing the sample selection problem with the Buchinsky (1998a, 2002) model.[4] The testing problem is given by

$$\begin{aligned} &H_0 : \boldsymbol{\beta_\tau} = \boldsymbol{\beta_{0.5}}, \quad \forall \tau \in \mathcal{T}, \\ &H_1 : \boldsymbol{\beta_\tau} \neq \boldsymbol{\beta_{0.5}}, \quad \text{for some } \tau \in \mathcal{T}, \end{aligned} \tag{6.16}$$

where $\mathcal{T}$ is a closed subset of $\big[e, 1-e\big]$, $0 < e < 1$, and β_τ denotes the true τ quantile regression coefficient defined as

$$\boldsymbol{\beta_\tau} = \underset{\beta}{\operatorname{argmin}}\, E\Big[\rho_\tau\Big(y_i - \boldsymbol{x'_{2i}}\boldsymbol{\beta} - h_\tau\big(\boldsymbol{x'_{1i}}\boldsymbol{\alpha_0}\big)\Big)\Big]. \tag{6.17}$$

6.2.3. The data

The data are drawn from a random sample of 1 % of the American Community Survey (ACS) database provided by the Integrated Public Use Microdata Series (IPUMS) (Ruggles et al., 2013). We restrict our sample to individuals who were between 25 and 55 years old, worked 60 hours or less weekly during the year preceding the census and entered the USA when they were between 17 and 49 years of age. This last criteria corresponds to the age group most likely to migrate for economic reasons (Bertoli et al., 2013) and we sought to exclude people who completed their education in the USA (Lowell et al., 2008).

[4] See Huber and Melly (2015, Section 3.2 and 3.3) for details.

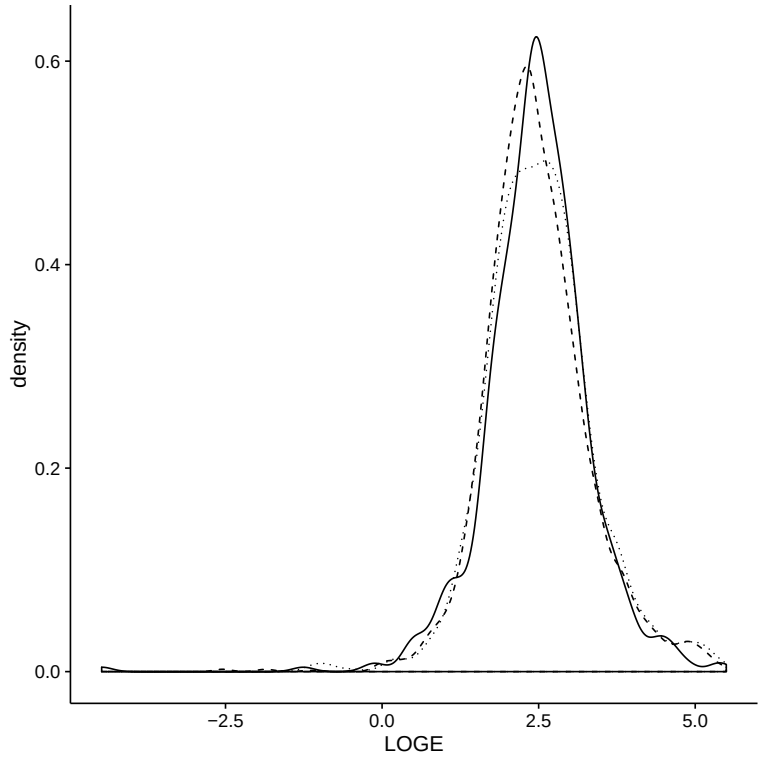
In order to estimate Eq. (6.15), we use repeated cross-sections between 2000–2007 dividing the sample in three cohorts: people who migrated to the USA arriving in the 1980s, the 1990s and the 2000s; implementing Buchinsky's correction for quantile regression (Buchinsky, 1998a, 2002) and testing the assumption of independence as in Huber and Melly (2015).

The vector of regressors $\boldsymbol{x_2}$ contains indicators for being a black person, being a citizen of the USA, proficiency in English, years of education, years of education squared, potential experience[5] and potential experience squared. The vector $\boldsymbol{x_1}$ contains the same variables as $\boldsymbol{x_2}$ plus an indicator for being a woman or being married, an interaction variable that reflects the individual being married and a woman, and age at time of migration.

Figure 6.3 shows the density of (log) hourly earnings for the different cohorts in the sample. The (log) hourly earnings distributions for the cohorts corresponding to the most recent decades, the 1990s and 2000s, are skewed compared to the (log) hourly earnings distribution of the 1980s cohort and there is a slight decrease in individuals in the right tail of the distribution corresponding to the 2000s, which means that there is a decrease in the number of persons who earn the highest amounts.

Summary descriptive statistics of the variables included in the study are listed in Table 6.4a. We can observe that Cubans who came to the USA in the 1980s have a higher mean salary per hour that can be explained because they are residents in the USA for more than 20 years. In contrast, the median value of the hourly earnings has remained between 11 and 12 dollars across the three cohorts considered. Overall, the 25 % of individuals who earned the least in the three cohorts studied obtained hourly earnings above the federal minimum wage, which was \$5.15 in the period of the repeated cross-section data (US Bureau of Labor Statistics, 2015).

[5] Potential experience has been calculated as: Age − Years of Education − 6.



Note: The dotted, dashed and solid lines corresponds to the 1980s, 1990s and 2000s, respectively.

Figura 6.3: *Kernel Density estimates for (log) hourly earnings.*

(a) *Cubans.*

	Mean			1st Quartile			Median			3rd Quartile		
	1980*s*	1990*s*	2000*s*	1980*s*	1990*s*	2000*s*	1980*s*	1990*s*	2000*s*	1980*s*	1990*s*	2000*s*
Hourly earnings	18.836	17.549	16.803	7.192	7.075	7.693	12.037	10.909	11.765	19.886	18.519	18.750
Years of education	11.35	12.65	13.34	9.00	12.00	12.00	12.00	12.00	12.00	13.00	14.00	16.00
Experience	28.18	19.31	17.73	23.00	13.00	13.00	29.00	18.00	17.00	34.00	25.00	22.00
Age at migration	27.01	32.01	34.04	22.00	26.00	29.00	27.00	31.00	34.00	31.00	37.00	38.00
Is black	0.036	0.039	0.035									
Is an American citizen	0.465	0.166	0.039									
Has proficiency in English	0.555	0.392	0.335									
Is a woman	0.433	0.447	0.497									
Is married	0.584	0.596	0.567									

(b) *Indians and Chinese (2000s).*

	Mean		1st Quartile		Median		3rd Quartile	
	Indians	Chinese	Indians	Chinese	Indians	Chinese	Indians	Chinese
Hourly earnings	19.814	14.206	8.354	5.769	17.752	9.615	25.888	17.382
Years of education	15.77	14.88	16.00	13.00	16.00	16.00	17.00	17.00
Experience	11.11	14.62	5.00	7.00	9.00	12.00	14.00	21.00
Age at migration	29.50	32.09	25.00	26.00	28.00	31.00	33.00	38.00
Is an American citizen	0.053	0.062						
Has proficiency in English	0.928	0.697						
Is a woman	0.349	0.512						
Is married	0.691	0.638						

Tabla 6.3: *Descriptive Statistics of relevant variables for Cubans, Indians and Chinese in the USA.*

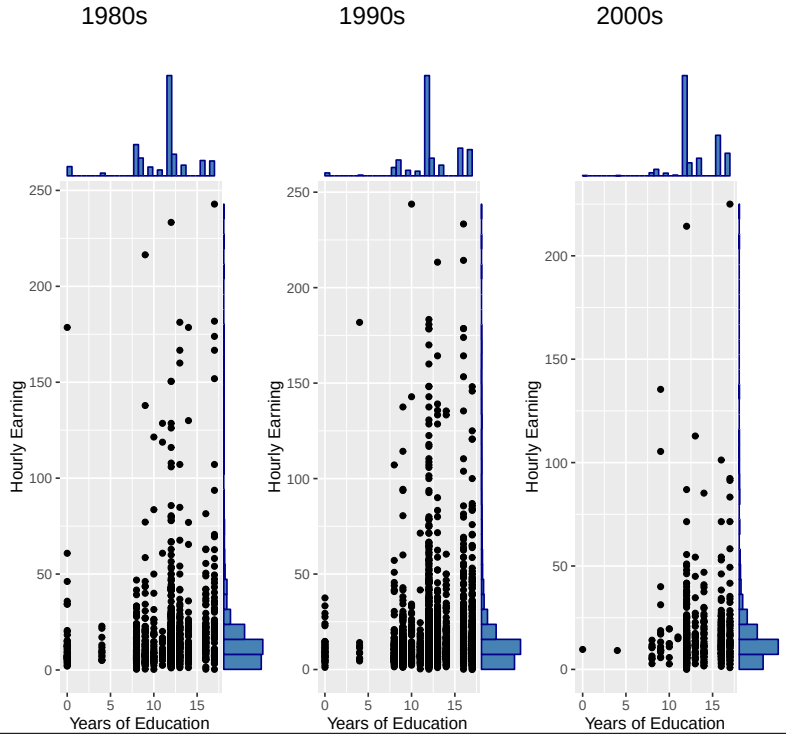


Figura 6.4: *Marginal Histogram Plot.*

As can be observed, workers who arrived to the USA in the 1980s are, on average, much more experienced than the rest. The potential experience has markedly decreased, and this can be explained, in part, by an increase in the years of education. The statistics highlight the small proportion of individuals of black race who have migrated from Cuba to the USA, a number that has remained steady across the three cohorts. The number of Cubans who are naturalized citizens is reduced among those who entered the United States in the 1990s to less than half of those who arrived in the 1980s from 47 % to 17 %. By contrast, among those who arrived in the 2000s, only 4 % have US citizenship. Also, a marked characteristic of the cohort corresponding to the 2000s is the large number of women. The percentage of married people has remained almost constant, on average, across all samples, while the age at the time of migration has been increasing over the years. The tendency to be proficient in English has decreased considerably, which is consistent with Borjas (2015a), where it was found that, without taking sample selection bias into account, English language proficiency is significantly lower for larger national origin groups such as Cuban immigrants.

Figure 6.4 shows the scatter plot between years of education and hourly earnings as well as the marginal distribution of each variable. In all cohorts considered, 50 % of Cuban

immigrants have 12 years of education or more. We can also observe that as the year of the cohort increases the values get more concentrated in the lower hourly earnings and the higher years of education.

Table 6.4b reports the descriptive statistics for Indians and Chinese in USA who arrived in 2000s. We note that the average earnings of Indian immigrants are significantly higher than those of Cubans and Chinese. However, their experience is the lowest among the three groups. On the other hand, the average years of education and knowledge of the English language of Cubans is markedly lower than in the other groups. Finally, the percentage of men and women in the case of Cubans and Chinese is around 50 %, while in the case of Indians the ratio is 65 %–35 % in favor of men, although no substantial differences are appreciated in their married rate (Fairlie et al., 2010).

The Spearman's rank correlation matrices for all variables of interest appear in Table 6.6 in the Appendix. We note that for all ethnic groups and cohorts hourly earnings are significantly and positively correlated with years of education and proficiency in English while negatively correlated with being a woman. In the 2000s cohort, significant correlations between hourly earnings and American citizenship or marital status are also found for Indian and Chinese immigrants while for Cubans these correlations are not significant although they were in previous cohorts. Among the explanatory variables, the most significant correlations, in all ethnic groups and cohorts, were observed between potential experience and age at time of migration followed by the correlation between years of education and proficiency in English, positive in both cases. Finally, there is a negative significant correlation between years of education and experience again in all ethnic groups and cohorts.

6.2.4. Results

In this section results from the estimation of Eq. (6.15) are presented. Since the main goals of our article are to describe the distribution of (log) hourly earnings and to study trends in socioeconomic characteristics over time in the history of migration from Cuba to the USA, a quantile regression model corrected for sample selection serves our purposes in that it allows us to observe the contribution of all variables included across the distribution of (log) hourly earnings.

In a first step, we have estimated the selection equation using the quasi-maximum likelihood estimator of Klein and Spady (1993). Results from this estimation (see Table 6.7 in the Appendix) suggest that the included variables are important factors for the participation

decision in the labor market. For the 1980s cohort, gender is the only significant variable with a negative impact on the participation probability, however it became a nonrelevant variable later on in the 1990s and 2000s. In fact, for the 2000s cohort, all the covariates are significant at the 5 % level except gender and marital status (the latter only significant at the 10 % level). Being an American citizen, being black and being proficient in English seem to be the most important variables in the participation probability. All variables except experience have the expected signs.

In a second step, we have estimated the outcome equation including only individuals who were working at the time of the census. The estimated coefficients are provided in the Appendix (Table 6.7). In our estimated model the socioeconomic characteristics included are statistically significant for some percentiles but not for others in all cohorts studied. Figure 6.5 in the Appendix shows the estimated quantile coefficients corrected for sample selection for four covariates together with the corresponding 95 % confidence intervals.

In a final third step we test the conditional independence assumption for the cohorts studied over the range of quantiles $\tau = \left[0.05, 0.95\right]$. The second row of Table 6.5 lists the p-values of Huber and Melly (2015)'s tests for the validity of the model for the whole quantile regression vector and, in the rest, the p-values associated to each of the explanatory variables separately. If the null hypothesis was not rejected this would imply that independence between the error term and the covariates in any of the quantiles of the distribution holds, so that the coefficients are consistently estimated by Buchinsky's method (Buchinsky, 1998a, 2002). As can be seen in Table 6.5 for Cuban immigrants the p-values imply the non rejection of the null hypothesis and, therefore, the estimated coefficients from the output equation in the previous step (Table 6.7) are consistent and can be interpreted in terms of influences of the corresponding covariates on the hourly earnings of Cuban immigrants across the different cohorts.

It can be noted (Fig. 6.5) that the conditional quantiles change differently across the three cohorts considered. For Cubans that arrived in the 1980s, proficiency in English is the only statistically relevant variable across a wide range of percentiles with a positive effect except for the highest hourly earnings. However, the effects of this variable become nonrelevant for more recent immigrants.

For the 1990s cohort, being an American citizen appears as a new relevant factor for explaining hourly earnings with a positive effect on approximately the lower to median values of the distribution. This variable becomes very relevant for the most recent arrivals with an important negative effect on the upper tail of the distribution only. On the other hand, the impact of experience on earnings becomes statistically significant in

the third quartile with a positive effect that disappears for the later arrivals. Being a black immigrant in the 1990s has a negative effect on wages for people at the bottom of the earnings distribution. However, this effect becomes most pronounced for the Cuban workers who entered the USA in the 2000s. In fact the curve follows an inverted U-shape meaning that the effect of this variable on hourly earnings is similar at the extremes of the distribution.

The returns on education show little influence for workers who migrated in the 1980s and 1990s and an ascending significant pattern in the first quartile of the earnings distribution for workers who migrated in the 2000s.

For Indian and Chinese immigrants the p-values in Table 6.5 imply the rejection of the null hypothesis of conditional independence between errors and covariates and, therefore, their estimated coefficients from the output equation are inconsistent and Buchisnky sample selection model is not valid (Huber and Melly, 2015). Consequently, no valid inference can be drawn and therefore no meaningful comparison can be made with the Cubans' results.

	Kolmogorov-Smirnov				
	Cubans			Indians	Chinese
	1980s	1990s	2000s	2000s	1990s
All variables	0.621	0.789	0.562	0.000	0.024
Years of education	0.974	0.595	0.788	0.005	0.818
Years of education squared	0.997	0.784	0.929	0.002	0.903
Experience	0.919	0.022	0.198	0.015	0.014
Experience squared	0.941	0.020	0.180	0.010	0.024
Is black	0.474	0.462	0.933	...	...
Is an American citizen	0.973	0.924	0.495	0.232	0.829
English Proficiency	0.147	0.922	0.688	0.046	0.568
	Cramér - Von Mises				
	Cubans			Indians	Chinese
	1980s	1990s	2000s	2000s	1990s
All variables	0.710	0.892	0.825	0.000	0.010
Years of education	0.967	0.778	0.863	0.000	0.842
Years of education squared	1.000	0.903	0.879	0.001	0.868
Experience	0.765	0.207	0.170	0.002	0.011
Experience squared	0.735	0.170	0.190	0.083	0.023
Is black	0.318	0.368	0.826	...	...
Is an American citizen	0.803	0.896	0.687	0.129	0.792
English Proficiency	0.090	0.926	0.710	0.043	0.412

Tabla 6.5: *Independence test results.*

6.2.5. Conclusions

This paper contributes to the migration literature with the use of a quantile regression model with sample selection adapted from the work of Buchinsky (1998a, 2002). It also considers the test proposed by Huber and Melly (2015) to test the independence between error terms and regressors conditional on the selection probability, without which Buchinsky's method would not be consistent.

This model allows us to analyze the hourly earnings of immigrants in the USA taking into account a range of observable characteristics, namely, years of education, potential work experience, ethnicity, citizenship status and proficiency in English. The coefficients obtained along the different quantiles of the earnings distribution have been compared,

both within specific cohorts as well as between the different cohorts and ethnic groups considered.

The results show that the hypothesis of conditional independence is not rejected for Cuban immigrants and increments in their earnings associated with the usual socioeconomic characteristics in labor studies vary across the cohorts analyzed. However, this hypothesis is rejected for Indian and Chinese immigrants, so that no valid inference can be drawn and therefore no meaningful comparison can be made with the Cubans' results.

For Cuban immigrants, two findings are particularly important. First, the decline in returns from education for recent cohorts may be a sign that a high level of education no longer provides a competitive advantage, since Cubans with higher levels of education are those who migrate and once in the USA not all work in jobs commensurate with their level of education. Second, being a black person is associated with significantly lower earnings across the working population, regardless of the individuals' position in the earnings distribution. This may indicate that black Cubans lack an economic incentive to emigrate and may explain the fact that, historically, very few black Cubans have made the decision to emigrate to the USA.

	Hourly earnings	Years of education	Experience	Black	American citizen	English Proficiency	Woman	Married	Age at migration
Cubans 1980s									
Hourly earnings	1.00	0.22***	−0.12***	−0.05*	0.09***	0.19***	−0.13***	0.08***	−0.07**
Years of education		1.00	−0.52***	−0.05	0.34***	0.37***	0.10***	−0.01	−0.10***
Experience			1.00	0.05*	−0.15***	−0.33***	−0.12***	0.00	0.73***
Black				1.00	−0.06**	−0.05	−0.05	−0.10***	0.02
American citizen					1.00	0.25***	0.16***	0.02	−0.08***
English proficiency						1.00	−0.01	−0.02	−0.25***
Woman							1.00	0.01	−0.06**
Married								1.00	0.00
Age at migration									1.00
Cubans 1990s									
Hourly earnings	1.00	0.16***	−0.06***	−0.01	0.14***	0.18***	−0.14***	0.01	−0.07***
Years of education		1.00	−0.30***	−0.03	0.14***	0.26***	0.08***	0.06***	0.02
Experience			1.00	0.05***	0.05***	−0.23***	−0.03*	0.01	0.84***
Black				1.00	0.01	0.02	−0.01	−0.08***	0.04**
American citizen					1.00	0.29***	0.05***	0.01	−0.06***
English proficiency						1.00	−0.01	−0.07***	−0.26***
Woman							1.00	0.08***	0.00
Married								1.00	0.07***
Age at migration									1.00
Cubans 2000s									
Hourly earnings	1.00	0.18***	−0.06	−0.04	0.06	0.13***	−0.15***	0.04	−0.01
Years of education		1.00	−0.21***	0.00	0.08**	0.27***	0.06	−0.07*	0.09**
Experience			1.00	0.04	0.04	−0.11***	−0.03	0.17***	0.90***
Black				1.00	−0.04	0.01	0.03	0.07	0.07*
American citizen					1.00	0.20***	0.05	0.02	0.01
English proficiency						1.00	−0.05	−0.03	−0.10**
Woman							1.00	0.10**	−0.06
Married								1.00	0.15***
Age at migration									1.00
Indians 2000s									
Hourly earnings	1.00	0.28***	−0.08***		−0.06***	0.24***	−0.23***	0.12***	−0.10***
Years of education		1.00	−0.35***		−0.09***	0.34***	−0.07***	0.00	−0.19***
Experience			1.00		0.09***	−0.28***	−0.02	0.30***	0.88***
American citizen					1.00	−0.04***	0.06***	0.02*	0.01
English proficiency						1.00	−0.10***	−0.01	−0.23***
Woman							1.00	0.16***	−0.07***
Married								1.00	0.24***
Age at migration									1.00
Chinese 2000s									
Hourly earnings	1.00	0.38***	−0.06		0.01***	0.29***	−0.16***	0.07***	−0.01
Years of education		1.00	−0.53***		−0.13***	0.60***	−0.13***	0.00	−0.31***
Experience			1.00		0.09***	−0.45***	0.00	0.22***	0.91***
American citizen					1.00	−0.01	0.06***	0.02*	0.02
English proficiency						1.00	−0.03**	0.00	−0.35***
Woman							1.00	0.03**	−0.04***
Married								1.00	0.21***
Age at migration									1.00

Note: *, ** and *** indicate correlations significant at the 10 %, 5 % and 1 % significance levels respectively.

Tabla 6.6: *Spearman rank correlation matrix between variables of interest.*

	Outcome Equation (Buchinsky, 1998b): $\tau = 0.05$			$\tau = 0.50$			$\tau = 0.90$		
	1980*s*	1990*s*	2000*s*	1980*s*	1990*s*	2000*s*	1980*s*	1990*s*	2000*s*
Years of education	−0.062	0.051	−0.193***	0.007	−0.033	−0.032	0.087	−0.033	0.075
	(0.061)	(0.066)	(0.057)	(0.046)	(0.035)	(0.077)	(0.069)	(0.040)	(0.060)
Years of education squared	0.004	0.000	0.008**	0.003	0.002	0.004	0.000	0.002	0.001
	(0.003)	(0.003)	(0.003)	(0.002)	(0.002)	(0.003)	(0.003)	(0.002)	(0.003)
Experience	−0.024	−0.002	−0.014	−0.021	0.005	0.002	−0.061	0.041*	0.073**
	(0.050)	(0.018)	(0.057)	(0.034)	(0.013)	(0.030)	(0.081)	(0.022)	(0.033)
Experience squared	0.000	0.000	0.000	0.000	0.000	0.000	0.001	−0.001**	−0.002*
	(0.001)	(0.000)	(0.001)	(0.001)	(0.000)	(0.001)	(0.001)	(0.000)	(0.001)
Is black	−1.263	−0.679***	−1.295*	−0.225	−0.069	−0.405	0.273	−0.139	−0.678***
	(1.077)	(0.229)	(0.687)	(0.235)	(0.135)	(0.346)	(0.319)	(0.167)	(0.191)
Is an American citizen	−0.105	0.072	0.336	0.033	0.157**	−0.054	0.169	0.198	−0.806***
	(0.147)	(0.113)	(0.268)	(0.096)	(0.067)	(0.212)	(0.161)	(0.136)	(0.237)
Has proficiency in English	0.085	0.172**	0.357**	0.334***	0.097	0.016	0.330**	0.159	0.249
	(0.138)	(0.081)	(0.153)	(0.098)	(0.067)	(0.104)	(0.156)	(0.110)	(0.168)
	Selection equation Klein and Spady (1993):								
	1980*s*	1990*s*	2000*s*						
Years of Education	0.149	0.141	0.229***						
	(0.137)	(0.264)	(0.017)						
Years of Education squared	−0.002	0.028	−0.012***						
	(0.006)	(0.019)	(0.0007)						
Experience	0.115	0.467	−0.132***						
	(0.092)	(0.307)	(0.009)						
Experience squared	−0.002	−0.0006	0.005***						
	(0.001)	(0.002)	(0.0003)						
Is black	0.584	0.609	1.029***						
	(0.603)	(0.746)	(0.06)						
Is an American citizen	0.421	−0.097	1.363***						
	(0.269)	(0.377)	(0.078)						
English proficiency	0.398	1.791	0.331***						
	(0.264)	(1.137)	(0.025)						
Woman and married	1.000	1.000	1.000						
Is a woman	−0.519***	−0.908	−1.732						
	(0.206)	(0.414)	(0.052)						
Is married	−0.250	−1.794	−0.043*						
	(0.214)	(0.877)	(0.024)						
Age at migration	−0.046	−0.315	−0.028***						
	(0.032)	(0.199)	(0.005)						

Note: Standard errors in parentheses. *, ** and *** indicate coefficients significant at the 10 %, 5 % and 1 % significance levels respectively.

Tabla 6.7: *Hourly earnings distribution for Cuban immigrants in the USA.*

Outcome Equation (Buchinsky, 1998b):	$\tau = 0.05$		$\tau = 0.50$		$\tau = 0.90$	
	Indian	Chinese	Indian	Chinese	Indian	Chinese
Years of education	-0.256^{***}	-0.100^{***}	-0.163^{***}	-0.092^{***}	-0.139^{**}	-0.088
	(0.029)	(0.022)	(0.0186)	(0.019)	(0.058)	(0.037)
Years of education squared	0.012^{***}	0.007^{***}	0.009^{***}	0.008^{***}	0.008^{***}	0.007^{***}
	(0.001)	(0.001)	(0.001)	(0.001)	(0.002)	0.001
Experience	0.057^{***}	0.044^{***}	0.032^{***}	0.0719^{***}	0.044^{***}	0.048^{***}
	(0.0097)	(0.012)	(0.004)	(0.005)	(0.0067)	(0.008)
Experience squared	-0.0018^{***}	-0.0012^{***}	-0.0014^{***}	-0.0017^{***}	-0.0012^{***}	-0.0011
	(0.0009)	(0.0003)	(0.0001)	(0.0001)	(0.0002)	(0.0002)
Is an American citizen	-0.074	0.187^{**}	-0.230^{***}	0.216^{***}	-0.108^{**}	0.276^{***}
	(0.0726)	(0.090)	(0.041)	(0.052)	(0.0534)	(0.086)
Has proficiency in English	0.259^{***}	0.128	0.389^{***}	0.205^{***}	0.303	0.215^{***}
	(0.0719)	(0.079)	(0.045)	(0.033)	(0.085)	(0.054)
Selection equation Klein and Spady (1993):						
	Indian	Chinese				
Years of Education	-0.423^{***}	-0.367^{***}				
	(0.0023)	(0.0242)				
Years of Education squared	0.0005^{***}	-0.013^{***}				
	(0.00005)	(0.00081)				
Experience	-0.271^{***}	-0.573^{***}				
	(0.0013)	(0.0301)				
Experience squared	-0.0015^{***}	-0.002^{***}				
	(0.00001)	(0.00018)				
Is an American citizen	0.009^{***}	0.697^{***}				
	(0.0033)	(0.056)				
English proficiency	-0.032^{***}	0.206^{***}				
	(0.0027)	(0.0319)				
Woman and married	1.000	1.000				
Is a woman	0.473^{***}	0.562^{***}				
	(0.0057)	(0.066)				
Is married	0.267^{***}	-0.494^{***}				
	(0.0035)	(0.0248)				
Age at migration	0.286^{***}	0.605^{*}				
	(0.0013)	(0.0315)				

Note: Standard errors in parentheses. *, ** and *** indicate coefficients significant at the 10 %, 5 % and 1 % significance levels respectively.

Tabla 6.8: *Hourly earnings distribution for Indian/Chinese immigrants in the USA.*

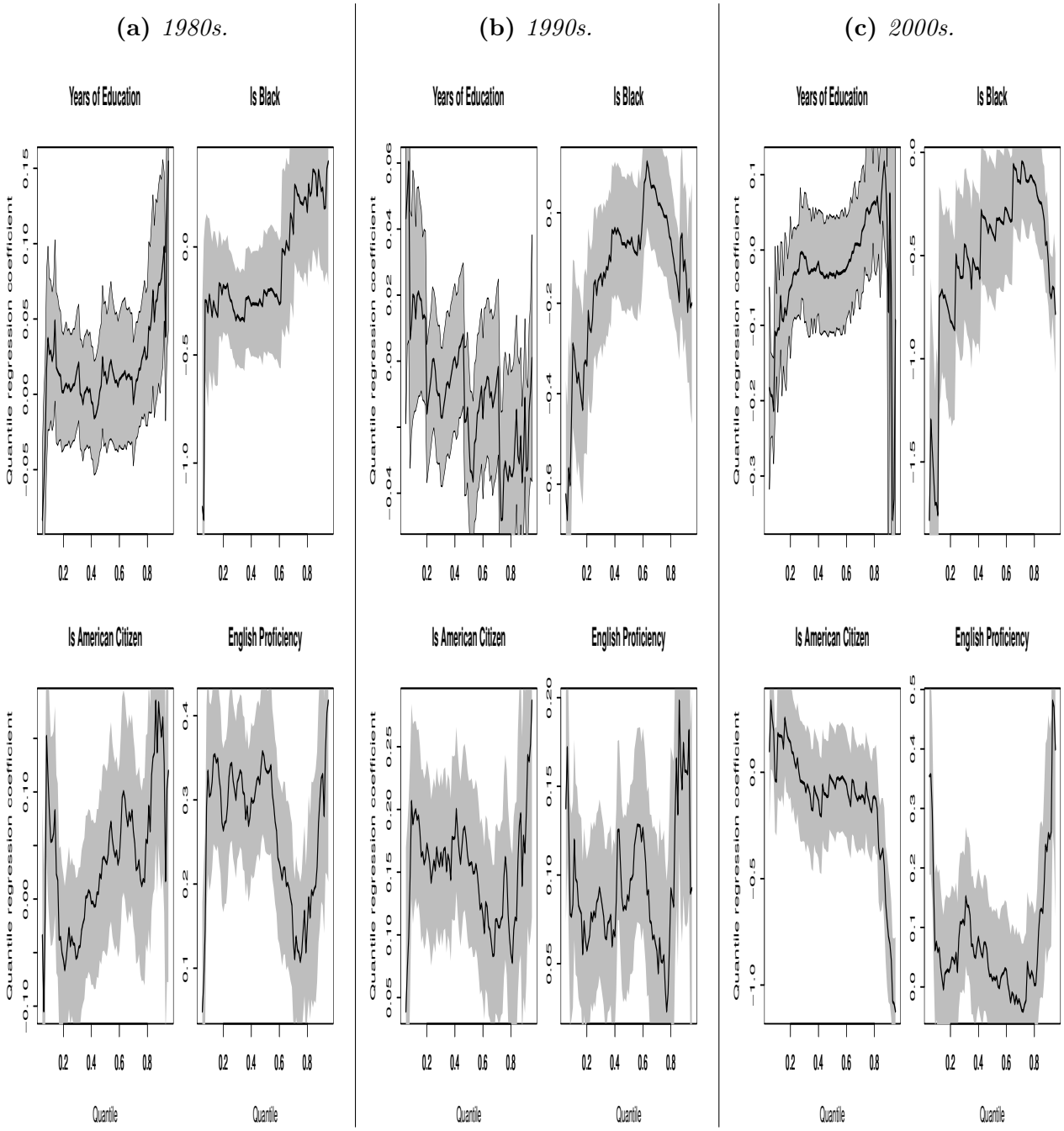


Figura 6.5: *Quantile Regression Estimations for Cubans.*

Parte IV

La regresión cuantílica flexible

CAPÍTULO 7

De los Modelos Aditivos Generalizados a los Modelos Cuantílicos Estructurados

Este Capítulo es una extensión del Capítulo 3 para dar continuidad a otros modelos de regresión con respecto a la media, con el fin de conectar con la regresión cuantílica aditiva estructurada. Para ello, en la Sección 7.1 hablaremos de los modelos aditivos generalizados (GAM) y en la Sección 7.2 de los modelos aditivos estructurados (STAR) para explicar en la Sección 7.3 los modelos cuantílicos aditivos estructurados (STAQ), definiendo cada tipo de modelo así como su forma de especificación y estimación.

7.1. Los Modelos Aditivos Generalizados (GAM)

Un Modelo Aditivo Generalizado (GAM) (Hastie and Tibshirani, 1990) es un GLM en el cual parte del predictor lineal η_i es especificado en términos de la suma de funciones suaves de los regresores sin tener que establecer el supuesto de que se conoce la forma paramétrica exacta de esas funciones ni de cuál es exactamente el grado de suavidad más apropiado para cada una de ellas.

Los modelos aditivos generalizados constituyen la más importante extensión de los GLM donde se ha incluido la aproximación paramétrica y no paramétrica en aras de aportar flexibilidad al efecto no lineal de las covariables.

$$g(\mu_i) = \eta_i = \alpha + \sum_{j=1}^{p} f_j(x_j) \tag{7.1}$$

Supongamos que tenemos las observaciones (y_i, x_i, v_i) desde $i = 1, \ldots, n$ donde y_i es el valor de la variable respuesta para la observación i, $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ip})^\intercal$ es un vector de regresores continuos y $\boldsymbol{v}_i = (v_{i1}, \ldots, v_{iq})$ es un vector de regresores categóricos.

Hastie and Tibshirani (1990) asumen que, dado (x_i, v_i):

$$\begin{aligned} g(\mu_i) &= f_1(x_{i1}) + \cdots + f_p(x_{ip}) + v_i'\gamma \\ \mu_i &= E(y_i) \end{aligned} \tag{7.2}$$

donde $f_1, \ldots, f_p$ son funciones suaves desconocidas de los regresores continuos y $v_i'\gamma$ representa la relación estrictamente lineal entre las variables (en su mayoría categóricas) y la variable respuesta. Para asegurar identificabilidad se asume que $E(f_j) = 0$ (Brezger and Lang, 2006).

La distribución de y_i pertenece a la familia exponencial, como en los GLM y la media $\mu_i = E(y_i|x_i, v_i)$ se relaciona con el predictor aditivo semiparamétrico η_i a través la Ecuación 7.2.

Existen numerosos trabajos encaminados a especificar y modelizar esas funciones desconocidas de los regresores continuos utilizando los denominados splines de suavizado (*smoothing splines*) (Hastie and Tibshirani, 1990; Green and Silverman, 1993); polinomios locales (Fan et al., 1996); splines de regresión (*regression splines*) (Friedman and Silverman, 1989; Friedmann, 1991; Stone et al., 1997; Eubank, 1999); los B-splines (De Boor, 1977; Dierckx, 1995) y los P-splines (Eilers and Marx, 1996; Marx and Eilers, 1998). También se utilizan modelos en los que los parámetros suavizados se intentan estimar conjuntamente con las funciones de regresión (Wand, 1999; Currie and Durbán, 2002).

Los splines de suavizado utilizan tantos parámetros como observaciones lo que tiene el inconveniente de que si el número de observaciones es muy elevado, será ineficiente su implementación. Los splines de regresión se pueden ajustar mediante mínimos cuadrados ordinarios una vez que se ha seleccionado el número de nodos pero para la selección del número de nodos es necesaria la utilización de complejos algoritmos. Los B-splines univariantes son construidos a partir de polinomios o más bien de trozos de polinomios.

Un B-spline de grado q consiste en $q+1$ piezas o trozos de polinomios de grado q que se han unido de forma suave a través de p_i puntos (llamados nodos internos) entre las fronteras p_{min} y p_{max} (llamados nodos externos). En esos puntos internos las derivadas hasta el orden $q-1$ son continuas y el B-spline es positivo en el dominio expandido por $q+2$ nodos y cero en el resto. Excepto en los extremos, se solapa con $2q$ trozos de polinomios de sus vecinos y para cada valor de x, $q+1$ B-splines son no nulos (Durbán et al., 2008).

En la elección de los nodos hay que tener en cuenta que si se eligen muchos nodos, se sobreestima el modelo y si se seleccionan pocos nodos, se infraestima el modelo. Friedman and Silverman (1989) y Stone et al. (1997) proponen métodos para optimizar el número de posiciones de los nodos. O'Sullivan (1988) propone seleccionar un número relativamente largo de nodos y fijar una penalización sobre la segunda derivada con lo que se impone una restricción a la flexibilidad.

Los P-splines utilizan una base para la regresión y lo que se hace es modificar la función de verosimilitud a través de la introducción de una penalización que se basa en las diferencias entre los coeficientes adyacentes. Los splines con penalizaciones combinan los splines de suavizado y los splines de regresión utilizando menos parámetros que los primeros y con la ventaja sobre los segundos de que la selección de los nodos no es tan determinante, el tamaño de la base utilizada es mucho menor que la dimensión de los datos y el número de nodos no es superior a 40, lo que hace más eficiente el proceso.

(Breiman and Friedman, 1985) ya plantearon la idea de un modelo aditivo:

$$\theta(y_i) = \alpha + \sum_{j=1}^{p} f_j(x_j) + \varepsilon \tag{7.3}$$

$E(\varepsilon) = 0$ y ε es independiente de las x_{js}. La transformación $\theta(y_i)$ de la variable respuesta $\boldsymbol{y}$ así como las funciones $f_j(x_j)\, s$ de los predictores son funciones arbitrarias suaves.

La principal razón para transformar la variable respuesta mediante la función $\theta(\boldsymbol{y})$ es que en algunas situaciones un modelo aditivo simple puede no ser apropiado para describir $E\{\boldsymbol{y}/x_1, \ldots, x_p\}$ y podría ser muy apropiado para describir $E\{\theta(\boldsymbol{y})/x_1, \ldots, x_p\}$. Por ejemplo, si $\boldsymbol{y} = exp(x_1 + x_2^2)\,\boldsymbol{\varepsilon}$, un modelo aditivo simple describiría $\log(\boldsymbol{y})$ pero no describiría $\boldsymbol{y}$.

Las funciones θ^* y $f_1^*, \ldots f_p^*$ serán tales que minimicen la SCR de la regresión de $\theta(\boldsymbol{y})$ sobre $\sum_{j=1}^{p} f_j(\boldsymbol{x}_j)$ definida como:

$$e^2\left(\theta, f_1, \ldots f_p\right) = \frac{E\left\{\left[\theta\left(\boldsymbol{y}\right) - \sum_{j=1}^{p} f_j\left(x_j\right)\right]^2\right\}}{Var\left[\theta\left(\boldsymbol{y}\right)\right]} \tag{7.4}$$

Para encontrar θ^*, $f_1^*, \ldots, f_p^*$ se define un simple proceso iterativo tal que asumiendo que $\boldsymbol{y}$, $x_1, x_2 \ldots x_p$ tienen distribuciones conocidas y que, sin pérdida de generalidad, $E[\theta^2\left(\boldsymbol{y}\right)] = 1$ y todas las funciones tienen valor esperado igual a cero. A este proceso se le conoce como el algoritmo ACE *(Alterning Conditional Expectation)*.

ACE fue el primero de una serie de artículos que Leo Breiman escribió sobre suavidad y modelos aditivos. Este algoritmo, que proporciona el primer método para estimar modelos aditivos, tuvo algunas dificultades, sobre todo para muestras pequeñas en las que los resultados no eran los esperados (Cutler, 2010).

Por ello, Breiman adaptó el método basado en splines usando *Stepwise Deletion of Knots* (Smith, 1982) y de ahí surgió el algoritmo mejorado que plantea en Breiman (1993). Una de las grandes aportaciones de este trabajo de Breiman ha sido que en él se pueden encontrar los primeros intentos de utilizar validación cruzada para medir la inestabilidad.

Supongamos que tenemos dos variables aleatorias $\boldsymbol{y}$ y $\boldsymbol{x}$ así como las transformaciones $\theta\left(\boldsymbol{y}\right)$ y $f\left(\boldsymbol{x}\right)$ de forma que $E\left\{\theta\left(\boldsymbol{y}\right)/\boldsymbol{x}\right\} \approx f\left(\boldsymbol{x}\right)$. El algoritmo ACE pretende minimizar la función de pérdida al cuadrado:

$$E\left\{\theta\left(y\right) - f\left(\boldsymbol{x}\right)\right\}^2 \tag{7.5}$$

de tal forma que para un valor θ establecido, la función f minimizadora es aquella tal que:

$$f\left(\boldsymbol{x}\right) = E\left\{\theta\left(\boldsymbol{y}\right)/\boldsymbol{x}\right\} \tag{7.6}$$

A su vez, para una función f establecida, el valor θ minimizador sería aquel tal que:

$$\theta\left(\boldsymbol{y}\right) = E\left\{f\left(\boldsymbol{x}\right)/\boldsymbol{y}\right\} \tag{7.7}$$

En la Figura 7.1 se resume la forma en la que funciona el algoritmo ACE cuando se cuenta con un único predictor o con varios predictores.

Algoritmo ACE para un único predictor

(i) Inicializamos: Sea $\theta(\boldsymbol{y}) = \{\boldsymbol{y} - E(\boldsymbol{y})\} / \{Var(\boldsymbol{y})\}^{1/2}$.

(ii) Computamos: $f(\boldsymbol{x}) = E\{\theta(\boldsymbol{y}) / \boldsymbol{x}\}$ para obtener una nueva función f.

(iii) Computamos: $\hat{\theta}(\boldsymbol{y}) = E\{f(\boldsymbol{x}) / \boldsymbol{y}\}$ y estandarizamos $\theta(\boldsymbol{y}) = \hat{\theta}(\boldsymbol{y}) / \left\{Var\left[\hat{\theta}(\boldsymbol{y})\right]\right\}^{1/2}$ para obtener un nuevo θ.

(iv) Iteramos: Se repiten los pasos **(ii)** y **(iii)** hasta que $E\{\theta(\boldsymbol{y}) - f(\boldsymbol{x})\}^2$ no cambie.

Algoritmo ACE para varios predictores

(i) Inicializamos: Sea $\theta(\boldsymbol{y}) = \{\boldsymbol{y} - E(\boldsymbol{y})\} / \{Var(\boldsymbol{y})\}^{1/2}$

(ii) Ajustamos: Ajustamos un modelo aditivo para $\theta(\boldsymbol{y})$ para obtener nuevas funciones $f_1, \ldots, f_p$

(iii) Computamos: $\hat{\theta}(\boldsymbol{y}) = E\left\{\sum_{j=1}^{p} f_j(\boldsymbol{x}) / \boldsymbol{y}\right\}$ y estandarizamos $\theta(\boldsymbol{y}) = \hat{\theta}(\boldsymbol{y}) / \left\{Var\left[\hat{\theta}(\boldsymbol{y})\right]\right\}^{1/2}$ para obtener un nuevo θ.

(iv) Iteramos: Se repiten los pasos **(ii)** y **(iii)** hasta que $E\left\{\theta(\boldsymbol{y}) - \sum_{j=1}^{p} f_j(\boldsymbol{x})\right\}^2$ no cambie.

Figura 7.1: *Algoritmo ACE .*

Cuando tenemos varios predictores, el único cambio que se hace en el procedimeinto anterior es en el paso (ii) donde ahora se ajustaría un modelo aditivo para $\theta(\boldsymbol{y})$ que nos permita obtener las nuevas funciones $f_1, \ldots, f_p$.

Así mismo, Engle et al. (1986) propusieron lo que se conoce como modelos parcialmente lineales, siendo modelos que se mueven entre los modelos lineales y los modelos no paramétricos debido a su flexibilidad puesto que permiten incluir variables con efectos lineales y no lineales sobre la variable respuesta.

Dadas las observaciones $(y_i, \boldsymbol{x_i}, \boldsymbol{z})$, $i = 1, \ldots, n$ donde y_i es la variable respuesta, $\boldsymbol{x_i} = (x_{i1}, \ldots, x_{ip})^{'}$ y $\boldsymbol{z_i} = (z_{i1}, \ldots, z_{ip})^{'}$ son los vectores de las covariables. El modelo parcialmente lineal asume que:

$$y_i = \beta_0 + \boldsymbol{x_i}^{'}\beta_i + f(\boldsymbol{z_i}) + \varepsilon_i \tag{7.8}$$

donde $\boldsymbol{\beta}$ es el vector de parámetros desconocidos, siendo β_0 el intercepto; f es una función desconocida, cuyo dominio de definición va desde $\mathbb{R}^q$ hasta $\mathbb{R}$ y ε_i es el término de error *iid* con media cero y varianza constante. El modelo más utilizado, en la práctica es el univariante, en el que asume que $q = 1$.

Para este caso, diferentes modelos se han especificado teniendo en cuenta funciones suaves. Heckman (1986) utiliza splines de suavizados; Speckman (1988) emplea kernels suavizados; Li and Liang (2008) se apoya en los polinomios locales; Zhang et al. (2012) emplea splines de suavizados en el contexto de los modelos ANOVA; Lou et al. (2015) introducen el *Sparse Partially Linear Additive Model* (SPLAM) y Cheng et al. (2015) utiliza splines de regresión penalizados.

Aplicaciones de estos modelos[1]encontramos en Schmalensee and Stoker (1999) para estudiar el consumo de gasolina en Estados Unidos, en Dinse and Lagakos (1983) para bio-ensayos, entre otros.

En este tipo de modelos es crucial la selección de las covariables que van a tener un efecto lineal y las que van a tener un efecto no lineal. La validez del modelo especificado así como la inferencia dependen de la estructura que se le de al modelo porque en estos modelos la estructura es muy cambiante al depender de múltiples funciones lineales o no (Zhang et al., 2012).

[1]Para profundizar en estos modelos, una buena referencia es (Härdle and Liang, 2007).

7.2. Los Modelos Aditivos Generalizados Estructurados (STAR)

Los Modelos Aditivos Generalizados Estructurados (STAR) fueron introducidos por Fahrmeir et al. (2004) y Brezger and Lang (2006), basados en los trabajos previos de Fahrmeir and Lang (2001a); Fahrmeir and Lang (2001b), Brezger and Lang (2003) y Lang and Brezger (2004).

Fahrmeir et al. (2004) extendieron los Modelos Aditivos Generalizados (GAM) de manera que el efecto no lineal de las covariables continuas así como el efecto suave de las tendencias en el tiempo fuesen modeladas a través de una versión bayesiana de los P-splines propuestos por Eilers and Marx (1996) y por Marx and Eilers (1998).

En un STAR pueden coincidir términos que recojan efectos lineales, efectos no lineales, espaciales, uno o más términos con coeficientes cambiantes e interacciones no lineales entre variables. Los modelos STAR permiten una mayor flexibilidad a la hora de intentar modelizar la relación no lineal entre los regresores y la variable respuesta, siendo los modelos GLM y GAM casos particulares (Umlauf et al., 2015).

La variable respuesta puede ser continua, categórica o de recuento dependiendo de la escala en la que se midan los datos. Dependiendo de los supuestos que se establezcan sobre su distribución, la variable respuesta puede ser Gaussiana, Binomial, Multinomial, Poisson o Gamma.

Los modelos STAR son estimados usando la inferencia bayesiana empírica (Fahrmeir et al., 2004).

En los Modelos Aditivos Generalizados Estructurados (STAR) (Fahrmeir et al., 2004) y (Brezger and Lang, 2006) se asume que, dadas las covariables $\boldsymbol{x}$, la distribución de la variable respuesta y_i, $i = 1, \ldots, n$ pertenece a la familia exponencial. La media condicional $\mu_i = E(y_i)$ está relacionada con un predictor aditivo semiparamétrico η_i de manera que $\mu_i = h(\eta_i)$, siendo h una función respuesta conocida. El predictor η_i es de la forma:

$$\eta_i = f_1(x_{i1}) + \cdots + f_q(x_{iq}) + \boldsymbol{v}_i'\boldsymbol{\gamma}, \qquad i = 1, \ldots, n \tag{7.9}$$

donde $f_1, \ldots, f_q$ son funciones no lineales de las covariables $x_1, \ldots, x_q$ y $\boldsymbol{v}_i'\boldsymbol{\gamma}$ es la parte lineal del modelo.

Las funciones no lineales en la Ecuación 7.9 son modeladas a través de funciones bases, así

una particular función de f para la covariable $\boldsymbol{x}$ es aproximada por la combinación lineal de las bases o funciones indicador:

$$f(\boldsymbol{x}) = \sum_{k=1}^{K} \beta_k B_k(\boldsymbol{x}) \tag{7.10}$$

Las B_k son conocidas como funciones bases y $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_k)^{'}$ es un vector de coeficientes de regresión desconocidos que deben ser estimados. Si definimos la matriz $\boldsymbol{X}$ de orden $n \times K$ con los elementos $\boldsymbol{X}[i,k] = B_k(\boldsymbol{x_i})$, entonces el vector $\boldsymbol{f} = (f(\boldsymbol{x}_1), \ldots, f(\boldsymbol{x_n}))$, puede ser escrito en forma matricial como $\boldsymbol{f} = \boldsymbol{X\beta}$ y teniendo en cuenta el predictor definido en la Ecuación 7.9 obtenemos:

$$\boldsymbol{\eta} = \boldsymbol{X}_1\boldsymbol{\beta}_1 + \cdots + \boldsymbol{X}_q\boldsymbol{\beta}_q + \boldsymbol{v}^{'}\boldsymbol{\gamma} \tag{7.11}$$

En el contexto de un STAR se pueden estimar modelos muy versátiles, considerando los *Generalized Additive Mixed Models* (GAMM) como el trabajo de Kuhlenkasper and Steinhardt (2012) y el trabajo de Shadish et al. (2014), Modelos de Regresión Geo-Aditivos como los empleados en Kandala et al. (2009, 2011), los *Varying Coefficient Models* (Hastie and Tibshirani, 1993) y (Fahrmeir et al., 2003) así como los *Geographically Weighted Regression Models* estudiados en Harris et al. (2010) y en Wheeler (2014) como casos particulares.

La principal ventaja del uso de estos modelos es que permiten combinar los efectos no lineales de los regresores sobre la variable respuesta y que además permiten controlar la autocorrelación en el espacio funcional al permitir incluir en el modelo los efectos espaciales.

En un modelo de regresión lineal generalizado (GLM) asumimos que, dado el vector de covariables $\boldsymbol{x}$ y el vector de parámetros desconocidos $\boldsymbol{\beta}$, la distribución de la variable respuesta $\boldsymbol{y}$ pertenece a la familia exponencial con media $\mu = E(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\beta})$ y predictor lineal dado por la Ecuación 3.8 siendo la *link function* h y los coeficientes de la regresión $\boldsymbol{\beta}$ desconocidos. Al llegar al modelo STAR, tenemos que sustituir al predictor lineal por un predictor de forma más flexible que sea aditivo estructurado, tal que:

$$\begin{aligned} \eta &= f_1(\boldsymbol{x}) + \cdots + f_p(\boldsymbol{x}) + \boldsymbol{v}'\boldsymbol{\gamma} \\ \mu &= E(y|\boldsymbol{x}, \boldsymbol{v}, \boldsymbol{\gamma}, \boldsymbol{\theta}) \end{aligned} \tag{7.12}$$

$\boldsymbol{x}$ representa a todas las covariables que se modelarán a través de una relación no lineal

con $\boldsymbol{y}$ y el vector $\boldsymbol{\theta}$ incluye a todos los parámetros de las funciones $f_1, \ldots, f_p$. La parte estrictamente lineal entre la variable respuesta y los regresores se recogerá en $\boldsymbol{v}'\boldsymbol{\gamma}$.

Las funciones $f_j(\cdot)$ serán posiblemente suaves y tendrán en cuenta varios tipos de efectos (Umlauf et al., 2015). En la Figura 7.2 aparecen las más utilizadas.

Efecto no lineal de algunas covariables continuas:

$$f_j(\boldsymbol{x}) = f(x_1)$$

Efecto en dos dimensiones (*Two-dimensional surfaces*):

$$f_j(\boldsymbol{x}) = f(x_1, z_2)$$

Efectos correlados espacialmente (*Spatially correlated effects*):

$$f_j(\boldsymbol{x}) = f_{spat}(x_s)$$

Coeficientes cambiantes (*Varying Coefficients*):

$$f_j(\boldsymbol{x}) = x_1 f(x_2)$$

Efectos cambiantes espacialmente (*Spatially correlated effects*):

$$f_j(\boldsymbol{x}) = x_1 f_{spat}(x_s)$$
$$f_j(\boldsymbol{x}) = x_1 f_{spat}(x_2, x_3)$$

Intercepto aleatorio con índice clúster c (*Random intercepts with cluster index c*):

$$f_j(\boldsymbol{x}) = \beta_c$$

Pendientes aleatorias con índice clúster c (*Random slopes with cluster index c*):

$$f_j(\boldsymbol{x}) = x_1 \beta_c$$

Figura 7.2: *Efectos recogidos en las funciones de un STAR.*

Modelos similares en cuanto a complejidad han sido desarrollados desde el punto de vista frecuentista por Wood (2003, 2006); Ruppert et al. (2003); Rigby and Stasinopoulos (2005) y Rue et al. (2009). Estos modelos fueron concebidos como modelos de regresión semiparamétrica con variables dependientes discretas o continuas intentando combinar diferentes efectos de las covariables en una única estructura, extendiendo de esta forma los modelos aditivos generalizados (GAM) a una situación más compleja (Fahrmeir and Kneib, 2009).

Debido a que las funciones $f_j(\boldsymbol{x})$ son, a menudo centradas alrededor de cero, se debería probar si estas funciones difieren significativamente de cero. Además, al poder incluir efectos lineales y no lineales, se tendría que probar que el efecto recogido es lineal, con lo que $f_j(\boldsymbol{x}) = x_j\beta_j$. En caso de que el efecto sea no lineal debería ocurrir que $f_j(\boldsymbol{x}) = f_j(x_j)$. En ambos casos debería contrastarse si esas funciones difieren significativamente de una línea recta.

Por otro lado, como podemos recoger términos cambiantes, tendríamos que preguntarnos si la función que hemos seleccionado es una buena aproximación o si necesitamos incorporar términos de coeficientes cambiantes tales como: $f_1(x_1)\,x_2$ ó interaciones tales como: $f_{12}(x_1, x_2)$.

Según Scheipl et al. (2013) la mayoría de las aproximaciones frecuentistas y bayesianas para la selección de funciones se basa en la selección de variables en los modelos de regresión con predictores lineales altamente dimensionados[2]. Dentro de estas propuestas cabe destacar los trabajos de Tibshirani (1996) utilizando Lasso como método de penalización; Fan and Li (2001) donde se emplea SCAD penalty; Zou (2006) con su propuesta de *Adaptive Lasso*, así como Bühlmann and Yu (2003) y Bühlmann and Hothorn (2007) con el método Boosting. Scheipl et al. (2013) proponen Mínimos Cuadrados Penalizados o Métodos de Verosimilitud por un lado ó Métodos Bayesianos por otro basados en la selección de indicadores.

Los modelos STAR más recientes son los denominados modelos *Multilevel STAR* (Lang et al., 2014) los cuales asumen que los coeficientes β_j en la Ecuación 7.11 son en sí mismos modelos de regresión con un predictor estructurado, o sea:

$$\beta_j = \nu_j + \varepsilon_j = \boldsymbol{X}_{j1}\boldsymbol{\beta}_{j1} + \cdots + \boldsymbol{X}_{jq_j}\boldsymbol{\beta}_{jq_j} + \boldsymbol{v}_j^{'}\boldsymbol{\gamma}_j + \boldsymbol{\varepsilon_j} \tag{7.13}$$

siendo los términos $\boldsymbol{X}_{j1}\boldsymbol{\beta}_{j1} + \cdots + \boldsymbol{X}_{jq_j}\boldsymbol{\beta}_{jq_j}$ correspondientes a funciones $f_{j1}, \ldots, f_{jq_j}$ adicionales y $\boldsymbol{v}_j^{'}\boldsymbol{\gamma}_j$ comprende efectos lineales adicionales; $\boldsymbol{\varepsilon_j} \sim N\left(\boldsymbol{0}, \sigma_j^2\boldsymbol{I}\right)$.

[2] Los modelos altamente dimensionados son modelos donde la cantidad de parámetros a estimar p es superior al número de datos n.

7.3. Los Modelos Cuantílicos Aditivos Estructurados (STAQ)

La utilización de la regresión cuantílica aditiva parte de la necesidad de permitir mayor flexibilidad en el planteamiento de la relación entre los regresores y la variable respuesta en el contexto de la regresión cuantílica.

Koenker (2005) plantea que si se conoce a priori la forma funcional de esta relación lo más lógico resultará en utilizar la regresión cuantílica lineal (paramétrica). Si la forma funcional no se conoce, entonces podemos recurrir a la estimación no paramétrica como el trabajo de Takeuchi et al. (2006) o la estimación semiparamétrica como lo propuesto en Bühlmann and Hothorn (2007).

En la regresión cuantílica lineal, partíamos de plantear la relación entre la variable respuesta y las covariables como

$$y_i = \beta_{\tau 0} + \beta_{\tau 1} x_1 + \cdots + \beta_{\tau k} x_k + \varepsilon_{\tau i} \qquad i = 1, \ldots n \tag{7.14}$$

Con $\tau \in (0,1)$, $\varepsilon_{\tau i} \sim H_{\varepsilon_{\tau i}}$ comprobándose que $H_{\varepsilon_{\tau i}}(0) = \tau$.

Ahora lo que se hace es sustituir la relación lineal definida en (7.14) por una estructura no paramétrica o semiparamétrica

$$y_i = f_\tau(x_i) + \varepsilon_{\tau i} \qquad i = 1, \ldots, n \tag{7.15}$$

donde $f_\tau(x_i)$ será una función suave desconocida, el τ - cuantil del témino de error $\boldsymbol{\varepsilon}$ condicionado al vector de covariables $\boldsymbol{x}$ es igual a cero, $Q_\tau(\varepsilon_{\tau i}/x_i) = 0$. La estimación de f se obtiene a partir del uso de alguna función suave con la forma

$$\hat{f}_\tau(x) = \sum_{i=1}^{n} w_{\lambda,\tau}(x_i)\, y_i \tag{7.16}$$

siendo λ el parámetro de suavidad y $w_{\lambda,\tau}$ una función de pesos pudiendo ser un spline, o un tipo kernel, etc. Esto nos indica que la función $f_\tau(x)$ se puede estimar de forma no paramétrica estableciendo únicamente supuestos sobre su grado de suavidad.

Supongamos que tenemos como único regresor a una variable continua, el problema de minimización planteado en la Ecuación 5.20 puede escribirse como

$$argmin_{\beta_\tau} \sum_{i=1}^{n} \rho_\tau \left(y_i - f_\tau \left(x_i \right) - \lambda V \left(f'_\tau \right) \right) \tag{7.17}$$

$V\left(f'_\tau\right)$ denota la variación total de la derivada f'_τ la cual es definida como

$$V\left(f'_\tau\right) = sup \sum_{i=1}^{n} \left| f'_\tau \left(x_{i+1} \right) - f'_\tau \left(x_i \right) \right| \tag{7.18}$$

siendo $\lambda > 0$ el parámetro que controla la suavidad de la función estimada. Para la función f'_τ continua y diferenciable, la variación total puede ser planteada como

$$V\left(f'_\tau\right) = \int \left| f''_\tau \left(x \right) \right| dx \tag{7.19}$$

Esto es la L_1 norm de f''_τ. En los modelos de regresión clásica, este término es el L_2 norm de f''_τ. [3]

Si λ toma un valor muy bajo, incluso cercano a cero, la función estimada provee una interpolación de los datos. En cambio si λ toma un valor alto, lo que obtenemos es un ajuste por mínimos cuadrados ya que la penalización dominará a la fidelidad. El parámetro λ está relacionado con el número de puntos de interpolación (p_λ) los cuales se consideran una medida de los grados de libertad del ajuste (Hendricks and Koenker, 1992). Si λ crece, p_λ decrece y viceversa.

En particular, $p_\lambda - 1$ representa el número de segmentos ajustados que juntos conformarán el ajuste final. El parámetro de suavidad, por tanto sirve de balance entre la fidelidad de los datos y la suavidad del ajuste.

Los Modelos Cuantílicos Aditivos Estructurados (STAQ) que se deben a los trabajos de Fahrmeir et al. (2004), Kneib et al. (2009) y Fenske et al. (2012) consideran que dadas las observaciones $(y_i, \boldsymbol{x}_i)$ donde y_i denota la variable respuesta y $\boldsymbol{x}_i$ es el vector de covariables para la observación i, la relación entre la función cuantílica de y_i y el predictor estructurado aditivo $\eta_{\tau i}$ para el cuantil específico τ puede ser establecida como:

[3] En la regresión cuantílica clásica, el L_2norm no es apropiado puesto que impide el uso de la programación lineal para obtener los valores estimados óptimos.

$$Q_{\boldsymbol{y}_i}(\tau|\eta_{\tau i}) = \eta_{\tau s}(\boldsymbol{x}_i) \tag{7.20}$$

Esta notación es similar a la planteada para el modelo de regresión cuantílica lineal en el Capítulo 5 donde se definía a $Q_{\boldsymbol{y}_i}(\tau|\boldsymbol{x}_i) = \boldsymbol{x}_i'\boldsymbol{\beta}_\tau = \eta_{\tau i}$ en la Ecuación 5.19 pero ahora permitiendo una estructura flexible para el predictor $\eta_{\tau i}$. Sobre el término de error se asume que $F_{\varepsilon_{\tau i}}(\tau) = 0$

Sabiendo que los parámetros son cuantiles-específicos, podemos plantear el predictor estructurado aditivo $\eta_{\tau i}$ como:

$$\eta_i = \beta_0 + \sum_{d=1}^{D} h_d(\boldsymbol{x}_i) \tag{7.21}$$

donde β_0 es el intercepto y h_d son funciones cuya flexibilidad en cuanto a la variedad de funciones que se pueden tener en cuenta, permite la inclusión de una gran variedad de modelos (Kneib et al., 2009):

- Componentes lineales (*Linear components*)

  $$h_d(\boldsymbol{x}_i) = \beta_d x_{ik}$$

  con parámetro de regresión lineal β_d para covariables categóricas o continuas x_k.

- Componentes no lineales suaves (*Smooth nonlinear components*)

  $$h_d(\boldsymbol{x}_i) = f_d(x_{ik})$$

  con covariable x_k continua y función f_d suave y potencialmente no lineal.

- Términos de Coeficientes Cambiantes (*Varying coefficient terms*)

  $$h_d(\boldsymbol{x}_i) = x_{ik} \cdot f_d(x_{il})$$

  con x_k covariables categóricas o continuas, f_d una función suave de la covariable x_l. De esta forma el efecto de la covariable x_l varía sobre el dominio de x_k de acuerdo a la función f_d.

- Superficies bivariantes (*Bivariate surfaces*)

  $$h_d(\boldsymbol{x}_i) = f_d(\boldsymbol{x}_{ik}, \boldsymbol{x}_{il})$$

siendo f_d una función bivariante suave de dos variables continuas, x_k y x_l denotan la longitud y la latitud de datos que están orientados espacialmente.

- Componentes espaciales discretos (*Discrete spatial components*)

$$h_d(\boldsymbol{x}_i) = f_d(\boldsymbol{x}_{ik})$$

la variable x_k contiene información espacial discreta, como puede ser un municipio dentro de una provincia y la función f_d recoge los efectos espaciales del conjunto de vecinos, por ejemplo el conjunto de los municipios de la provincia.

- Componentes cluster (*Cluster-specific components*)

$$h_d(\boldsymbol{x}_i) = x_{il} \cdot ([I(x_{ik} \in G_1), \ldots, I(x_{ik} \in G_K)]^\intercal \gamma_d)$$

$I(\cdot)$ es la función indicador, x_l puede ser una variable categórica o continua, x_k es una variable categórica con K grupos diferentes o clúster $G_1, \ldots, G_K$ y γ_d es un vector $(K \times 1)$ que contiene los parámetros cluster-específicos; de esta forma el efecto de x_l difiere a través de los grupos $G_1, \ldots, G_K$ definidos por el factor de grupo x_k.

Cada uno de los componentes de un modelo STAQ constituyen bases. Fenske et al. (2012) propusieron estimar estos modelos a través del *Component-wise Functional Gradient Descent Boosting* (Bühlmann and Yu, 2003) utilizando P-splines como bases. Este método de optimización proviene del Aprendizaje Automático y permite obtener la estimación de un modelo aditivo estructurado a través de la técnica del gradiente funcional descendente. En las siguientes secciones se describe el origen de este método y su utilización en la Estadística.

CAPÍTULO 8

El Boosting en la Regresión Cuantílica

El Boosting (Schapire, 1990) es un algoritmo que originalmente estaba limitado a la clasificación en la comunidad del Aprendizaje Automático pero que ha trascendido este campo del conocimiento hasta ser utilizado en problemas de optimización y de estimación de modelos de regresión. En este capítulo nos centramos en el Boosting desde sus orígenes hasta su utilización en la Estadística para estimar modelos de regresión. Para ello se realiza una panorámica de la utilización del Boosting en el análisis de regresión y su desarrollo hasta llegar a su empleo en los modelos cuantílicos aditivos estructurados. Breiman (1998, 1999) estableció la conexión entre estos algoritmos de clasificación y la Estadística, sembrando las bases de lo que posteriormente se ha venido a establecer como una potente herramienta para la estimación.

8.1. El Aprendizaje Automático y la Estadística

Breiman (2001b) divide a los estadísticos según la manera de abordar los principales objetivos de la Estadística, a saber la predicción y la inferencia, en dos grupos: aquellos que utilizan los Modelos Estocásticos para plantear la relación entre la variable respuesta y las covariables y los que utilizan el Aprendizaje Automático o Machine Learning (ML), por su nombre en inglés, en el que se asume que la relación entre las variables es desconocida y lo que tenemos que hacer es encontrar una función $f(\boldsymbol{x})$ entendiendo ésta como un algoritmo matemático que opera sobre el vector de covariables $\boldsymbol{x}$ para predecir la variable respuesta $\boldsymbol{y}$.

El Aprendizaje Automático es una rama dentro de la Inteligencia Artificial[1] en la que básicamente lo que se pretende es aprender de los datos mediante la creación de algoritmos capaces de generalizar comportamientos y reconocer patrones a partir de la información suministrada por los datos a través de la utilización de un ordenador.

El Aprendizaje Automático está presente en nuestras vidas sin que seamos conscientes de ello. Cuando entramos en páginas de compra por internet como Amazon, Booking, etc es el Aprendizaje Automático el que hace que nos aparezcan recomendaciones de productos similares a los que hemos adquirido o simplemente revisado; cuando pagamos con nuestra tarjeta de crédito, permite comparar la transacción que hemos realizado con una base de datos de transacciones y así detectar fraudes; cuando en Google realizamos búsquedas de información relacionada con un tema de nuestro interés permite que se filtren billones de páginas para encontrar las más relevantes.

En los correos electrónicos, es por medio del Aprendizaje Automático que se clasifican los mensajes como spam o no y con su utilización se puede predecir el número de personas que comprarán un determinado producto o servicio y también el número de personas que se darán de baja de un determinado servicio.

Los pioneros en utilizar estos algoritmos matemáticos en la Estadística lo hicieron con el propósito de mejorar la predicción de los modelos, si existían algoritmos capaces de combinar múltiples modelos y al final ofrecer la mejor predicción posible de la variable respuesta, por qué no utilizarlos.

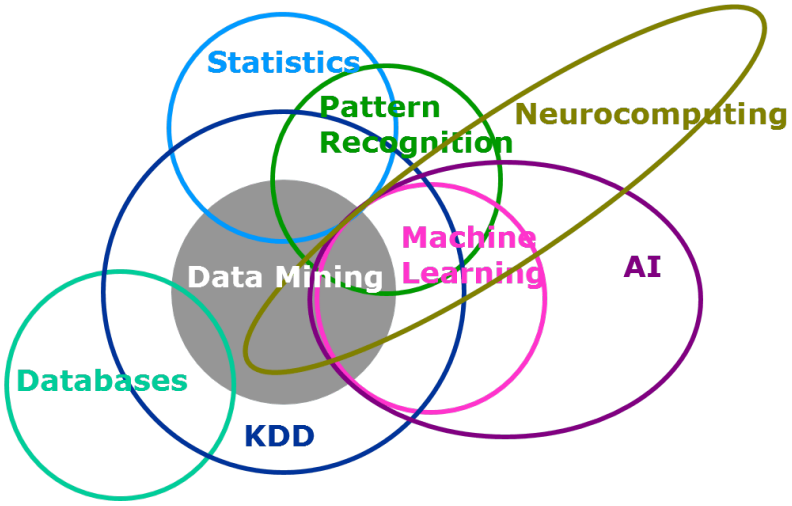
Las tres contribuciones más importantes del Aprendizaje Automático a la Estadística son: *Support Vector Machines* (SVM) (Cortes and Vapnik, 1995), *Boosting* (Schapire 1990; Freund 1995; Freund and Schapire 1996) y el *Random Forests* (Breiman, 2001a).

El diagrama que aparece en la Figura 8.1 (aunque de 1978) sintetiza cómo se entremezclan las diferentes ramas de la Inteligencia Artificial y la Estadística.

Support Vector Machines (SVM) (Cortes and Vapnik, 1995) está formado por un grupo de métodos de aprendizaje que pueden ser aplicados a la clasificación o a la regresión. Es una extensión a los modelos no lineales del *Generalized Portrait Algorithm* desarrollado por Vapnik and Lerner (1963) y Vapnik and Chervonenkis (1964) y está basado en la teoría del Aprendizaje en Estadística (*Statistical Learning*) y la denominada dimensión Vapnik-Chervonenkis (VC)[2] desarrollada por Vladimir Vapnik y Alexey Chervonenkis

[1]La Inteligencia Artificial puede definirse como un conjunto de métodos de Ingeniería que simulan la estructura y los principios operacionales del cerebro humano (Cleophas and Zwinderman, 2013).

[2]Podemos encontrar una definición de la VC dimension en Denison et al. (2013).



Fuente: Looking backwards, looking forwards: SAS, data mining, and machine learning. Polly Mitchel Guthrie. SAS.

Figura 8.1: *Diagrama de Venn para representar la relación entre la Estadística y el Machine Learning.*

entre 1967-1968 (Vovk et al., 2015). Vladimir N. Vapnik, uno de los fundadores de la teoría del Aprendizaje en Estadística, relaciona la Estadística con la teoría del aprendizaje de la forma siguiente:

> *The problem of learning is so general that almost any question that has been discussed in statistical science has its analog in learning theory. Furthermore, some very important general results were first found in the framework of learning theory and then formulated in the terms of statistics (Vapnik, 2013).*

El Support Vector Machine y sus variantes: *Least Squares SVM* (LS-SVM) (Suykens and Vandewalle, 1999); *Proximal SVM* (PSVM) (Fung and Mangasarian, 2001) y *Reduced SVM* (RSVM) (Lee and Mangasarian, 2001) ha sido ampliamente utilizado en las últimas décadas en aplicaciones de clasificación debido a su capacidad de clasificación y con la introducción de la función de pérdida, se ha extendido a solucionar problemas de regressión (Drucker et al., 1997).

Bajo el SVM se han empleado diferentes aproximaciones para modelar la distribución de probabilidades de la variable respuesta tomando como referencia la regresión cuantílica: *Quantile Regression Forest* (Meinshausen, 2006), *Quantile Regression Neural Networks* (Taylor, 2000) y *Kernel-Based Quantile Regression* (Christmann and Hable, 2012).

En el ***Boosting*** (Schapire, 1990) un número de clasificadores débiles o bases (*base learners*

por su nombre en inglés)[3] tales como los árboles y los *stumps* (árboles con un split y dos nodos terminales) son combinados (boosted) para producir un clasificador de conjunto (ensemble classifier) con un error generalizado menor (Kuhn and Johnson, 2013).

El Boosting es uno de los más significativos avances en el *Machine Learning* para clasificación y regresión; lo que básicamente hace es llamar a esos clasificadores débiles repetidamente sobre diferentes conjuntos de entrenamiento generando cada vez una regla de decisión débil, la cual combinada con las otras reglas generadas, producen un clasificador que predice mejor que todas esas reglas de forma separada. (Schapire, 2003).

El ***Random Forests*** (Breiman, 2001a) es una modificación importante del Bagging (Breiman, 1996). Construye una gran colección de árboles decorrelados y los promedia. En algunas situaciones, funciona de manera similar al Boosting, lo que le ha hecho muy popular. Lo que se intenta con este algoritmo es mejorar la reducción de varianza que ofrece el Bagging a partir de disminuir la correlación entre los árboles (Friedman et al., 2001).

Cuando Breiman dividió a los estadísticos en estas dos ramas no existía el *Extreme Machine Learning* (ELM)(Huang et al., 2006b) el cual también se ha extendido a la regresión como puede verse en los trabajos de Li et al. (2005); Huang et al. (2006a); Liang et al. (2006); Feng et al. (2009); Rong et al. (2009) y Huang et al. (2012). El *Extreme Machine Learning* puede usarse además en el análisis clúster y en la clasificación binaria o multiclase y el término extremo significa avanzar más allá de las técnicas de aprendizaje convencionales (Huang, 2015).

Utilizar la Estadística desde el punto de vista del Machine Learning ha adquirido gran utilidad ya que actualmente se utilizan conjuntos de datos más complejos y más grandes aunque hay quienes afirman que tiene el inconveniente de que se basa solo en la predicción y hay campos como la medicina donde no sólo es necesario predecir los valores de la variable respuesta sino también es preciso establecer el mecanismo que asocia a unas variables con otras y los modelos algorítmicos derivados del Random Forest, Support Vector Machines o el Boosting no necesariamente proveen a los estadísticos de las herramientas necesarias para la inferencia estadística.

Sin embargo, se han realizado y se realizan esfuerzos por combinar la Estadística basada en el Aprendizaje Automático y la Estadística tradicional basada en los Modelos Estocásticos.

Muestra de ello es la utilización del *Functional Gradient Descent Boosting* (Friedman

[3]Se consideran calsificadores débiles aquellos que sólo predicen marginalmente mejor que los clasificadores aleatorios.

et al., 2000) y del *Component-Wise Functional Gradient Descent Boosting* (Bühlmann and Yu, 2003) como algoritmos de optimización en los modelos aditivos generalizados con lo que se pueden interpretar los coeficientes estimados obtenidos o ver la contribución parcial de cada parte aditiva considerada en el modelo planteado.

De esta forma se pueden combinar las propiedades del Boosting como algoritmo matemático y la parte de inferencia de un modelo estocástico. Breiman (2001b) apunta que el Boosting podría ser el método que cierre la brecha existente entre las Estadística y el Machine Learning.

En la literatura encontramos la utilización del Boosting en la regresión cuantílica en diferentes contextos. Destacado el trabajo de Fenske et al. (2012) sobre la malnutrición infantil de los niños en la India, siendo uno de los trabajos pioneros en utilizar el *Component-Wise Functional Gradient Descent Boosting*; el trabajo de Mayr et al. (2012b) donde se utiliza esta técnica para predecir los valores de los índices BMI en los niños. En el mismo año Zheng (2012) también utiliza el Boosting en la regresión cuantílica a través *Component-Wise Functional Gradient Descent Boosting*. Fenske et al. (2013a) hacen una comparación entre el Boosting y el Bagging en la regresión cuantílica para ampliar su análisis sobre la malnutrición infantil y Martínez-Silva et al. (2013) lo utilizan para estudiar las barreras de coral.

8.2. Los fundamentos del Boosting

El Boosting (Schapire, 1990) junto con el Bagging (Breiman, 1996) es una de las técnicas que más se ha popularizado en la construcción de Conjuntos de Clasificadores (*Ensembles Methods*). Los Conjuntos de Clasificadores son generalmente variantes de un mismo clasificador.

Para entender mejor cómo funcionan estos algoritmos, debemos comprender primero qué es la Clasificación. Por Clasificación se entiende el proceso que ubica nuevas observaciones en las clases existentes a través de una regla previamente establecida. Es necesario especificar que se le denomina Clasificación en el lenguaje estadístico pero dentro de la Comunidad de Aprendizaje Automático se le conoce como aprendizaje supervisado.

Se tiene un conjunto de registros de entrada $\mathcal{D} = X_1, ..., X_n$, al que se le denomina ***conjunto de entrenamiento*** y cada registro pertenece a una clase k del conjunto de valores discretos $\{1, \ldots, k\}$. Este ***conjunto de entrenamiento*** se utiliza para construir un ***modelo de clasificación*** que relaciona cada uno de los registros con una clase,

dependiendo de las características subyacentes de ese registro. Se dice que con el conjunto de entrenamiento el algoritmo lo que hace es aprender la regla de clasificación.

Para Dietterich (2000b) un clasificador es una hipótesis para la verdadera función $f(\cdot)$, así dando nuevos valores a $\boldsymbol{x}$, el clasificador predice el correspondiente valor y.

En el proceso de validación de un clasificador, se utiliza por otro lado, otro conjunto de registros que no es conocido en el proceso de aprendizaje, denominado el ***conjunto de prueba*** y con el que se comprueba la precisión del clasificador.

Para evaluar un clasificador se tienen en cuenta las siguientes propiedades (Alberto Fernández, PhD, 2010):

- **Precisión**: representa el nivel de confianza del clasificador, usualmente representado como la proporción de clasificaciones correctas que es capaz de producir.
- **Velocidad**: es el tiempo de respuesta desde que se presenta un nuevo registro a clasificar hasta que se obtiene la clase que el clasificador predice. La velocidad suele ser tan importante como la precisión.
- **Interpretabilidad**: claridad y credibilidad, desde el punto de vista humano de la regla de clasificación.
- **Velocidad de Aprendizaje**: es el tiempo requerido por un clasificador para obtener la regla de clasificación desde un conjunto de registros.
- **Robustez**: número mínimo de ejemplos necesarios para obtener una regla de clasificación fiable y precisa.

En la Figura 8.2 se resume el modo de funcionamiento de un clasificador.

Los métodos de clasificación siguen la idea de que una potente herramienta de predicción permite agregar la información que se posee de manera eficiente. Se contruyen múltiples funciones estimadas o predicciones a partir de ir reponderando los datos y usando una combinación de ellos, que puede ser lineal o de otro tipo, para llegar a un estimador o predictor agregado. Las aproximaciones que se agrupan dentro de estos métodos difieren básicamente en tres aspectos:

- Las bases (*base-learners*) empleadas.

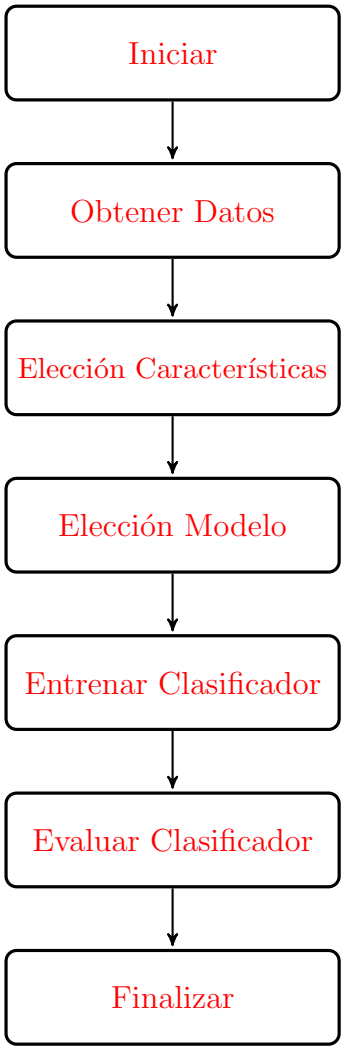


Figura 8.2: *Método de clasificación.*

- La forma en la que se va agregando la información (voto mayoritario contra voto ponderado).
- La forma en la que se generan submuestras o se re-ponderan las observaciones en cada iteración.

En los algoritmos de clasificación hay que distinguir dos tipos de errores existentes:

- Error de entrenamiento: es la fracción de errores sobre el conjunto de entrenamiento.
- Error generalizado: es la probabilidad de clasificar mal un nuevo ejemplo o registro siendo por tanto igual al error de prueba esperado (*expected test error*). El error de prueba (*test error*) es, a su vez, la fracción de errores sobre un nuevo conjunto de prueba.

Una vez entendida la Clasificación, podemos explicar qué es un Conjunto de Clasificadores tomando como referencia lo que sobre ellos apuntaron Bühlmann and Hothorn (2007). Lo primero que debemos especificar es el procedimiento base (*base procedure*)[4] con el cual construiremos la función estimadora $\hat{g}(\cdot)$ definida en $\mathbb{R}$ y basada en un conjunto de datos $(x_1, y_1), \ldots, (x_n, y_n)$:

$$(x_1, y_1), \ldots, (x_n, y_n) \overset{\text{base procedure}}{\rightarrow} \hat{g}(\cdot)$$

donde $(x_1, y_1), \ldots, (x_n, y_n)$ son las observaciones originales de las variables aleatorias x e y. Para cada iteración, diferentes conjuntos de datos son generados a través de re-ponderar los datos originales.

La manera en la que trabaja un Conjunto de Clasificadores sería el siguiente:

Re-ponderación de la primera observación $\overset{\text{base procedure}}{\rightarrow} \hat{g}^{[1]}(\cdot)$

Re-ponderación de la segunda observación $\overset{\text{base procedure}}{\rightarrow} \hat{g}^{[2]}(\cdot)$

$\ldots \quad \ldots$

$\ldots \quad \ldots$

Re-ponderación de la M observación $\overset{\text{base procedure}}{\rightarrow} \hat{g}^{[M]}(\cdot)$

Agregación: $\hat{f}_A(\cdot) = \sum_{m=1}^{M} \alpha_m \hat{g}^m(\cdot)$

Figura 8.3: *La esencia de los Conjuntos de Clasificadores.*

El término **dato reponderado** (*reweighted data*) significa que se asigna un peso individual a cada una de las n observaciones, lo que conduce a basar la estimación en una muestra ponderada.

Los *Ensemble Methods* fueron introducidos en la Comunidad del Aprendizaje automático para los problemas de clasificación donde las decisiones individuales son combinadas de alguna forma (típicamente por voto ponderado o no) para clasificar nuevos ejemplos.[5]

[4] Un procedimiento base puede ser, por ejemplo, una regresión lineal.

[5] (Dietterich, 2000b) realiza un excelente análisis de las propiedades de los Ensemble Methods.

El Bagging y el Boosting tienen en común que toman un algoritmo de aprendizaje (base learning algorithm) y lo ejecutan muchas veces con diferentes conjuntos de entrenamiento (Dietterich, 2000a).

En el Bagging (acrónimo de Bootstrap Aggregating) el conjunto de entrenamiento es construido a partir de replicar el conjunto de entrenamiento original mediante la técnica Bootstrap (Efron, 1979); (Friedman et al., 2001). El Bagging fue introducido por Breiman (1996) con el objetivo de disminuir la varianza de un predictor insesgado o de bajo sesgo pero con una varianza alta, siendo los árboles ideales para conseguir este objetivo porque son capaces de capturar estructuras de interacción en los datos.

Bühlmann and Yu (2002) definen el Bagging de la siguiente forma:

(i) Se construye una muestra bootstrap: $L_i^* = (\boldsymbol{y}_i^*, X_i^*)$ de acuerdo a la distribución empírica de $L_i = (\boldsymbol{y}_i, X_i)$ donde $\boldsymbol{y}_i$ es el valor de la variable respuesta, X_i es un vector p-dimensional para las covariables $i = 1, \ldots, n$.

(ii) Se computa el predictor bootstrapped $\hat{\theta}_n^*(x)$ por el principio plug-in, o sea:

$$\hat{\theta}_n^*(x) = h_n(L_1^*, \ldots, L_n^*)(x)$$

donde $\hat{\theta}_n(x) = h_n(L_1, \ldots, L_n)(x)$

(iii) El bagged predictor será:

$$\hat{\theta}_{n;B}(x) = \mathsf{E}^*\left[\hat{\theta}_n^*(x)\right]$$

En el Bagging cada observación (y_i, x_i) para $i = 1, \ldots, n$ tiene la misma probabilidad de entrar en la muestra iteración tras iteración. Lo que hace esta técnica es generar múltiples versiones de una base y combinarlas para obtener un predictor agregado combinando los métodos boostraps con los de agregación pues a través del muestreo aleatorio con reemplazamiento se obtienen B muestras de igual tamaño que el conjunto original $X^b = \left(X_b^1, X_2^b, \ldots X_n^b\right)$ y se construyen los clasificadores $C_b(X)$. La agregación se puede realizar a través de promediar las salidas obtenidas en la regresión o a través del voto mayoritario o ponderado en los problemas de clasificación.

Con el Boosting, en cambio, nos fijamos iteración tras iteración en las observaciones que han sido problemáticas en el pasado, con el objetivo de mejorar la predicción, lo que

significa que no usamos el Bootstrap para generar los datos sino que re-ponderamos las observaciones en cada iteración. Este algoritmo genera un conjunto de clasificadores y elige entre ellos de forma secuencial porque lo hace en cada iteración, a diferencia del Bagging que lo hace de forma paralela.

La idea básica del Boosting es detectar aquellas observaciones que menos influyen en la predicción y asignarles un gran peso en la siguiente iteración con lo que se va asignando peso en cada iteración a los clasificadores más débiles, siendo un método efectivo en la reducción del sesgo y de la varianza (Bauer and Kohavi, 1999).

El Boosting, fue introducido por (Schapire, 1990) dentro de la comunidad del Aprendizaje Automático (*Machine Learning Community*) como un algoritmo de clasificación binaria. Su posterior desarrollo ha sido auspiciado por los trabajos de Freund (1995) y Freund and Schapire (1997), bajo la influencia de la teoría del aprendizaje promovida por Valiant (1984) y Kearns and Valiant (1994).

Para Friedman et al. (2000) el Boosting es uno de los algoritmos más importantes de clasificación y básicamente lo que hace es aplicar algoritmos de clasificación secuenciales a diferentes versiones reponderadas del conjunto de entrenamiento tomando un voto mayoritario (ponderado) de la secuencia de clasificadores que produce.

En el Boosting se mantienen los pesos del conjunto de entrenamiento original y esos pesos se van ajustando iteración tras iteración, dando mayor peso a los elementos que han sido mal clasificados y menor peso a aquellos que fueron correctamente clasificados en la iteración anterior.

Schapire (1990) demostró que un clasificador (*learner*) débil se puede reconvertir en un learner fuerte en el sentido de la *Probably Approximately Correct* (PAC) de la teoría del aprendizaje (Shrestha and Solomatine, 2006). Un clasificador débil [6] es aquel en el que pequeños cambios de los datos, provocan grandes cambios en el modelo planteado (Galar et al, 2007).

El trabajo de Schapire (1990) fue la primera propuesta de un algoritmo Boosting. Más tarde Freund (1995) desarrolló un algoritmo Boosting más eficiente. El primer experimento con estos algoritmos Boosting fue realizado por Drucker et al. (1993).

Freund and Schapire (1996, 1997) introdujeron el algoritmo AdaBoost (Adaptative Boosting) para clasificación, resolviendo alguna de las dificultades de los algoritmos Boosting

[6] La definición de aprendizaje débil es dada por Kearns and Valiant (1994) de manera que se pueda entender por qué un clasificador débil que sólo predice mejor que uno aleatorio puede ser combinado (boosted) dentro de un algoritmo para producir un clasificador fuerte.

previos y convirtiéndole en uno de los algoritmos de clasificación más utilizados y conocidos debido a su solidez, precisión y simplicidad siendo su intención inicial, reducir el error de entrenamiento a cero (Wu et al., 2008).[7]

AdaBoost fue la primera aplicación del Boosting. Se utilizó en la comunidad del Aprendizaje Automático como un procedimiento que agregaba clasificadores débiles. Una de las razones de su éxito es que se centra en las observaciones que son difíciles de clasificar.

Utiliza todo el conjunto de datos pero iteración tras iteración se centra en las observaciones que han sido problemáticas en la iteración anterior, dándoles mayor ponderación, obligando de esta forma al clasificador, a tener más en cuenta estas observaciones, en la siguiente iteración. Esa ponderación inicialmente es la misma para todas las observaciones.

Cuando se pasa a la siguiente iteración, los pesos de las observaciones peor clasificadas, son mayores y los pesos de las observaciones mejor clasificadas, son inferiores. Por otro lado se mira el clasificador en su conjunto y se le da una ponderación eligiéndose al final, por voto mayoritario, el más preciso (Galar et al., 2012).

De este algoritmo se han realizado diferentes versiones, utilizadas en distintas aplicaciones. Dentro de estas versiones, se pueden citar AdaBoost.M1 (Freund and Schapire, 1997), AdaBoost.M2 (Schapire and Singer, 1999); AdaCost (Fan et al., 1999); RareBoost (Joshi et al., 2001); AdaC1, AdaC2 y AdaC3 (Sun et al., 2007).

En la Figura 8.4 se detalla el algoritmo AdaBoost.M1 (Freund and Schapire, 1997) utilizado en los problemas de clasificación de clases no balanceadas [8] al ser el algoritmo AdaBoost más intuitivo en la clasificación binaria y el más conocido. En cada iteración, el peso o ponderación que se le asigna a cada observación depende de la base anterior, de manera que los pesos son actualizados en cada iteración dando mayor peso a las observaciones que mostraron un peor comportamiento en la iteración anterior.

El Boosting permite la selección de variables así como la selección del modelo que mejor se ajuste a cada covariable. A través del Boosting podemos predecir la variable respuesta a partir de un conjunto de predictores y para ello se realiza una combinación de diferentes bases (*base-learners*) de manera que obtengamos la mejor predicción de la variable respuesta en lugar de la que obtendríamos si consideráramos una única base.

[7] Para saber más sobre los errores en el algoritmo AdaBoost nos podemos remitir a Schapire and Singer (1999) y a Freund and Schapire (1997).

[8] Un problema de clasificación de clases no balanceadas, se entiende aquel que se caracteriza por tener una distribución muy distinta entre sus clases. Para profundizar en este tema, una importante referencia la encontramos en Galar et al. (2012).

Algoritmo AdaBoost.M1

(1) Inicializamos: Establecemos los pesos individuales w_i para las observaciones:

$$w_i^{[0]} = 1/n \qquad i = 1, \ldots, n$$

y fijamos $m = 0$.

(2) Computamos la Base: $m = m + 1$. Computamos la base para el conjunto de datos ponderados usando los pesos de la iteración anterior:

$$w_i^{[m]} y_i \overset{\text{base procedure}}{\rightarrow} \hat{g}^{[m]}(\boldsymbol{x}_i) \qquad \forall i$$

$$err^{[m]} = \frac{\sum_{i=1}^{n} w_i^{[m-1]} I\left\{y_i \neq \hat{g}^{[m]}(\boldsymbol{x}_i)\right\}}{\sum_{i=1}^{n} w_i^{[m-1]}}$$

En este paso, las observaciones mal clasificadas que tengan un peso w_i alto contribuyen más a la tasa de error que observaciones con un peso w_i bajo. Se calcula entonces el peso de cada iteración para el proceso final de agregación:

$$\alpha^{[m]} = \log\left(\frac{1-err^{[m]}}{err^{[m]}}\right)$$

y se actualizan los pesos:

$$\tilde{w}_i = w_i^{[m-1]} \exp\left(\alpha^{[m]} I\left\{y_i \neq \hat{g}^{[m]}(\boldsymbol{x}_i)\right\}\right)$$

Esto significa que si la observación i estaba mal clasificada, su peso w_i quedará multiplicado por el factor $\left(\frac{1-err^{[m]}}{err^{[m]}}\right)$ con lo que se consigue que su peso adquiera mayor relevancia en la siguiente iteración:

$$w_i^{[m]} = \frac{\tilde{w}_i}{\sum_{j=i}^{n} \tilde{w}_j}$$

(4) Iteramos: Repetimos los pasos **(2)** y **(3)** hasta que $m = m_{\text{stop}}$ y computamos el clasificador final para una nueva observación $\boldsymbol{x}_{new}$ a través del voto mayoritario ponderado:

$$\hat{f}_{AdaBoost}(\boldsymbol{x}_{new}) = \; argmin_{y \in \{0,1\}} \sum_{m=1}^{m_{\text{stop}}} \alpha^{[m]} I\left\{y = \hat{g}^{[m]}(\boldsymbol{x}_{new})\right\}$$

$$\alpha^{[m]} = \log\left(\frac{1-err^{[m]}}{err^{[m]}}\right) \begin{cases} > 0 & \text{para } err^{[m]} < 0.5 \\ = 0 & \text{para } err^{[m]} = 0.5 \\ < 0 & \text{para } err^{[m]} > 0.5 \end{cases}$$

Figura 8.4: *Algoritmo AdaBoost.M1.*

8.3. La relación del Boosting con la Estadística

En Breiman (1998) y Breiman (1999) se demuestra que el algoritmo AdaBoost puede ser utilizado en problemas de optimización en el espacio funcional, estableciendo de esta forma el nexo entre los algoritmos del aprendizaje automático y la Estadística al tomar como referencia la optimización numérica y la estimación estadística. Estos trabajos de Breiman permitieron mostrar posteriormente que el Boosting se puede interpretar como un *Functional Gradient Descent Algorithm*.

Breiman (1999) y Friedman et al. (2000) mostraron que AdaBoost es equivalente a *Forward Stagewise Additive Model*, al plantear un modelo como la combinación aditiva de clasificadores débiles y no como la combinación aditiva de covariables, demostrando a su vez que el AdaBoost es un método Newton de optimización que permite minimizar una particular función de pérdida de la familia exponencial.

Mason et al. (2000) paralelamente desarrollaron el algoritmo MarginBoost para elegir una combinación de clasificadores que optimice la media muestral de cualquier función de coste utilizando el *Gradient Descent* en el espacio funcional. En Friedman (2001) y Friedman (2002) se hace una extensión del trabajo de Friedman et al. (2000) y se demuestra que el Boosting puede interpretarse como un algoritmo gradiente descendente. Friedman (2002) tiene en cuenta una muestra aleatoria de observaciones, derivando así la formulación del método Boosting y los correspondientes modelos de lo que se denominó el Stochastic Gradient Boosting Machine.

Esta estructura generalmente se conoce como ***Functional Gradient Descent Boosting*** (Friedman, 2002); lo que hace es adaptar el conocido algoritmo Gradiente-Descendente, también conocido como *Steepest-Descent*, para obtener un procedimiento Boosting más general desde el punto de vista de la Estadística al que denomina *Gradient Boosting*. En este contexto, el procedimiento de aprendizaje lo que hace es consecutivamente ajustar nuevos modelos que provean la mejor estimación de la variable respuesta construyendo nuevas bases que estén máximamente correladas con el gradiente negativo de la función de pérdida (Natekin and Knoll, 2013).

El *Gradient Boosting* derivó en la implementación de diferentes variantes del original AdaBoost utilizando diferentes funciones de pérdida y siendo más accesibles a los análisis. Los algoritmos del *Gradient Boosting* permitieron utilizar reglas de predicción que tienen la misma interpretación que la alcanzada a través de la estimación de modelos en la Estadística.

Hastie et al. (2005) relacionan el Boosting con la estimación $\mathcal{L}_1$-penalizada. Bühlmann and Hothorn (2007) hacen un recorrido por la utilización estadística del Boosting donde se especifican los diferentes tipos de distribuciones que pueden ser utilizadas.

Bühlmann and Yu (2003) introdujeron el **Component-Wise Functional Gradient Descent Boosting** donde cada base depende sólo de un conjunto de posibles predictores y en cada iteración sólo se actualiza la mejor base lo que conduce a la selección de las bases y por lo tanto a la selección de las variables.

Ridgeway (2010) implementa el paquete de R ***gbm*** el cual tiene su origen en el trabajo de Ridgeway (1999) donde se utiliza el método de tree-based Boosting[9]. Bühlmann and Hothorn (2007); Hothorn et al. (2010) y Hothorn et al. (2015) desarrollaron el paquete ***mboost*** en R [10], con el que se ha implementado la estimación a través del *Component-Wise Gradient Descent Boosting* indistintamente para pequeñas y grandes dimensiones, utilizando bases para estimar modelos lineales y aditivos generalizados y modelos aditivos estructurados. Dentro de las bases pueden utilizarse modelos de regresión lineal univariantes, árboles de clasificación y regresión o splines penalizados.

8.3.1. Functional Gradient Descent (FGD) Boosting

En Breiman (1998) y Breiman (1999) se demuestra que el algoritmo AdaBoost se puede interpretar como un algoritmo gradiente-descendente en el espacio funcional, lo que es denominado *Functional Gradient Descent* (FDG).

El FDG Boosting que se debe a los trabajos de Friedman et al. (2000) y Friedman (2001) ha sido utilizado con éxito para estimar modelos de regresión aditivos generalizados, lo que le ha hecho ganar popularidad dentro de la estadística del aprendizaje.

El algoritmo *Gradient Boosting* (Friedman, 2002) se basa en querer estimar una función, de la que se tiene un sistema formado por una variable aleatoria respuesta y y un conjunto de variables explicativas o covariables $\boldsymbol{x} = \{x_1, \ldots, x_n\}$.

Teniendo la muestra de entrenamiento $\{y_i, \boldsymbol{x}\}$ del conjunto conocido de valores $(y_i, \boldsymbol{x})$ el objetivo es encontrar una función $F^*(\boldsymbol{x})$ que represente la influencia de $\boldsymbol{x}$ sobre y, tal que sobre la función de distribución conjunta de $(y, \boldsymbol{x})$, el valor esperado de la función de

[9] Muy útil para entender más sobre este paquete estadístico, el artículo de Ridgeway (2007).

[10] Hofner et al. (2014) realiza un detallado compendio de la utilización de este paquete estadístico en R y las ventajas que el mismo tiene.

pérdida especificada $\Psi\left(y, F\left(\boldsymbol{x}\right)\right)$ es minimizada

$$F^{*}\left(\boldsymbol{x}\right) = \underset{F(x)}{\operatorname{argmin}} E_{y,\boldsymbol{x}} \Psi\left(y, F\left(\boldsymbol{x}\right)\right) \tag{8.1}$$

A través del Boosting, se aproxima $F^{*}\left(\boldsymbol{x}\right)$ mediante una expansión "aditiva"de la forma:

$$F\left(\boldsymbol{x}\right) = \sum_{m=0}^{M} \beta_m h\left(\boldsymbol{x}; \boldsymbol{a}_m\right) \tag{8.2}$$

siendo la función $h\left(\boldsymbol{x}, \boldsymbol{a}\right)$ las bases (*base-learners*) que usualmente son seleccionadas como simples funciones de $\boldsymbol{x}$ con parámetros $\boldsymbol{a} = \{a_1, a_2, \cdots\}$. Los coeficientes de la expansión $\{\beta_m\}_0^M$ y los parámetros $\{a_m\}_0^M$ son conjuntamente estimados a través de un método por etapas (*stagewise methods*). Estableciendo valores iniciales $F_0\left(\boldsymbol{x}\right)$ y para $m = 1, 2, \ldots, M$

$$\left(\beta_m, \boldsymbol{a}_m\right) = \underset{\beta, \boldsymbol{a}}{\operatorname{argmin}} \sum_{i-1}^{N} \Psi\left(y_i, F_{m-1}\left(\boldsymbol{x}_i\right) + \beta h\left(\boldsymbol{x}_i; \boldsymbol{a}\right)\right) \tag{8.3}$$

y

$$F_m\left(\boldsymbol{x}\right) = F_{m-1}\left(\boldsymbol{x}\right) + \beta_m h\left(\boldsymbol{x}; \boldsymbol{a}_m\right) \tag{8.4}$$

Friedman (2001) a través del Gradient Boosting soluciona la Ecuación 8.3 en dos pasos:

Paso 1: Estima $h\left(\boldsymbol{x_i}; \boldsymbol{a}\right)$ por mínimos cuadrados:

$$\boldsymbol{a}_m = \underset{\boldsymbol{a}, \rho}{\operatorname{argmin}} \sum_{i=1}^{N} \left[\tilde{y}_{im} - \beta h\left(\boldsymbol{x}_i; \boldsymbol{a}\right)\right]^2$$

donde $\tilde{y}_{im} = -\left[\dfrac{\partial \Psi\left(y_i, F\left(\boldsymbol{x}_i\right)\right)}{\partial F\left(\boldsymbol{x}_i\right)}\right]_{F(\boldsymbol{x}) = F_{m-1}(\boldsymbol{x})}$

es el gradiente negativo de la función de pérdida.

Paso 2: Dado $h\left(\boldsymbol{x_i}; \boldsymbol{a}_m\right)$, el valor óptimo para los coeficientes β_m es obtenido mediante

$$\beta_m = \underset{\beta}{\operatorname{argmin}} \sum_{i=1}^{N} \Psi\left(y_i, F_{m-1}\left(\boldsymbol{x}_i\right) + \beta h\left(\boldsymbol{x}_i; \boldsymbol{a}_m\right)\right)$$

El objetivo, en definitiva, de este algoritmo es modelar la relación entre una variable respuesta $\boldsymbol{y}$ y un conjunto de covariables $\boldsymbol{x} = \{x_1, \ldots, x_n\}$ de manera que se obtenga una predicción "óptima"de y dado $\boldsymbol{x}$. La función de pérdida Ψ se asume diferenciable y convexa con respecto a x. La Figura 8.5 nos muestra la manera en la que funciona este algoritmo[11].

Se pueden considerar diferentes funciones de pérdida. La función de pérdida puede ser especificada teniendo en cuenta el tipo de variable respuesta. Natekin and Knoll (2013) hacen un resumen de las funciones de pérdidas que se pueden utilizar dependiendo de la familia de la variable respuesta, a saber:

(1) Variable respuesta continua, $y \in \mathbb{R}$:

- Función de Pérdida L_2 Gausiana
- Función de pérdida L_1 Laplace
- Función de pérdida Hube, con δ especificada
- Función de pérdida cuantílica, con α especificada

(2) Variable respuesta categórica, $y \in (0, 1)$:

- Función de Pérdida Binomial
- Función de Pérdida AdaBoost

(3) Otras familias de variables respuestas:

- Función de Pérdida para Modelos de Supervivencia
- Función de pérdida para datos de recuento

Si comparamos los algoritmos AdaBoost y FGD observamos que en éste último se cumplen los aspectos fundamentales del algoritmo AdaBoost, lo que se resume en la Tabla 8.1

[11] La descripción de los pasos a seguir en el Functional Descent Gradient Boosting así como la notación empleada ha sido tomada de Bühlmann and Hothorn (2007).

Functional Gradient Descent Algorithm

(1) Inicializamos: Establecemos $m = 0$. Inicializamos la función a estimar $\hat{f}^{[0]}(\cdot)$ con un valor offset que puede una constante c que minimiza la función de pérdida:

$$\hat{f}^{[0]}(\cdot) \equiv \text{argmin}_c \frac{1}{n} \sum_{i=1}^{n} \rho(y_i, c)$$

Si estaríamos en el caso del modelo de regresión lineal y por tanto la pérdida sería cuadrática, $c = \bar{y}$. Un valor de desplazamiento (offset value) que se utiliza especialmente para datos centrados es:

$$\hat{f}^{[0]}(\cdot) \equiv 0$$

(2) Gradiente negativo: Incrementamos m en 1; $m = m + 1$. Obtenemos el gradiente negativo de la función de pérdida $\rho(y_i, f)$ evaluada en la estimación obtenida en la iteración anterior $\hat{f}^{[m-1]}(\boldsymbol{x}_i)$:

$$\mathrm{u}_i^{[m]} = -\frac{\partial \rho(y_i, f)}{\partial f}_{f=\hat{f}^{[m-1]}(\boldsymbol{x}_i)} \qquad i = 1, \ldots, n$$

(3) Estimamos: Se ajusta el vector gradiente negativo $\boldsymbol{u}^{[m]} = \left(u_1^{[m]}, \ldots, u_n^{[m]}\right)$ usando el procedimiento base (por ejemplo, la regresión) para $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$:

$$\mathrm{u}_i^{[m]} \overset{\text{base procedure}}{\rightarrow} \hat{g}^{[m]}(\boldsymbol{x}_i) \qquad \forall i$$

$\hat{g}^{[m]}(\boldsymbol{x}_i)$ no estima el valor de y_i sino que estima el valor del gradiente negativo para la observación i, por ello denota la función que aproxima el vector gradiente negativo $\boldsymbol{u}^{[m]}$.

(4) Actualizamos: Se actualiza la estimación de la función $\hat{f}^{[m]}(\cdot)$ utilizando el factor de longitud de paso ν siendo $0 < \nu \leq 1$:

$$\hat{f}^{[m]}(\cdot) = \hat{f}^{[m-1]}(\cdot) + \nu \cdot \hat{g}^{[m]}(\cdot)$$

Figura 8.5: *Functional Gradient Descent Algorithm.*

Algoritmo AdaBoost	**Algoritmo FGD**
1.1 En cada iteración, los pesos de cada observación son actualizados.	1.1 En cada iteración, cada observación (x_i) recibe una ponderación individual dependiendo de la iteración anterior.
1.2 Aquellas observaciones que han sido peor clasificadas, adquieren un mayor peso.	1.2 Cada y_i es sustituido por el gradiente $u_i^{[m]}$ el cual es la derivada negativa de la función de pérdida evaluada en y_i y la función f estimada en la iteración anterior, o sea $f = \hat{f}^{[m-]}(\boldsymbol{x}_i)$.
	1.3 De esta forma, las observaciones que han dado mayores problemas en esa iteración anterior, son los que adquieren mayor importancia.
1.3 Al final del proceso, las bases que mostraron los mejores resultados serán las que jueguen un papel principal en la predicción final.	1.4 En cada iteración cuando actualizamos (Paso 4) lo que estamos haciendo es asegurarnos de que las mejores bases serán las más importantes en la predicción final
	1.5 Sólo el base procedure $\hat{g}^{[m]}(\cdot)$ que sea una buena aproximación del efecto de $\boldsymbol{X}$ sobre $\boldsymbol{Y}$ ofrecerá los valores más altos y por tanto contribuirá al descenso más marcado de la función de pérdida.

Tabla 8.1: *Comparación entre los algoritmos AdaBoost y FGD .*

8.3.2. Component-wise FGD Boosting

Bühlmann and Yu (2003) desarrollaron el **Component-Wise Functional Gradient Descent Boosting** a partir del algoritmo *Functional Gradient Descent* (FGD) *Boosting*. La principal diferencia con el FGD es que usan diferentes bases (n_B) en lugar de una sola. Se puede especificar una base para cada una de las covariables, incluido el intercepto siendo en este caso el número de bases igual al número de covariables o podemos especificar también una base que incluya varias covariables.

En cada iteración se selecciona una base, lo que conlleva a que se produzca la selección de las variables una vez que se han alcanzado el número óptimo de iteraciones. Esto se traduce en que no todas las bases y por tanto, no todas las variables serán seleccionadas para el ajuste final. La Figura 8.6 explica en detalle este algoritmo.[12]

Un parámetro importante en este algoritmo es el número total de iteraciones (m_{stop}), valores altos para m_{stop} conducen a un modelo con un sesgo pequeño en el efecto estimado de las covariables pero con una varianza grande. En cambio valores pequeños para m_{stop} conducen a modelos con una varianza pequeña y un sesgo grande (Mayr et al., 2012a).

Para seleccionar la m_{stop} óptimo se han utilizado en la literatura métodos de remuestreo o criterios de información como el Criterio de Información de Akaike (AIC) y el Criterio Bayesiano de Información (BIC) o Métodos de Validación Cruzada.

Mayr et al. (2012a) demostraron que el método AIC tiende a encontrar un valor demasiado grande para la iteración m_{stop} y plantean que además este criterio tiene como inconvenientes que es necesario conocer a priori un valor inicial (muy grande) para m_{stop}, siendo computacionalmente inefectivo porque en muchos casos el número de iteraciones óptimo es más pequeño que el valor planteado como valor inicial.

Hofner et al. (2014) proponen utilizar el K-Fold-Cross Validation, método que evita el sesgo en la estimación de los grados de libertad del modelo.

[12] La descripción de los pasos a seguir en el algoritmo Component-Wise FDG Boosting así como la notación empleada ha sido tomada de Bühlmann and Hothorn (2007) y de Hofner et al. (2014).

Component-wise FGD Boosting

(1) Inicializamos: Establecemos $m = 0$. Inicializamos la función a estimar y el predictor aditivo con $\hat{f}_j^{[0]}(\cdot) \equiv 0$ para $j = 1, \ldots, n_B$ que puede ser una constante c que minimiza la función de pérdida:

$$\hat{\eta}^{[0]}(\cdot) = \text{argmin}_c \frac{1}{n} \sum_{i=1}^{n} \rho(y_i, c)$$

(2) Gradiente negativo: Incrementamos m en 1: $m = m + 1$. Obtenemos el gradiente negativo de la de pérdida $\rho(y_i, f)$ evaluada en la estimación obtenida en la iteración anterior $\hat{\eta}^{[m-1]}(\boldsymbol{x}_i)$:

$$\mathrm{u}_i^{[m]} = -\frac{\partial \rho(y_i, \eta)}{\partial \eta}_{\eta = \hat{\eta}^{[m-1]}(\boldsymbol{x}_i)}, \qquad i = 1, \ldots, n$$

(3) Estimamos: Se ajusta el vector gradiente negativo $\boldsymbol{u}^{[m]} = \left(u_1^{[m]}, \ldots, u_n^{[m]}\right)$ usando la base $g_j(\boldsymbol{x}) \, \forall j \in \{1, \ldots, n_B\}$. Cada base g_j depende de un conjunto de covariables o de una sola covariable:

$$\mathrm{u}_i^{[m]} \overset{\text{base procedure}}{\rightarrow} \hat{g}^{[m]}(\boldsymbol{x}_i) \;\; \forall i$$

Se selecciona el mejor ajuste para el base-learner g_{j*}, es decir, el base-learner que minimiza la suma de cuadradros de los residuos:

$$\mathrm{g}_{j*} = argmin_j \sum_{i=1}^{n} \left(u_i^{[m]} - \hat{g}_j^{[m]}(\boldsymbol{x}_i^{\intercal})\right)^2$$

(4) Actualizamos: Se actualiza la estimación del predictor aditivo η y de la función $\hat{f}_{j*}^{[m]}(\cdot)$ utilizando el factor ν (factor de longitud de paso) siendo $0 < \nu \leq 1$:

$$\hat{\eta}^{[m]}(\cdot) = \hat{\eta}^{[m-1]}(\cdot) + \nu \cdot \hat{g}_{j*}^{[m]}(\cdot)$$
$$\hat{f}_{j*}^{[m]}(\cdot) = \hat{f}_{j*}^{[m-1]}(\cdot) + \nu \cdot \hat{g}_{j*}^{[m]}(\cdot)$$

(5) Finalizamos: Si $m = m_{stop}$ paramos, para un m_{stop} previamente establecido.

Figura 8.6: *Component-wise FGD Boosting.*

8.4. Utilización del Boosting en los Modelos Cuantílicos Aditivos Estructurados (STAQ)

Una de las principales razones que nos llevan a especificar un modelo, es cuantificar la influencia de las covariables en la variable respuesta. En ello juega un papel importante la selección de variables, o sea debemos determinar cuáles son realmente las covariables que influyen de manera significativa en la variable respuesta y también cuál es la forma de esa influencia lo que se traduce en que no sólo seleccionamos variables sino también el modelo.

En este objetivo, tiene un específico e importante papel en los modelos STAQ, el Boosting. El Boosting en los modelos STAQ, también conocido como Quantile Boosting fue introducido por Fenske et al. (2012) basado en el trabajo de Kneib et al. (2009) siendo una utilización innovadora de los algoritmos de aprendizaje aplicados en la Estadística al abrir una puerta alternativa a la estimación de los modelos de regresión cuantílica aditivos.

Esta forma de ver la regresión cuantílica se enmarca dentro de los algoritmos del Aprendizaje en Estadística sin tener que establecer supuestos sobre la distribución de probabilidades de la variable respuesta o el término de error. A este tipo de algoritmos se les conoce como *Distribution-free Statistical Learning Algorithms*.

Langford et al. (2012) propusieron el algoritmo Quanting con el que intentaron resolver el problema de la clasificación vía la selección de predictores en la regresión cuantílica pero la principal desventaja de esta propuesta comparado con el Boosting es que computacionalmente requiere mucho tiempo.

Leathwick et al. (2006) y Elith and Leathwick (2009) utilizaron la técnica del *Boosted Regression Tree*s (BRT) la cual difiere de los tradicionales métodos de regresión en que utiliza la técnica del Boosting para combinar un número importante de modelos de árboles (*tree models*) simples. Kriegler and Berk (2010) también utilizaron el Boosting empleando árboles como bases. Esta técnica tiene la desventaja de que no resulta sencillo cuantificar la influencia de una determinada covariable en la variable respuesta.

Retomando la Ecuación 7.21 donde definimos un predictor aditivo estructurado, tenemos que para algunas covariables vamos a predecir su efecto sobre la variable respuesta a partir de un efecto lineal $h_d(\boldsymbol{x}_i) = \beta_d x_{il}$ teniendo que estimar el coeficiente β_d y para otras covariables vamos a predecir su efecto a través de una relación no lineal $h_d(\boldsymbol{x}_i) = f_d(x_{il})$ a partir de funciones suaves f_d que tenemos que estimar.

Si se utiliza una apropiada base para cada una de las covariables, la regresión cuantílica

a través del Boosting permitirá seleccionar los predictores que más información brindan dentro del conjunto de predictores que tengamos incluidos en el modelo, lo cual es heredado de la técnica AdaBoost.

En un modelo lineal generalizado donde

$$E(y|\boldsymbol{x}) = h(\eta(\boldsymbol{x})) \tag{8.5}$$

con variable respuesta y, función de respuesta h y predictor lineal η como habíamos ya visto en la Sección 3.2 del Capítulo 3; el predictor aditivo se puede definir como

$$\eta\left(\boldsymbol{x}\right) = \beta_0 + \sum_{i=1}^{n} \beta_i x_i \qquad i = 1, \ldots, n \tag{8.6}$$

con vector de covariables $\boldsymbol{x}$ y sus correspondientes efectos β_i. En estos modelos lo que hacemos es minimizar la función de pérdida esperada $E(\rho(y, \eta(\boldsymbol{x})))$ con una apropiada función de pérdida $\rho(y, \eta(\boldsymbol{x}))$ la cual es previamente definida y corresponde a la familia exponencial. En los modelos GLM, se puede maximizar el logaritmo de la función de verosimilitud o minimizar el valor negativo del logaritmo de la función de verosimilitud

$$\rho(y, \eta(\boldsymbol{x})) = -ylog(P(y = 1|\eta(\boldsymbol{x}))) + (1 - y)log(1 - P(y = 1|\eta(\boldsymbol{x}))) \tag{8.7}$$

como una función de pérdida.

A través del Boosting como algoritmo de optimización, se minimiza una determinada función de pérdida y paso a paso se va actualizando el estimador dependiendo del gradiente descendente más pronunciado para la función de pérdida.

En la práctica lo que en realidad se minimiza no es la función de pérdida esperada (*expected loss function*) sino que lo que hacemos es optimizar la función de riesgo empírica (*empirical risk function*) a través del *Component-wise Functional Gradient Descent Boosting*:

$$\mathcal{R}\left(\boldsymbol{y}, \boldsymbol{x}\right) = n^{-1} \sum_{i=1}^{n} \rho\left(y_i, \eta\left(x_i\right)\right) \tag{8.8}$$

siendo $\boldsymbol{y} = (y_1, \ldots, y_n)$ y $\boldsymbol{x} = (x_1, \ldots, x_n)$.

Koenker(2011) apunta que utilizar esta técnica en la regresión cuantílica provee una

aproximación natural a la selección del modelo (*model selection*) y dada una muestra relativamente grande tiene ventajas computacionales.

Si la relación que existe entre los regresores y la variable respuesta no es lineal, es necesario aportar más flexibilidad al modelo y esto es posible a través de la regresión cuantílica aditiva y siguiendo lo que habíamos planteado en la Ecuación 7.20 y en la Ecuación 7.21:

$$Q_\tau\left(\boldsymbol{y}/x_i, z_i\right) = x_i'\beta_\tau + \sum_{j=1}^{q} f_{\tau j}\left(z_{ij}\right) \tag{8.9}$$

En esta ecuación hemos delimitado una parte lineal $x_i'\beta_\tau$ que incluye al intercepto y donde podemos incluir a las variables categóricas o ficticias y una parte no lineal $\sum_{j=1}^{q} f_{\tau j}\left(z_{ij}\right)$ donde incluimos a los regresores continuos y donde las $f_{\tau j}$ son las funciones suaves de esas variables continuas. Con el Boosting no vamos a obtener un ajuste de un modelo aditivo en las covariables, sino que el Boosting es una combinación aditiva y lineal de esas bases que tienen el papel de ser estimadores simples.

Una base (*base learner*) puede ser cualquier tipo de herramienta de regresión donde la variable respuesta es modelada por una o más covariables:

$$\text{covariable(s)} \overset{\text{base learner}}{\rightarrow} \text{predicción de la respuesta}$$

Existen dos grandes grupos en los que se agrupan las bases que son: P-Splines y B-Splines, existiendo otras alternativas como los *Thin Plate Regression Splines*. Una ventaja de utilizar P-splines en la regresión cuantílica Boosting es que se cuenta con una buena herramienta para estimar intervalos de predicción basados en los cuantiles. Los P-splines fueron introducidos por Eilers and Marx (1996) y en los modelos aditivos fueron utilizados por primera vez por Schmid and Hothorn (2008).

Utilizando P-splines como bases en la regresión cuantílica aditiva se deben especificar los grados de libertad ($df(\lambda_d)$) de las bases. Bühlmann and Yu (2003) afirman que las bases deben tener un sesgo grande y una varianza pequeña recomendando un valor establecido de grados de libertad para las bases de cada variable que se incluya como regresor en el STAQ. Es de señalar que en el algoritmo Boosting, el parámetro de suavidad λ_d de las bases penalizadas $d = 1, \ldots, D$ no son tratadas como parámetros que deben ser optimizados y esta es una de las diferencias del algoritmo Boosting con otros algoritmos.

También debemos especificar el número de nodos equidistantes. Schmid and Hothorn (2008) demuestran que deben estar entre 20 y 50 para alcanzar un buen ajuste, en el

paquete *mboost* este número por defecto es de 20. También es necesario seleccionar el factor de longitud de paso $\nu \in (0, 1)$ siendo el único requerimiento que el mismo sea pequeño, por ejemplo $\nu = 1$ para garantizar que no se exceda el mínimo del riesgo empírico $\mathcal{R}$ (Schmid and Hothorn, 2008) y evitar el esfuerzo computacional.

En la Figura 8.7 se describe cómo funciona el algoritmo *Component-wise FGD Boosting* en los modelos cuantílicos aditivos.

Component-wise FGD Boosting en la Regresión Cuantílica

(1) Inicializamos: Establecemos $m = 0$. Inicializamos el predictor aditivo y la función a estimar con los valores iniciales que usualmente puede ser la mediana de la variable respuesta

$$\hat{\eta}_i^{[0]} = \text{argmin}_c \sum_{i=1}^{n} \rho_{0.5}\left(y_i - c\right) \text{ y } \hat{\boldsymbol{h}}_j^{[0]} = \mathbf{0} \qquad j = 1, \ldots, n_B.$$

(2) Gradiente negativo: Incrementamos m en 1; $m = m + 1$. Obtenemos los residuos gradiente de la función de pérdida evaluada en la estimación obtenida en la iteración anterior $\hat{\eta}^{[m-1]}$:

$$\mathrm{u}_i^{[m]} = -\frac{\partial \rho(y_i, \eta)}{\partial \eta}_{\eta = \hat{\eta}^{[m-1]}} \qquad _{i=1,\ldots,n}$$

En el caso del Boosting en la regresión cuantílica, tenemos que insertar la check function para la función de pérdida y con ello obtenemos los residuos gradientes negativos:

$$\mathrm{u}_i^{[m]} = -\rho_\tau^{'}\left(y_i - \hat{\eta}_i^{[m-1]}\right) = \begin{cases} \tau & y_i - \hat{\eta}_i^{[m-1]} \geq 0 \\ \tau - 1 & y_i - \hat{\eta}_i^{[m-1]} < 0 \end{cases}$$

(3) Estimación: Ajustamos el vector gradiente negativo $\boldsymbol{u}^{[m]} = \left(u_1^{[m]}, \ldots, u_n^{[m]}\right)$ para cada una de las bases y así obtenemos el mejor ajuste para la base $\eth_{j*}$ que minimiza la L_2

$$\mathrm{j}^* = \text{argmin}_j \left[\left(\boldsymbol{u}^{[m]} - \hat{\boldsymbol{g}}_j^{[m]}\right)^{'} \left(\boldsymbol{u}^{[m]} - \hat{\boldsymbol{g}}_j^{[m]}\right)\right]$$

donde $\boldsymbol{u}^{[m]} = \left(u_1^{[m]}, \ldots, u_n^{[m]}\right)^{'}$ es el vector de residuos gradientes de la presente iteración.

(4) Actualización: Se actualiza la mejor base

$$\hat{\boldsymbol{h}}_{j^*}^{[m]} = \hat{\boldsymbol{h}}_{j^*}^{[m-1]} + \nu \cdot \hat{\boldsymbol{g}}_{j^*}^{[m]}$$

donde $\nu \in (0, 1]$ es el factor de longitud de paso establecido.
Se mantienen los demás efectos constantes, es decir $\hat{\boldsymbol{h}}_j^{[m]} = \hat{\boldsymbol{h}}_j^{[m-1]}$ para todo $j \neq j^*$ y computamos y actualizamos el predictor $\hat{\eta}_i^{[m]}$ para todo $i = 1, \ldots, n$.

(5) Finalizamos: Si $m = m_{stop}$ paramos, para un m_{stop} dado.

Figura 8.7: *Component-wise FGD Boosting en la Regresión Cuantílica.*

CAPÍTULO 9

Un modelo STAQ estimado por Boosting para estudiar el salario de los inmigrantes cubanos en Estados Unidos

9.1. Introducción

La regresión cuantílica (Koenker and Bassett, 1978) ha ido ganando terreno en el análisis de la relación entre una variable respuesta y un conjunto de covariables debido a su robutez ante la presencia de outliers o valores extremos y a la posibilidad que ofrece de analizar cómo varían los cuantiles condicionados de la variable respuesta ante el conjunto de covariables incluidas en el estudio.

Su utilización abarca campos como la Medicina, en los que podemos encontrar trabajos como los artículos de Petscher and Logan (2013), Briollais and Durrieu (2014), Hofner et al. (2015) y Sun et al. (2016); las Ciencias Sociales con aplicaciones como las de Gaglianone et al. (2011), Wong et al. (2016) y Autor et al. (2017); el estudio del Medio Ambiente con ejemplos como los trabajos de Cade (2011), Phelan et al. (2017) y Chen et al. (2017), entre otros.

Otra de las razones de la extensión en el uso de la regresión cuantílica es el hecho de no se puede asumir que el efecto de las covariables sea el mismo en cada uno de los cuantiles de la distribución de la variable respuesta. De ahí, que los coeficientes estimados a través de la regresión cuantílica ofrecen una información adicional que no es posible obtener a partir de la regresión con respecto a la media.

Tampoco se puede asumir que ese efecto tiene que ser necesariamente lineal, lo que permitiría introducir en el modelo de regresión cuantílica cierto grado de flexibilidad.

Fenske et al. (2012) propusieron utilizar un modelo de regresión cuantílica aditivo estructurado (STAQ) utilizando el Component-wise FGD Boosting (Bühlmann and Yu, 2003) en la estimación, basándose en el trabajo de Kneib et al. (2009). Estos modelos tienen la ventaja de que permiten especificar el efecto de las diferentes covariables de manera flexible, lo que se traduce en poder recoger la relación entre las covariables y la variable respuesta de una forma lineal o no lineal a partir del uso de funciones suaves como los P-splines.

El Boosting es un algoritmo de optimización que permite minimizar la función de pérdida esperada de acuerdo al criterio del gradiente descendente. Para ello se utilizan funciones de regresión simples como bases y de forma iterativa se van ajustando una a una al gradiente negativo de la función de pérdida. En cada iteración sólo el mejor ajuste de la base es incluido en el modelo, lo que conduce a la selección de variables.

Este algoritmo, tomado del Aprendizaje Automático, permite la selección de variables en el marco de modelos de alta dimensionalidad y es una útil herramienta estadística para el análisis de datos (Bühlmann et al., 2014).

Provee además una herramienta para la selección del modelo (Koenker, 2011). Kriegler and Berk (2010) ya usaron antes el Boosting en la regresión cuantílica pero no a través de un modelo aditivo estructurado sino mediante la utilización de árboles de regresión, lo que no permitía la cuantificación de la influencia parcial de cada covariable sobre el cuantil condicionado como se puede obtener a partir de un modelo STAQ.

La definición de la función de pérdida define el tipo de regresión, así si se establece una función de pérdida cuadrática L_2 estaríamos ante la regresión clásica con respecto a la media, en cambio la función de pérdida L_1 para la regresión con respecto a la mediana, puede ser extendia a los demás cuantiles a través de la denominada *check-function*.

Como plantean Fenske et al. (2012) y Bühlmann et al. (2014), las ventajas de la utilización del Boosting en un modelo STAQ quedan reflejadas en que: (i) la estimación de parámetros y la selección de variables es parte intrínseca de este procedimiento; (ii) los modelos de regresión cuantílica aditiva son mayoritariamente estimados a través de la programación lineal. En el caso del Boosting, se incrementa la flexibilidad en la estimación de efectos no lineales porque la especificación de la diferenciabilidad de los efectos no lineales es parte del modelo y no del método de estimación; (iii) se pueden estimar modelos complejos donde se combinan los efectos lineales con los no lineales y con términos de coeficientes cambiantes;

(iv) la inferencia puede ser obtenida a partir del criterio de selección de estabilidad de Meinshausen and Bühlmann (2010) y Shah and Samworth (2012), lo que permite aumentar el rendimiento de este algoritmo, basándose en la agregación de los resultados de aplicar un procedimiento de selección a submuestras de los datos.

Si bien existen aplicaciones de la utilización del Boosting en la estimación de modelos STAQ en el campo de la Bioestadística, como muestran los trabajos de Fenske et al. (2013a), Fenske et al. (2013b), Hofner et al. (2015) y Mayr et al. (2016) así como en el campo de las Finanzas como los trabajos de Demirer et al. (2016) y Pierdzioch et al. (2016); no existen trabajos dedicados al análisis de los salarios en los que se haya utilizado esta técnica a pesar de las grandes ventajas que para la estimación y el análisis posee. Por tanto, este trabajo es el primero, para nuestro conocimiento, en que se utiliza el Boosting en la estimación de un modelo STAQ en el análisis de los salarios.

En la siguiente sección se explica brevemente la metodología utilizada. En otra sección se describen los datos utilizados y se muestran los resultados obtenidos para exponer las principales conclusiones en una última sección.

9.2. Implementación

Para estudiar el impacto de las diferentes variables socioeconómicas en el salario por hora de los inmigrantes cubanos en Estados Unidos en un contexto de regresión cuantílica más flexible, vamos a especificar el siguiente predictor aditivo estructurado:

$$Q_{\boldsymbol{y}_i}(\tau|\boldsymbol{x}_i) = \eta_{\tau i} = \beta_{\tau 0} + \beta_{\tau 1} x_{i1} + \cdots + \beta_{\tau 6} x_{i6} + f_{\tau 1}(z_{i1}) + f_{\tau 2}(z_{i2}) + x_{i6} \cdot g_{\tau,1}(z_{i1}) + \varepsilon_{\tau i} \tag{9.1}$$

Siendo $x_{i1}, \ldots, x_{i6}$ para $i = 1, \ldots, n$ el conjunto de variables categóricas donde se incluirán variables ficticias para indicar si el individuo es mujer, si vive en el Estado de Florida, si no es blanco, si está casado, si tiene la nacionalidad norteamericana y si domina el idioma inglés. El conjunto de variables continuas se representa a través de las variables z_{i1}, z_{i2}, donde vamos a tener en cuenta la edad de individuo en el momento de emigrar y los años de educación. La variable dependiente y_i es el (log) del salario por hora.

Se utilizarán bases lineales para expresar el efecto lineal de las variables categóricas sobre la variable respuesta y este efecto lineal será recogido en el vector $\beta_{\tau 1}, \ldots, \beta_{\tau 6}$. Para las

variables continuas recogeremos el efecto lineal con la utilización de bases lineales y el posible efecto no lineal a través de las funciones (potencialmente) no lineales $f_{\tau 1}, f_{\tau 2}$ suaves.

La función $g_{\tau,1}(z_{i1})$ (potencialmente no lineal) mostrará el efecto de la edad en el momento de la emigración teniendo en cuenta el hecho de que el individuo domine el idioma inglés en el momento de realizarse la encuesta.

Para la estimación del modelo se ha utilizado el paquete ***mboost*** de R (Hothorn et al., 2015) y dentro de éste la función `gamboost` con la opción `family = QuantReg()`. Se han definido diferentes bases para cada uno de los regresores de manera que se pueda recoger el efecto ya sea lineal o no de los mismos y una base para el intercepto, lo que ha conllevado a que hemos quitado en cada base la posibilidad de contar con su propio intercepto.

A la hora de estimar con el algoritmo Boosting las variables continuas incluidas en el modelo podrían ser seleccionadas para entrar en él: (i) como un efecto lineal, (ii) como un efecto suave centrado alrededor de cero, (iii) como una combinación de un efecto lineal y una desviación suave, o (iv) podrían no ser seleccionadas para entrar en el modelo.

Para las variables categóricas usamos la función `bols()` para especificar las bases lineales y para las variables continuas definimos una base con la función `bols()` para medir el efecto lineal y otra base con la función `bbs()` para medir el efecto no lineal.

Las funciones `bols()` permiten definir una base de mínimos cuadrados ordinarios penalizados. Ejemplos de este tipo de bases son (a) los efectos lineales, (b) los efectos categóricos, (c) los efectos lineales de grupos de variables $\boldsymbol{x} = (x_1, \ldots, x_n)'$, (d) efecto *ridge*-penalizado para los casos (b) y (c), (e) término de coeficientes cambiantes como las interacciones entre variables.

Las funciones `bbs()` permiten definir efectos suaves a través de P-splines. Ejemplos de este tipo de bases son (a) efectos suaves, (b) efectos suaves bivariantes como los efectos espaciales, (c) término de coeficientes cambiantes, (d) efectos cíclicos, entre otros (Hofner et al., 2014).

La interacción entre la edad del individuo en el momento de emigrar y el nivel de inglés, que en el modelo viene expresada como $x_{i6} \cdot g_{\tau,1}(z_{i1})$ se ha especificado usando `bbs(x, by = z)` teniendo por tanto sólo en cuenta un efecto suave siendo x la edad en el momento de emigrar y llamado el efecto modificador del nivel de inglés z.

9.3. Datos y Resultados

En la estimación del modelo especificado hemos utilizado datos de sección cruzada repetida entre los años 2000-2007. Los datos han sido obtenidos de una muestra aleatoria del 1 % de la American Community Survey (ACS) facilitada por Integrated Public Use Microdata Series (IPUMS) (Ruggles et al., 2013) de la Universidad de Minnesota. Se cuenta con una muestra de 15.131 individuos que han emigrado a Estados Unidos entre los años 1962 y 2007.

A la hora de seleccionar las variables incluidas como regresores, se han tomado en consideración criterios similares a los asumidos en la estimación a través de la regresión cuantílica lineal y la regresión cuantílica lineal con selección muestral en cuanto a la edad del individuo en el momento de la emigración, la edad en el momento de la realización de la encuesta y las horas de trabajo a la semana, lo que significa que hemos restringido nuestra muestra a individuos con edades comprendidas entre los 25 y los 55 años, que trabajan 60 horas o menos a la semana y que tenían entre 17 y 49 años en el momento de llegar a Estados Unidos.

Así mismo, se ha sustituido la variable *Is Black* que toma el valor 1 si el individuo es de la raza negra por la variable *Is Not White* que toma el valor 1 si el individuo no es blanco, debido a que algunos individuos no declaran ser de raza negra; se ha considerado no incluir la variable Experiencia Potencial y se ha adicionado una variable ficticia para indicar si el individuo vive en el Estado de la Florida, característica que no hemos tenido en cuenta en los modelos de regresión cuantílica lineal.

En la Tabla 9.1 se describen las variables tenidas en cuenta en el modelo especificado. El 42 % de la muestra es mujer, un 72 % vive en el Estado de la Florida, un 13 % no es blanco, lo que contrasta con el porcentaje de individuos que se declararon negros en las muestras consideradas en las estimaciones de los capítulos anteriores donde sólo el 3 % era de raza negra; el 61 % de la muestra está casado, un 33 % tiene la nacionalidad americana en el momento de realizar la encuesta, un 50 % domina el inglés; la media de edad en el momento de arribar a Estados Unidos es de 29.73 años, la media de años de estudio es de 12.57 años y la media del (log) de salario por hora es de 2.391 lo que representa aproximadamente un salario medio por hora de 15.95 dólares.

Variable	*Description*	*Mean*	*Standard Deviation*
Is Woman	Dummy variable: 1 si el individuo es mujer	0.4169	0.4931
Live in Florida	Dummy variable: 1 si el individuo vive en Florida	0.7123	0.4194
Is Not White	Dummy variable: 1 si no es blanco	0.1269	0.3329
Is Married	Dummy variable: 1 si está casado	0.6084	0.4881
Is American Citizen	Dummy variable: 1 si tiene la nacionalidad americana	0.3287	0.4698
English Proficiency	Dummy variable: 1 si domina el Inglés	0.5000	0.5000
Age at Migration	Edad del individuo en el momento de la migración	29.7300	7.8931
Years of Education	Años de educación del individuo	12.5700	2.8464
(log) Hourly Earnings	Logaritmo del salario por hora. Variable Respuesta	2.3910	0.7241

Tabla 9.1: *Variables en el Predictor Aditivo Estructurado.*

Se han considerado diferentes cuantiles $\tau = (0.05, 0.10, 0.25, 0.50, 0.75, 0.80, 0.85, 0.90, 0.95)$ con el objetivo de analizar la forma en que las diferentes covariables se relacionan con el (log) del salario por hora en diferentes locaciones de la distribución de probabilidades de la variable respuesta y poder comparar a través de los diferentes cuantiles, prestando especial atención a la comparación entre los individuos que menos y que más salario ganan.

En las Figura 9.1 aparece el Diagrama de Dispersión entre el (log) del salario por hora y las covariables continuas del modelo y en la Figura 9.3 se recoge la relación entre estas variables por medio de una Matriz de Diagramas de Dispersión. En ambos gráficos no puede concluirse una clara relación lineal entre la variable (log) del salario por hora y las covariables continuas en el modelo lo que justifica el intento de estimar la función cuantílica de manera flexible teniendo en cuenta que la regresión cuantílica lineal no tendría que ser suficiente para expresar de manera adecuada la relación entre las covariables y la variable respuesta.

Previamente todas las covariables continuas han sido centradas con respecto a su media. Es necesario centrar los valores con respecto a la media al utilizar bases sin intercepto. Si no centramos las covariables con respecto a su media, esa base no recogería el verdadero

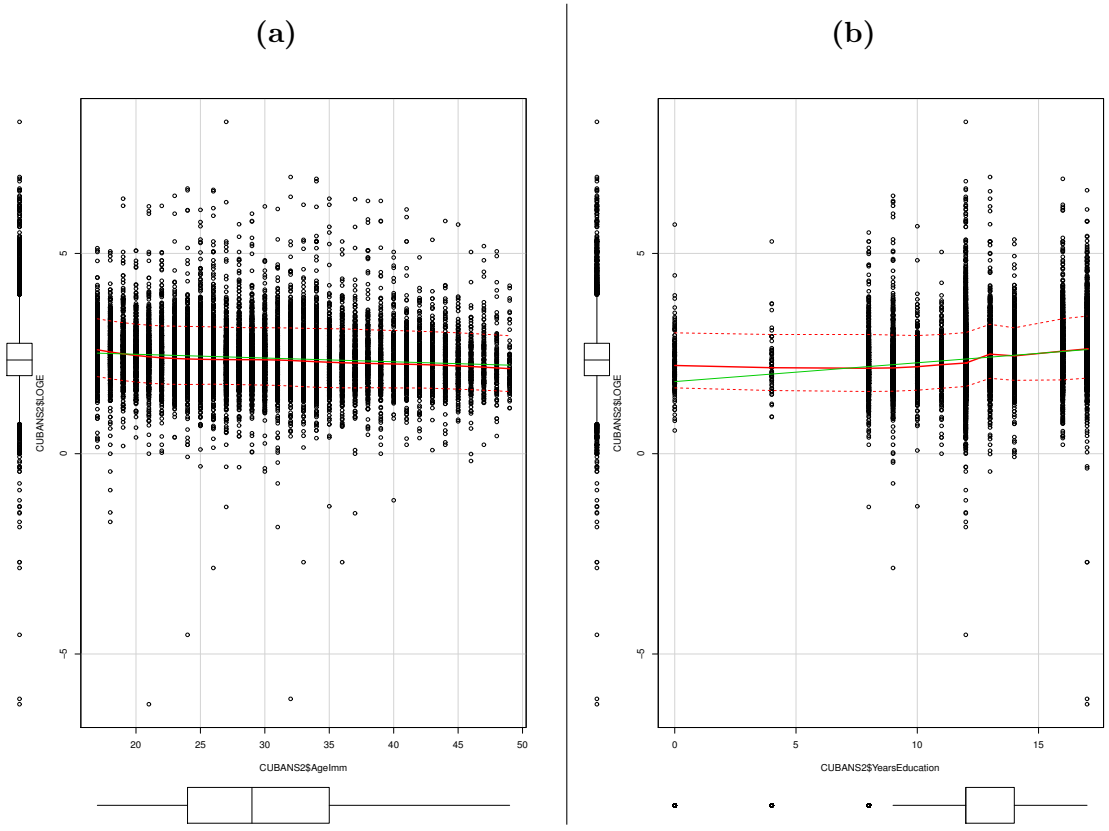
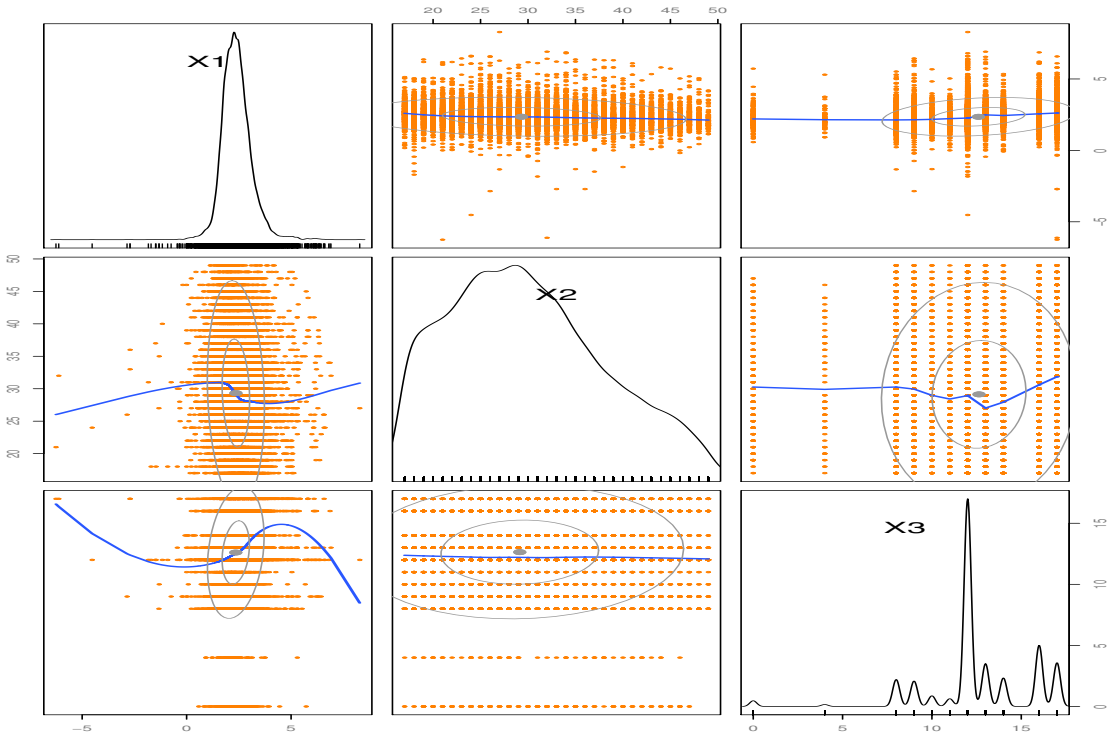


Figura 9.1: *Gráfico de dispersión entre las variables continuas en estudio y la variable respuesta.*



Note: $X1$ = (log) Hourly Earnings, $X2$ = Age at Migration, $X3$ = Years of Education

Figura 9.3: *Scatterplot Matrix de las covariables continuas y la variable respuesta.*

efecto de las mismas.

En la Figura 9.4 se aprecia la importancia de centrar las covariables cuando las bases son especificadas sin intercepto, apreciándose que cuando las covariables no están centradas, el origen y el centro de los datos no coinciden, en cambio cuando las covariables están centradas sí coinciden, capturándose la verdadera estructura de $\boldsymbol{x}$.

Centrar bases sin intercepto es de gran importancia porque en caso contrario, el efecto que resulta de esas bases sin intercepto sería forzado a través del origen (sin datos en esa zona) y de esa manera la convergencia sería muy lenta o el algoritmo Boosting no convergería a una solución correcta aún en un caso muy simple (Hofner et al., 2014).

Por ejemplo, si se considerara un predictor $\boldsymbol{x} = (x_1, \ldots, x_n)'$ distribuido normalmente y el modelo

$$\boldsymbol{y} = \beta \boldsymbol{x} + \boldsymbol{\varepsilon}$$

con $\beta = 1$ y $\varepsilon \sim N(0, 0.3^2)$. Si estimamos sin intercepto y se utilizara L_2 - Boosting [1], el gradiente negativo en la primera iteración será por defecto, la variable respuesta centrada $\boldsymbol{u}^{[1]} = \boldsymbol{y} - 1/n \sum_{i=1}^{n} y_i$ [2]. Ante esta situación, el uso de una base sin intercepto no es suficiente para recoger el efecto de β, por lo que resulta necesario y suficiente centrar el predictor $\boldsymbol{x}$.

Por tanto el intercepto necesita ser corregido, como se muestra a continuación

$$\begin{aligned} \hat{\boldsymbol{y}} &= \hat{\beta}_0 + \hat{\beta}_1 (\boldsymbol{x_1} - \bar{x_1}) + \hat{\beta}_2 (\boldsymbol{x_2} - \bar{x_2}) \\ \hat{\boldsymbol{y}} &= \underbrace{\left(\hat{\beta}_0 - \hat{\beta}_1 \bar{x}_1 - \hat{\beta}_2 \bar{x}_2\right)}_{\hat{\beta}_0^*} + \hat{\beta}_1 \boldsymbol{x_1} + \hat{\beta}_2 \boldsymbol{x_2} \end{aligned} \tag{9.2}$$

donde $\hat{\beta}_0^* = \hat{\beta}_0 - \sum_j \hat{\beta}_j \bar{x}_j$ sería el nuevo intercepto.

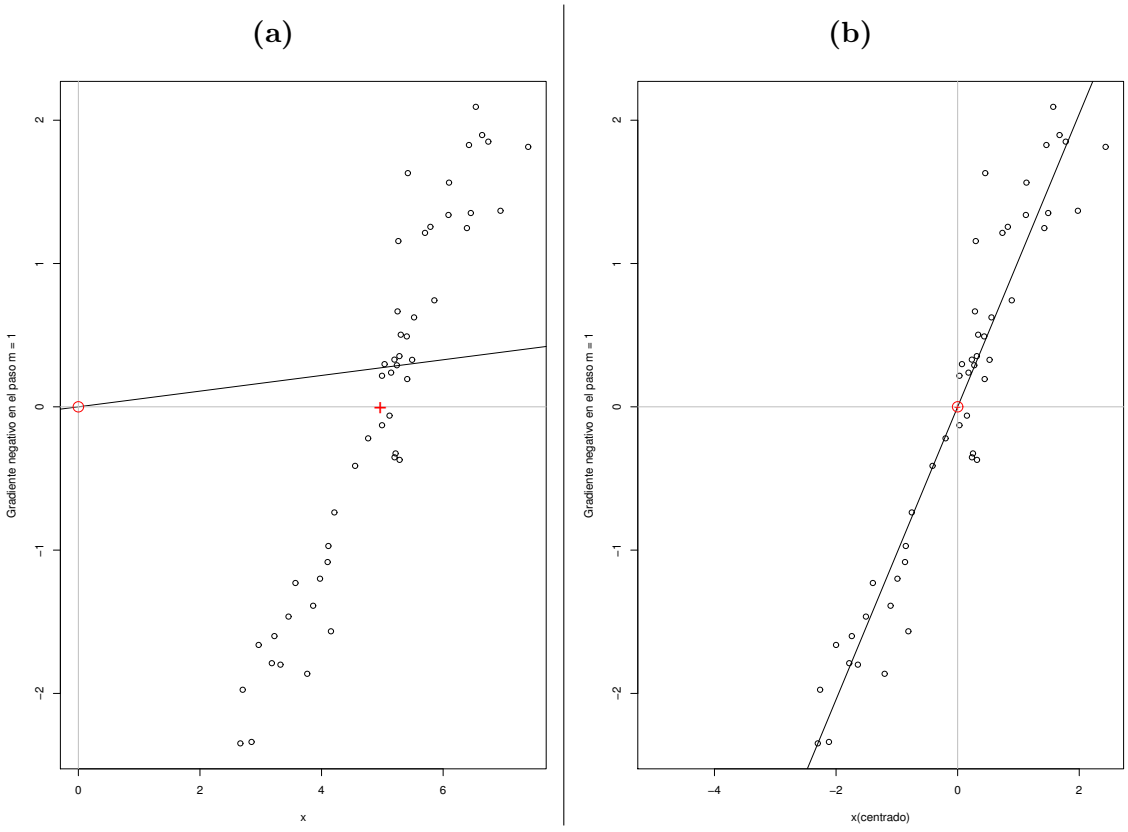


Figura 9.4: *Boosting en el primer paso de su algoritmo teniendo en cuenta (a) variables no centradas y (b) variables centradas.*

[1] L_2 - Boosting considera como función de pérdida el mínimo error cuadrático.

[2] Para otras funciones de pérdida el gradiente negativo en la primera iteración no tiene por qué ser exactamente la variable respuesta centrada con respecto a su media.

En el contexto de los P-splines lo que hacemos es descomponer el efecto suave $f(x)$ en un polinomio sin penalizar que correspondería con la parte lineal del efecto de la variable continua y una desviación suave de ese polinomio que sería una función de la variable centrada con respecto a su media.

Siguiendo esta pauta, para las covariables continuas hemos dividido el efecto suave como se especifica en la Ecuación 9.3. Así tendremos dividido el efecto de las covariables continuas en dos bases, una lineal y otra no lineal.

$$f(x) = \underbrace{\beta_0 + \beta_1 x}_{\text{polinomio sin penalizar}} + \underbrace{f_{centrada}(x)}_{\text{desviación suave}} \tag{9.3}$$

En los modelos GAM vistos en la Sección 7.1 del Capítulo 3 se asume que $E(f_j) = 0$, supuesto que se incorpora a la estimación centrando los valores de las variables con respecto a su media. Brezger and Lang (2006) explican que lo que se centra alrededor de la media son las funciones f_j en cada iteración del proceso de optimización para garantizar la identificabilidad del proceso porque las medias de las funciones f_j desconocidas son no identificables.

Un apartado importante en este algoritmo es definir el número de iteraciones óptimo (***mstop***) y este número lo hemos obtenido utilizando *K-Fold-Cross Validation* como sugieren Hofner et al. (2014). El número de iteraciones es el principal parámetro en el algoritmo Boosting ya que controla la complejidad del modelo y determina qué regresores van a ser incluidos.

Un valor alto para ***mstop*** conduce a modelos más complejos mientras que valores pequeños reducen la complejidad del modelo al dejar fuera del análisis a algunas covariables, por esa razón se optimiza el número de iteraciones para seleccionar las covariables más influyentes.

En la Figura 9.6 se puede apreciar el riesgo predicho en 25 muestras bootstraps. El número de iteraciones óptimo es aquel que minimiza el riesgo promedio de las 25 muestras resultando ser de 1.219 iteraciones.

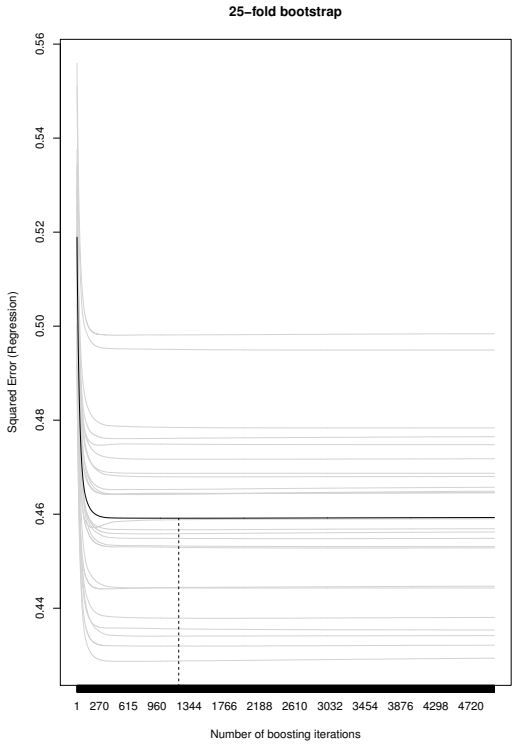


Figura 9.6: *Número de iteraciones óptimo.*

Además de obtener el número de iteraciones óptimas, es preciso establecer el factor de longitud de paso ν que debe ser pequeño para garantizar que el algoritmo no exceda el mínimo riesgo empírico definido en la Ecuación 8.8. Se ha seleccionado un valor $\nu = 0.25$.

Por defecto en el paquete ***mboost*** los grados de libertad [3] $df = 4$ pero vamos a utilizar $(df = 1)$ para todas las bases de manera que nos permita comparar el efecto de las mismas y que no se produzca un sesgo en la selección de las bases porque una base con un grado de libertad muy grande (que conlleva a una poca penalización) ofrece una mayor flexibilidad que una base con un grado de libertad muy pequeño (que conlleva una mayor penalización) y por tanto tendrá más oportunidad de ser seleccionada por el algoritmo Boosting. En este caso al definir los grados de libertad, el programa ***mboost*** ignora el valor de λ.

En la Figura 9.7 se aprecia el efecto de las covariables continuas: edad en el momento de emigrar, años de estudio y la interacción entre la edad en el momento de emigrar y el nivel

[3] Kneib et al. (2009) plantean que los grados de libertad $df(\lambda_d) = tr(S)$, mientras que Hofner et al. (2011) plantean que $df(\lambda_d) = tr(2S - S'S)$ siendo S la denominada matriz *hat* que une los residuos del gradiente estimados y observados en una base penalizada $\hat{u} = Su$.

de inglés del individuo en el momento en que se ha realizado la encuesta en diferentes cuantiles, una vez estimado el modelo aditivo estructurado planteado en la Ecuación 9.1.

En todas las funciones que reflejan el efecto de estas variables se aprecia un comportamiento no lineal y en forma de U, mientras que la función que representa el efecto de la edad en el momento de la emigración exhibe este comportamiento hasta el cuantil $\tau = 0.25$ porque a partir de la mediana muestra un comportamiento en forma de U-inversa.

El efecto de la edad en el momento de emigrar sobre el (log) del salario por hora es decreciente para aquellos individuos con edades inferiores a la media de la distribución de la variable edad en aquellos grupos de individuos que menos ganan, sin embargo este efecto se revierte cuando los individuos tienen salarios ubicados en el centro de la distribución o más altos.

Para aquellos que en el momento de la encuesta dominan el inglés, el efecto de la edad en el momento de la emigración tiene el mismo patrón de comportamiento en todos los cuantiles, o sea un efecto decreciente sobre el salario por hora cuando la edad es inferior a la media de edad y creciente para los que tenían una edad superior a la media de edad de todos los individuos en la muestra en el momento de entrar en Estados Unidos. Y lo mismo ocurre con los años de educación.

En la Tabla 9.2 se muestran los valores estimados para los coeficientes asociados a las variables categóricas del modelo y a los efectos lineales de las variables continuas. Los valores que aparecen en blanco se deben a que son variables que para el específico τ- cuantil no han sido seleccionadas por el algoritmo Boosting.

	Coeficientes para los diferentes valores de τ						
	0.05	0.10	0.25	0.50	0.75	0.90	0.95
Intercepto	−0.9039	−0.6647	−0.3269	0.0182	0.4101	0.8415	1.1337
Is Woman	−0.1221	−0.1756	−0.2349	-0.2821	−0.2885	−0.2746	−0.1703
Live in Florida	...	−0.0437	−0.0806	-0.0897	−0.1181	−0.1421	−0.1110
Is Not White	−0.1005	−0.0237	−0.0393	-0.0227	−0.0306	−0.0047	...
Is Married	0.0597	0.0676	0.0742	0.0724	0.0803	0.0571	0.0231
Is an American Citizen	0.1324	0.1365	0.2015	0.2249	0.2189	0.2129	0.1787
Proficiency in English	0.1056	0.1243	0.1653	0.1944	0.2026	0.1956	0.2156
Age at Migration	−0.0029	−0.0032	−0.0044	-0.0056	−0.0052	−0.0021	−0.0024
Years of Education	0.0110	0.0139	0.0183	0.0204	0.0203	0.0200	0.0129

Tabla 9.2: *Quantile Regression Estimation.*

Ser mujer presenta un efecto negativo en todos los cuantiles, efecto que se hace mayor en el centro de la distribución y que es prácticamente el mismo para el 10 % de los individuos que menos ganan y para el 90 % de los que más ganan.

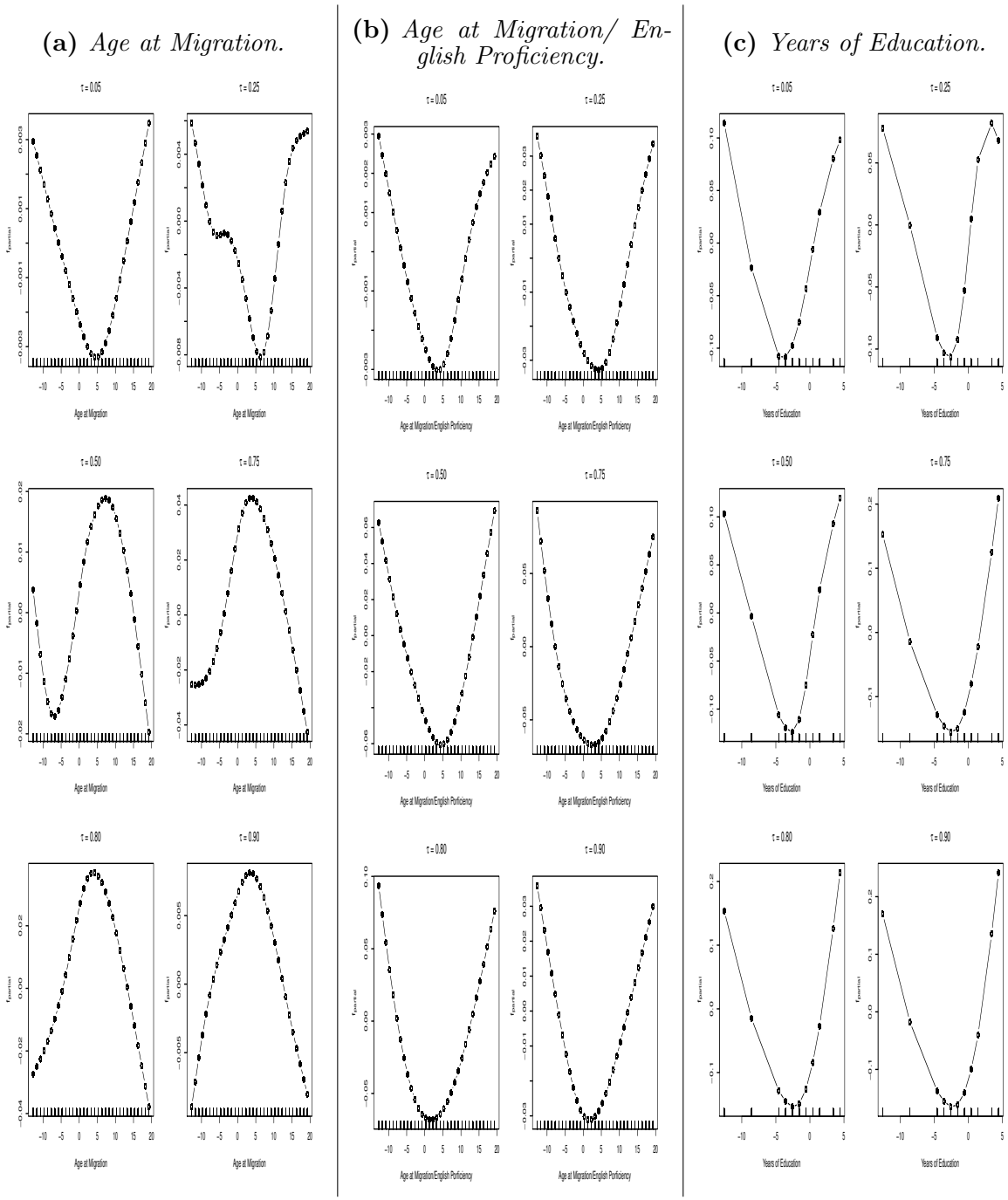


Figura 9.7: *Efectos suaves de los regresores continuos del modelo.*

Vivir en el Estado de Florida influye de manera negativa en el salario de todos los individuos independientemente del cuantil en el que nos encontremos, lo que en parte puede deberse al hecho de la gran concentración que existe en este Estado de individuos de origen cubano, como se ha visto en el Capítulo 2. El hecho de no ser blanco también influye de manera negativa en el salario de los cubanos, siendo más acusado este efecto negativo en los individuos que menos salario ganan.

El mayor efecto positivo sobre el salario lo muestra, en cada uno de los cuantiles estudiados, el hecho de poseer la nacionalidad norteamericana, seguido del hecho de dominar el idioma inglés.

En cuanto al nivel de inglés, se aprecia una tendencia creciente del rendimiento de esta característica sobre el salario a lo largo de la distribución, mostrándose también grandes diferencias entre las personas que ganan menos y las que ganan más. Para el 5 % de los cubanos que reciben los salarios más bajos, hablar fluidamente el inglés les repercute en sus salarios por hora con un aumento del 11 % mientras que para el 95 % de los que más ganan el incremento es del 22 %.

Si no hubiésemos incluido un efecto no lineal para las variables de edad en el momento de emigrar y años de educación, la relación lineal de ambas con el (log) del salario por hora sería en el sentido esperado, negativa para la edad y positiva para los años de educación.

9.4. Conclusiones

Con el interés de profundizar en el comportamiento de la distribución de salarios de los cubanos en Estados Unidos, se ha desarrollado en este Capítulo el novedoso algoritmo Boosting para estimar un modelo de regresión cuantílica aditivo estructurado (STAQ). Los datos utilizados han sido obtenidos de una muestra aleatoria del 1 % de la American Community Survey (ACS) facilitada por Integrated Public Use Microdata Series (IPUMS) (Ruggles et al., 2013) de la Universidad de Minnesota, contándose con una muestra de 15.131 individuos que han emigrado a Estados Unidos entre los años 1962 y 2007.

Para el análisis se ha tomado como variable respuesta el (log) del salario por hora obtenido por los individuos de la muestra en el ejercicio anterior al año en que se realizó la encuesta y como regresores a un conjunto moderado de variables lo que es deseable por razones de interpretabilidad (Wyatt and Altman, 1995) y que no contradice una de las ventajas en la utilización del Boosting relacionada con la selección de variables (Hofner et al., 2011). Las variables incluidas comprenden la edad del individuo en el momento de emigrar, los años

de educación y variables ficticias para indicar si el individuo es mujer, si vive en el Estado de Florida, si no es blanco, si está casado, si ha obtenido la nacionalidad norteamericana y si posee un alto nivel de inglés.

Se consideraron diferentes cuantiles a lo largo de la distribución del salario para observar las diferencias existentes en la influencia de los diferentes regresores en el salario por hora y sobre todo poder cuantificar estas diferencias entre los individuos que menos salarios reciben y los que más ganan. Los resultados obtenidos resultan interesantes y pueden extenderse a otros análisis de salarios una vez que hemos considerado en el modelo efectos lineales, efectos no lineales y la interacción entre el nivel de Inglés y la edad del individuo en el momento de emigrar.

Ser mujer presenta un efecto negativo en todos los cuantiles, siendo más negativo para las mujeres cuyos salarios se ubican en el centro de la distribución y no se aprecian grandes diferencias en el efecto de ser mujer sobre el salario entre los individuos que menos y más salarios perciben.

El hecho de vivir en el Estado de Florida, donde se concentran espacialmente los cubanos en Estados Unidos, influye de manera negativa en el salario de todos los individuos independientemente del cuantil en el que nos encontremos. Igual comportamiento presenta el hecho de no ser blanco, siendo más perjudicial para los individuos que menos salario ganan.

El mayor efecto positivo sobre el salario lo tiene, en cada uno de los cuantiles estudiados, la adquisición de la nacionalidad norteamericana, seguido del dominio del idioma inglés y la condición de estar casado. El nivel de inglés exhibe una tendencia creciente de su rendimiento sobre el salario a lo largo de la distribución.

Los cubanos llegados en las décadas recientes a Estados Unidos son más reticentes a la solicitud de la nacionalidad norteamericana y poseen niveles muy inferiores de inglés si se comparan con los cubanos que emigraron de Cuba en los años anteriores a 1980. Borjas (2011, 2015a) mostró cómo los inmigrantes recientes, en general, muestran menos habilidades en el uso del idioma inglés y el proceso de aprendizaje es más lento en comparación con los inmigrantes de décadas anteriores.

De no haber tenido en cuenta un efecto no lineal para las variables de edad en el momento de emigrar y años de educación, la relación lineal de ambas con el (log) del salario por hora sería en el sentido esperado, negativa para la edad y positiva para los años de educación. En relación a las variables continuas, el efecto no lineal de ambas sobre el salario resultó ser significativo.

La utilización de la regresión cuantílica aditiva es una contribución de gran importancia en el estudio del proceso de asimilación de los inmigrantes en el país de acogida, ya que se permite el uso de técnicas flexibles que se alejan del dogma de la linealidad, para describir la variación total que ocurre en la distribución de los salarios dado un conjunto de covariables.

CAPÍTULO 10

Conclusiones Generales

A lo largo del estudio realizado en esta tesis sobre la emigración cubana a Estados Unidos, se ha podido aportar a la investigación sobre este tema, importante material de análisis que permitirá el estudio de otros grupos de inmigrantes en Estados Unidos y definir nuevas líneas de investigación sobre los inmigrantes en general y sobre los cubanos en particular.

Pese a la relevancia de la inmigración cubana en Estados Unidos, poca literatura con carácter empírico se puede encontrar. La mayoría de los estudios existentes sobre este fenómeno versan sobre temas históricos y políticos.

El tratamiento a los cubanos como refugiados, amparados bajo la Ley de Ajuste Cubano de 1966, ha marcado diferencias entre la comunidad cubana y el resto de inmigrantes en Estados Unidos. No ha existido otro grupo de inmigrantes con similares privilegios.

Los cubanos han sido el único grupo a los que de forma automática se les ha otorgado el permiso de trabajo y ayudas para comida y alojamiento, una vez que están en suelo norteamericano; otorgándoseles además la residencia al año y un día. No existe una ley similar a la Ley de Ajuste Cubano que otorgue semejantes privilegios a un grupo de migrantes, a los que realmente no se les exige otro requisito que el de haber nacido en Cuba.

La política adoptada por los diferentes gobiernos de Estados Unidos, unido al descontento de la población cubana con el sistema económico y político cubano, ha sido un estímulo a la emigración desde Cuba y ha provocado el éxodo de cientos de miles de cubanos, lo que se ha materializado en las diferentes oleadas de emigración ocurridas desde 1959 hasta el presente.

Oeladas de emigración como las vividas entre 1958 y 1962 donde 248.870 personas salieron

de Cuba, incluidos niños que viajaron solos; el éxodo de Camarioca de 1965 con una salida de 300.000 cubanos; el del Mariel con 125.000 cubanos, el de 1994 denominado el éxodo de los balseros con más de 40.000 salidas y el éxodo actual con la travesía por diferentes países de Centroamérica que ha llevado a la llegada a Estados Unidos de casi 100.000 personas entre los años 2015 y 2016.

Tampoco se puede olvidar que la emigración cubana está fundamentalmente formada por personas con un nivel de educación alto, superior a la media de estudios de los que decidieron quedarse en Cuba y que es una comunidad que ha contribuido al desarrollo de la ciudad que ha sido su histórico enclave en Estados Unidos: Miami; cambiando a lo largo de décadas la composición y características de la misma (Grenier, 2015).

Ante estas premisas, resultan indispensables los análisis realizados en esta tesis, desde el momento de la toma de la decisión de emigrar hasta el proceso de incorporación al mercado laboral norteamericano.

El primer objetivo planteado fue el estudio de los cubanos que deciden emigrar, lo que se materializó a través de la estimación de un modelo logit que nos ha permitido analizar la influencia de diferentes variables socioeconómicas en la probabilidad de emigrar.

Este análisis se enmarca dentro de los estudios de la autoselección de los emigrantes, donde destacan trabajos como los de Borjas (1987, 1991), Chiswick (2000), Cattaneo (2007) y Grogger and Hanson (2011), entre otros. En el análisis de la autoselección de emigrantes hispanos en Estados Unidos existe una vasta bibliografía como los trabajos de Durand et al. (2001), Chiquiar and Hanson (2005), Cuecuecha (2010), Kaestner and Malamud (2014), entre otros. Pero sobre la autoselección de los emigrantes cubanos, lamentablemente no se cuenta con la misma actividad científica en este aspecto. Por lo que la contribución de esta tesis resulta importante.

Para el análisis de la autoselección de los cubanos se quiso hacer énfasis en la autoselección educativa, por ello se escogieron como variables explicativas del salario por hora, la edad del individuo en el momento de emigrar, los años de educación y el estatus ocupacional de la persona. Se utilizaron datos del censo de Población y Vivienda de Estados Unidos en el año 2010, facilitados por Ruggles et al. (2011a) y del Censo de Población y Vivienda de Cuba en el año 2002 obtenido a través de Ruggles et al. (2011b).

De la muestra limitada a personas que emigraron con edad superior a los 17 años para evitar la existencia de individuos que hayan completado su formación en Estados Unidos y que en el momento del censo tenían entre 16 y 49 años, se obtuvo que la probabilidad de emigrar de un individuo con más de 13 años de educación es considerablemente superior

a la probabilidad de emigrar si tuviera menos años de educación, resultando ser casi 15 veces mayor a la de un individuo que cuenta con menos de 9 años de estudios. En cuanto a la edad, se incrementa la probabilidad de emigrar para los que tienen edades entre los 41 y los 49 años y es más probable la emigración de aquellos que tienen categoría profesional de Obreros Cualificados.

Un segundo objetivo de esta tesis ha sido el análisis de los salarios de los inmigrantes cubanos en Estados Unidos, análisis que se ha realizado utilizando modelos estimados por regresión cuantílica desde tres perspectivas: regresión cuantílica lineal (Koenker and Bassett, 1978), regresión cuantílica lineal teniendo en cuenta la selección muestral (Buchinsky, 1998b, 2002) y la regresión cuantílica a través de un modelo aditivo estructurado (Fahrmeir et al., 2004), (Kneib et al., 2009) y (Fenske et al., 2012).

La necesidad de analizar cada uno de los puntos de la distribución de salarios y no limitarnos al análisis sobre la media ha sido una de las principales motivaciones de la utilización de la regresión cuantílica. Sólo así se puede llegar a una visión completa de cómo variables como la edad en el momento de emigrar, el hecho de ser mujer, de ser negro, de estar casado, de poseer la nacionalidad estadounidense, los años de educación, la experiencia potencial y el dominio del inglés influyen en cada uno de los cuantiles condicionados de la distribución de salarios.

En el análisis utilizando regresión cuantílica lineal se utilizaron datos de sección cruzada repetida entre los años 2000 y 2007, con individuos de entre 25 y 50 años en el momento del censo y que tenían entre 17 y 49 años cuando emigraron. Los resultados indicaron que ser mujer tiene un efecto negativo sobre los salarios en cada uno de los cuantiles de la distribución de salarios así como el hecho de ser negro para quienes ganan un salario inferior al salario mediano. Por cada año adicional en el momento de la emigración, el salario por hora también disminuye en todos los cuantiles.

Los mayores efectos positivos sobre el salario son atribuibles a la posesión de la nacionalidad norteamericana y al dominio del inglés. Los años de estudio mostraron un efecto positivo, pero inferior a lo que se esperaba. Sobre este aspecto, Machado and Mata (2005) indican que el exceso de mano de obra cualificada trae consigo la disminución de los salarios y Borjas (2014) señala que la educación sólo representa una pequeña parte de la varianza de los ingresos entre los trabajadores, lo que puede sugerir que el alto nivel de estudios adquirido por los cubanos en Cuba no se está transfiriendo a los salarios una vez que llegan a Estados Unidos.

Para el análisis de los salarios utilizando un modelo de regresión cuantílica teniendo en cuenta la selección muestral, se tomó como base los trabajos de Buchinsky (1998b, 2002)

realizando el test de independencia propuesto por Huber and Melly (2015). Los datos fueron divididos en cohortes según la década en la que los individuos arribaron a Estados Unidos: 1980, 1990 y 2000 para poder definir diferencias en cuanto a la influencia de las variables socioeconómicas seleccionadas sobre el salario. La muestra se restringió a individuos con edades entre los 17 y 49 años en el momento de emigrar y que tuviesen entre 25 y 55 años en el momento de la encuesta.

Del análisis realizado se pudo comprobar que sólo en la década de los años 2000 todas las variables excepto el estado civil y el género, son significativas en la ecuación de selección para un nivel de significación del 5 %, siendo las variables que identifican el tener la nacionalidad norteamericana, el ser negro y dominar el idioma el inglés las más importantes en la probabilidad de participar en el mercado laboral. En esta ecuación de selección se tuvieron en cuenta las mismas variables introducidas en la ecuación de salarios más la edad en el momento de emigrar, el sexo y el estado civil.

En la estimación de la ecuación de salarios, no todas las variables resultaron ser significativas en todos los cuantiles. Para los cubanos que llegaron a Estados Unidos en la década de los años 80, sólo el nivel de inglés resultó ser una variable relevante para los individuos ubicados en el centro y parte superior de la distribución de salarios, teniendo un efecto positivo para todos los cuantiles.

En la década de los 90, el tener la nacionalidad emergió como un nuevo factor determinate de los salarios, con un efecto positivo y ser negro resultó ser perjudicial para los individuos que ganan menos salarios.

En la década del 2000 el tener la nacionalidad apenas tiene efecto sobre el salario y el efecto se vuelve negativo para las personas que más salarios perciben, dominar el idioma inglés es significativo e importante para los que ganan menos y los años de educación sólo tienen un rendimiento positivo para los que más ganan pero exhiben un rendimiento muy bajo.

Una contribución destacada de esta tesis ha sido la utilización de un modelo de regresión cuantílica aditiva estructurada en el análisis de los salarios, aplicación que no ha sido utilizada antes en los estudios sobre salarios e inmigración. Los Modelos Cuantílicos Aditivos Estructurados (STAQ) se deben a los trabajos de Fahrmeir et al. (2004), Kneib et al. (2009) y Fenske et al. (2012) y su empleo parte de la necesidad de permitir mayor flexibilidad en el planteamiento de la relación entre los regresores y la variable respuesta.

En este modelo, como regresores se tomaron en consideración las mismas variables que en los análisis anteriores pero se sustituyó a la variable identificativa de ser negro por una

que definiera si el individuo no es blanco, por el hecho de que existen antecedentes de que muchos cubanos no se declaran negros en las encuestas. Se introdujo una nueva variable que represente si el individuo vive en la Florida, no se tuvo en cuenta la Experiencia Potencial y se planteó la interacción entre el nivel de inglés y la edad del individuo en el momento de emigrar.

Ser mujer presenta un efecto negativo en todos los cuantiles, siendo más negativo para las mujeres cuyos salarios se ubican en el centro de la distribución y no se apreciaron grandes diferencias en el efecto de ser mujer sobre el salario entre los individuos que menos y más salarios perciben. El hecho de vivir en el Estado de Florida, donde se concentran los cubanos en Estados Unidos, influye de manera negativa en el salario de todos los individuos independientemente del cuantil en el que nos encontremos. Igual comportamiento presenta el hecho de no ser blanco, siendo más perjudicial para los individuos que menos salario ganan.

El mayor efecto positivo sobre el salario lo mostró, en cada uno de los cuantiles estudiados, la adquisición de la nacionalidad norteamericana, seguido del dominio del idioma inglés y la condición de estar casado. El nivel de inglés exhibió una tendencia creciente de su rendimiento sobre el salario a lo largo de la distribución.

Comparando con los resultados obtenidos en la Sección 6.1 del Capítulo 6 donde se utilizaba la regresión cuantílica lineal; para los individuos cuyos salarios corresponden al valor mediano de la distribución y para las variables que coinciden en las dos especificaciones, no encontramos diferencias en cuanto a signos ni grandes diferencias en cuanto al valor de los coeficientes estimados a excepción del hecho de ser norteamericano, ya que el efecto sobre el salario es muy superior en el modelo estimado con Boosting que en el modelo de regresión cuantílica lineal.

Para los que menos ganan y cuyos salarios se ubican en el cuantil $\tau = 0.10$ no existen diferencias en cuanto al sentido de la relación lineal y la mayor diferencia en cuanto a efecto se presenta en el nivel de inglés, cuyo rendimiento es mayor en el modelo estimado con Boosting. Por otra parte, para los que más ganan y cuyos salarios se ubican en el cuantil $\tau = 0.90$ destaca las diferencia en cuanto a sentido del efecto de la raza, puesto que en la regresión cuantílica lineal el efecto era positivo para esos individuos y en el Boosting es negativo y el efecto de la nacionalidad que es mucho mayor cuando hemos utilizado el modelo aditivo estructurado estimado a través del Boosting.

10.1. Futuras Líneas de Investigación

El tema de estudio de esta tesis se enmarca en un ámbito que ha ganado atención en los últimos años a pesar de no ser un fenómeno nuevo en la historia de la Humanidad: la emigración. Con las tendencias actuales de movimientos migratorios, las políticas a favor y en contra de los mismos, es un tema de estudio abierto a la constante actualización y aplicación de los conocimientos adquiridos en las investigaciones llevadas a cabo con anterioridad.

En este sentido, desde el punto de vista del estudio de la emigración cubana hacia Estados Unidos y su repercusión en la vida cotidiana los cubanos dentro de Cuba, es necesario analizar en profundidad lo que implica que sean las personas con más años de estudio los que emigran.

Hay voces que advierten de la pérdida irreparable que trae consigo para los países emisores el que sean sus individuos más preparados los que emigren (Moraga, 2011) pero también otros autores defienden la idea de que la emigración de los más cualificados trae beneficios para los países emisores que se revierten luego en remesas e inversiones en el país de origen (Ratha et al., 2015).

Otro aspecto que invita a la reflexión es investigar de qué manera aquellos emigrantes cubanos que adquirieron sus estudios superiores en Cuba, transfieren sus conocimientos, perfeccionados y enriquecidos en el país de destino, a su país de origen. Esto en la literatura forma parte de lo que se denomina las "remesas del conocimiento" (Blanco, 2013).

En los dos análisis realizados de los salarios teniendo en cuenta la regresión cuantílica lineal, se ha obtenido un bajo rendimiento de la educación. Este resultado invita a realizar estudios futuros en este sentido. Matos and Liebig (2014) en un artículo sobre las habilidades de los inmigrantes en Europa y Estados Unidos, obtienen que los inmigrantes cuya formación proviene de los países de origen, tienen un rendimiento menor de la educación sobre los salarios y grandes desventajas en cuanto a calidad del empleo que se les ofrece.

Utilizando la regresión cuantílica aditiva estructurada, queda pendiente hacer el análisis de los salarios empleando los mismos cohortes que se utilizaron en la regresión cuantílica lineal teniendo en cuenta la selección muestral, o sea dividir a los individuos según la década en la que llegaron a Estados Unidos.

Otro campo de estudio futuro es profundizar en el análisis de salarios para aquellos sectores donde se han concentrado los emigrantes cubanos y ver sus diferencias con otros grupos

de inmigrantes que no han gozado de los privilegios que a los cubanos les ha ofrecido la política del gobierno de Estados Unidos, siguiendo las pautas del modelo de Machado and Mata (2005) y del modelo de Gelbach (2016), intentando extender ambos a un contexto de regresión cuantílica aditiva estructurada.

10.2. Divulgación de los resultados obtenidos en esta tesis

Artículos Publicados

Aleida Cobas-Valdés y Ana Fernández-Sainz (2013). La Emigración Cubana a Estados Unidos y la Autoselección en Educación. *Revista Vasca de Sociología y Ciencia Política*, Vol. 52-53, pp. 125-137.

Aleida Cobas-Valdés y Ana Fernández Sainz (2014). Los cubanos se autoseleccionan en su emigración? De Cuba a Estados Unidos. *Revista Internacional de Ciencias Sociales Interdisciplinares*, Vol 3, Nº 1, pp. 61-77.

Aleida Cobas-Valdés y Ana Fernández-Sainz (2014). Cuban Migration to the United States and the Educational Self-Selection Problem. *International Journal of Cuban Studies*, Vol. 6, Nº 1 , pp. 41-54.

Aleida Cobas-Valdés, Javier Fernández-Macho y Ana Fernández-Sainz (2016). Earnings Distribution of Cuban Immigrants in the USA: Evidence from Quantile Regression with Sample Selection, publicado online en la revista *Applied Economics*.

Aleida Cobas- Valdés, Ana Fernández-Sainz y Stephen Wilkinson (2016). Inmigrantes cubanos en Estados Unidos: ¿Qué determina su distribución de ingresos?. *Revista Semestre Económico*, Vol. 19, Nº 41, pp. 19-36.

Artículo bajo proceso de revisión

Aleida Cobas-Valdés, Javier Fernández-Macho y Ana Fernández-Sainz (2016). What determines the earnings distribution of Cuban immigrants in the United States? A Conditional Quantile Regression Analysis. Se encuentra en proceso de revisión en la revista *Applied Economics Letters*.

Congresos y Jornadas donde se han expuesto los resultados alcanzados

IX Congreso Vasco de Sociología y Ciencia Política, Bilbao. 16-18 de Julio de 2012. Comunicación Oral: La Emigración Cubana a Estados Unidos y la Autoselección en Educación.

VIII Congreso Internacional de Ciencias Sociales Interdisciplinares, Praga. Rep. Checa, 30 de Julio- 1 de Agosto de 2013. Comunicación Oral: La Emigración Cubana a Estados Unidos: El Problema de la Autoselección en Educación.

Workshop in Econometrics and Empirical Economics, Perugia, Italia, 28-29 Agosto de 2014. Comunicación Oral: Earnings Distribution of Cuban Immigrants in the United States. Evidence from Quantile Regressions.

Association for the Study of the Cuban Economy (ASCE). Twenty-Sixth Aannual Meeting, Miami, Estados Unidos. 28-30 de Julio de 2016. Comunicación Oral: Cuban Immigrants in the United States: What Determines their Earning Distributions? ***Trabajo Ganador del Primer Premio en***: The 2016 Jorge Pérez-López Graduate and Undergraduate Student Award Competition.

I Jornadas Doctorales de la UPV/EHU, Bilbao. 11-12 de Julio de 2016. Earnings Distribution of Cuban Immigrants in the U.S. Evidence from Quantile Regression with Sample Selection. Presentación en Forma de Póster y Comunicación Oral.

II Jornadas de Doctorados en Economía y Empresa de la Universidad del País Vasco UPV/EHU, Bilbao. 28-30 de Septiembre de 2016. Comunicación Oral: Earnings Distribution of Cuban Immigrants in the U.S. Evidence from Quantile Regression with Sample Selection.

Bibliografía

Abadie, A., Angrist, J., & Imbens, G. (2002). Instrumental Variables Estimates of the Effect of Subsidized Training on the Quantiles of Trainee Earnings. *Econometrica*, 70(1):91–117.

Abramitzky, R., Boustan, L. P., & Eriksson, K. (2012). Europe's Tired, Poor, Huddled Masses: Self-Selection and Economic Outcomes in the Age of Mass Migration. *The American Economic Review*, 102(5):1832–1856.

Abramitzky, R. & Braggion, F. (2006). Migration and Human Capital: Self-Selection of Indentured Servants to the Americas. *The Journal of Economic History*, 66(4):882–905.

Adsera, A. & Pytlikova, M. (2015). The Role of Language in Shaping International Migration. *Feature Issue*, 125(586):F49–F81.

Ahn, H. & Powell, J. L. (1993). Semiparametric estimation of censored selection models with a nonparametric selection mechanism. *Journal of Econometrics*, 58(1-2):3–29.

Aja Díaz, A. (2006). La migración desde Cuba. *Revista sobre Fronteras e Integración*, 11(22):7–16.

Aja Díaz, A. (2010). Los Estados Unidos-Cuba. Emigración y Relaciones Bilaterales. *Temas*, 62.

Albo, A. & Díaz, J. L. O. (2011). Migración Mexicana Altamente Calificada en EEUU y Transferencia de México a Estados Unidos a Través del Gasto en la Educación de los Migrantes. *Documentos de Trabajo, Análisis Económico, Servicio de Estudios Económicos. Madrid: Fundación BBVA*.

Albrecht, J., Van Vuuren, A., & Vroman, S. (2009). Counterfactual Distributions with Sample Selection Adjustments: Econometric Theory and an Application to the Netherlands. *Labour Economics*, 16(4):383–396.

Alhamzawi, R., Yu, K., & Benoit, D. F. (2012). Bayesian Adaptive Lasso Quantile Regression. *Statistical Modelling*, 12(3):279–297.

Andriyana, Y., Gijbels, I., & Verhasselt, A. (2016). Quantile Regression in Varying-Coefficient Models: Non-Crossing Quantiles Curves and Heteroscedasticity. *Statistical Papers*.

Antecol, H., Kuhn, P., & Trejo, S. J. (2006). Assimilation via Prices or Quantities? Sources of Immigrants Earnings Growth in Australia, Canada and the United States. *Journal of Human Resources*, 41(4):821–840.

Arango, J. (2003). La explicación teórica de las migraciones: luz y sombra. *Migración y desarrollo*, 1(1):1–30.

Arellano, M. & Bonhomme, S. (2017). Sample Selection in Quantile Regression: A Survey. *Handbook of Quantile Regression*, forthcoming.

Atun, R., de Andrade, L. O. M., Almeida, G., Cotlear, D., Dmytraczenko, T., Frenz, P., Garcia, P., Gómez-Dantés, O., Knaul, F. M., Muntaner, C., de Paula, J. B., Rígoli, F., Serrate, P. C.-F., & Wagstaff, A. (2015). Health-system reform and universal health coverage in Latin America. *The Lancet*, 385(9974):1230–1247.

Autor, D. H., Houseman, S. N., & Kerr, S. P. (2017). The effect of work first job placements on the distribution of earnings: An instrumental variable quantile regression approach. *Journal of Labor Economics*, 35(1):149–190.

B Dunson, D., Watson, M., & Taylor, J. A. (2003). Bayesian Latent Variable Models for Median Regression on Multiple Outcomes. *Biometrics*, 59(2):296–304.

Balibar, E. & Wallerstein, I. M. (1991). *Race, Nation, Class: Ambiguous Identities*. Verso.

Bauer, E. & Kohavi, R. (1999). An Empirical Comparison of Voting Classification Algorithms: Bagging, Boosting, and Variants. *Machine Learning*, 36(1-2):105–139.

Beine, M., Bertoli, S., & Moraga, J. F.-H. (2016). A Practitioners' Guide to Gravity Models of International Migration. *The World Economy*, 39(4):496–512.

Benoit, D. F. & Van den Poel, D. (2012). Binary Quantile Regression: A Bayesian Approach Based on the Asymmetric Laplace Distribution. *Journal of Applied Econometrics*, 27(7):1174–1188.

Bertoli, S. & Moraga, J. F.-H. (2015). The Size of the Cliff at the Border. *Regional Science and Urban Economics*, 51:1–6.

Bertoli, S., Moraga, J. F.-H., & Ortega, F. (2013). Crossing the Border: Self-Selection, Earnings and Individual Migration Decisions. *Journal of Development Economics*, 101:75–91.

Biavaschi, C. & Elsner, B. (2013). Let's Be Selective about Migrant Self-Selection. Discussion Paper 7865, IZA.

Blanco, J. A. (2013). *Remesas del Conocimiento: Del Brain Drain al Brain Gain*. Miami Dade College.

Blundell, R., Gosling, A., Ichimura, H., & Meghir, C. (2007). Changes in the Distribution of Male and Female Wages Accounting for Employment Composition Using Bounds. *Econometrica*, 75(2):323–363.

Bollaerts, K., Eilers, P. H., & Aerts, M. (2006). Quantile Regression with Monotonicity Restrictions Using P-splines and the L1-norm. *Statistical Modelling*, 6(3):189–207.

Bondell, H. D., Reich, B. J., & Wang, H. (2010). Noncrossing Quantile Regression Curve Estimation. *Biometrika*, 97(4):825–838.

Borjas, G. J. (1982). The Earnings of Male Hispanic Immigrants in the United States. *ILR Review*, 35(3):343–353.

Borjas, G. J. (1985). Assimilation, Changes in Cohort Quality, and the Earnings of Immigrants. *Journal of Labor Economics*, 3(4):463–489.

Borjas, G. J. (1987). Self-Selection and the Earnings of Immigrants. *The American Economic Review*, 77(4):531–553.

Borjas, G. J. (1989). Economic Theory and International Migration. *The International Migration Review*, 23(3):457–485.

Borjas, G. J. (1991). Immigration and Self-Selection. *In: Immigration, Trade, and the Labor Market*, pages 29–76.

Borjas, G. J. (1995). Assimilation and Changes in Cohort Quality Revisited: What Happened to Immigrant Earnings in the 1980s? *Journal of Labor Economics*, 13(2):201–245.

Borjas, G. J. (2011). *Heaven's Door: Immigration Policy and the American Economy*. Princeton University Press.

Borjas, G. J. (2014). *Immigration Economics*. Harvard University Press.

Borjas, G. J. (2015a). The Slowdown in the Economic Assimilation of Immigrants: Aging and Cohort Effects Revisited Again. *Journal of Human Capital*, 9(4):483–517.

Borjas, G. J. (2015b). The Wage Impact of the Marielitos: A Reappraisal. Technical report, National Bureau of Economic Research.

Borjas, G. J., Kauppinen, I., & Poutvaara, P. (2015). Self-Selection of Emigrants: Theory and Evidence on Stochastic Dominance in Observable and Unobservable Characteristics. Technical report, NBER.

Breiman, L. (1993). Fitting Additive Models to Regression Data: Diagnostics and Alternative Views. *Computational Statistics & Data Analysis*, 15(1):13–46.

Breiman, L. (1996). Bagging Predictors. *Machine Learning*, 24(2):123–140.

Breiman, L. (1998). Arcing Classifier (With Discussion and a Rejoinder by the Author). *The Annals of Statistics*, 26(3):801–849.

Breiman, L. (1999). Prediction Games and Arcing Algorithms. *Neural Computation*, 11(7):1493–1517.

Breiman, L. (2001a). Random Forests. *Machine Learning*, 45(1):5–32.

Breiman, L. (2001b). Statistical Modeling: The Two Cultures (with discussion). *Statistical Science*, 16(3):199–231.

Breiman, L. & Friedman, J. H. (1985). Estimating Optimal Transformations for Multiple Regression and Correlation: Rejoinder. *Journal of the American Statistical Association*, 80(391):614–619.

Brezger, A. & Lang, S. (2003). Generalized Additive Regression Based on Bayesian P-splines. In *SFB Discussion Paper 321, Deparment of Statistics, University of Munich*.

Brezger, A. & Lang, S. (2006). Generalized Structured Additive Regression Based on Bayesian P-splines. *Computational Statistics & Data Analysis*, 50(4):967–991.

Briollais, L. & Durrieu, G. (2014). Application of quantile regression to recent genetic and–omic studies. *Human Genetics*, 133(8):951–966.

Brown, A. & López, M. H. (2013). Mapping the Latino Population, by State, County and City. *Washington, DC: Pew Research Center*.

Brunner, L. & Pate, J. (2016). Promoting Entry of High-Quality Workers through US Immigration Policy. *Applied Economics*, 48(52):1–15.

Buchinsky, M. (1995). Estimating the Asymptotic Covariance Matrix for Quantile Regression Models: A Monte Carlo Study. *Journal of Econometrics*, 68(2):303–338.

Buchinsky, M. (1998a). Recent Advances in Quantile Regression Models: A Practical Guideline for Empirical Research. *Journal of Human Resources*, 33(1):88–126.

Buchinsky, M. (1998b). The Dynamics of Changes in the Female Wage Distribution in the USA: A Quantile Regression Approach. *Journal of Applied Econometrics*, 13(1):1–30.

Buchinsky, M. (2002). Quantile Regression with Sample Selection: Estimating Women's Return to Education in the U.S. *Empirical Economics*, 26(1):87–113.

Bühlmann, P., Gertheiss, J., Hieke, S., Kneib, T., Ma, S., Schumacher, M., Tutz, G., Wang, C., Wang, Z., & Ziegler, A. (2014). Discussion of the Evolution of Boosting Algorithms and Extending Statistical Boosting. *Methods of Information in Medicine*, 53(6):436–445.

Bühlmann, P. & Hothorn, T. (2007). Boosting Algorithms: Regularization, Prediction and Model Fitting. *Statistical Science*, 22(4):477–505.

Bühlmann, P. & Yu, B. (2002). Analyzing Bagging. *The Annals of Statistics*, 30(4):927–961.

Bühlmann, P. & Yu, B. (2003). Boosting with the L2 Loss: Regression and Classification. *Journal of the American Statistical Association*, 98(462):324–339.

Burgette, L. F. & Reiter, J. P. (2012). Modeling Adverse Birth Outcomes via Confirmatory Factor Quantile Regression. *Biometrics*, 68(1):92–100.

Cade, B. S. (2011). Estimating Equivalence with Quantile Regression. *Ecological Applications*, 21(1):281–289.

Cai, Y. & Jiang, T. (2015). Estimation of Non-Crossing Quantile Regression Curves. *Australian & New Zealand Journal of Statistics*, 57(1):139–162.

Card, D. (1990). The Impact of the Mariel Boatlift on the Miami Labor Market. *Industrial & Labor Relations Review*, 43(2):245–257.

Carliner, G. (1980). Wages, Earnings and Hours of First, Second and Third Generation American Males. *Economic Inquiry*, 18(1):87–102.

Castro, F. & Ramonet, I. (2006). *Cien Horas con Fidel*. Oficina de Publicaciones del Consejo de Estado.

Cattaneo, C. (2007). The Self-Selection in the Migration Process: What Can We Learn? Technical Report 199. Serie Economia e Impresa.

Chamberlain, G. (1994). Quantile Regression, Censoring, and the Structure of Wages. In Sims, C. A., editor, *Advances in Econometrics*, volume 2, pages 171–209. Cambridge University Press (CUP).

Chen, C. W., Hsu, Y.-T., & Taniguchi, M. (2017). Discriminant Analysis by Quantile Regression with Aapplication on the Climate Change Problem. *Journal of Statistical Planning and Inference*.

Chen, S. & Zhou, Y. (2010). Semiparametric and Nonparametric Estimation of Sample Selection Models Under Symmetry. *Journal of Econometrics*, 157(1):143–150.

Cheng, G., Zhang, H. H., & Shang, Z. (2015). Sparse and Efficient Estimation for Partial Spline Models with Increasing Dimension. *Annals of the Institute of Statistical Mathematics*, 67(1):93–127.

Chernozhukov, V., Fernandez-Val, I., & Galichon, A. (2009). Improving Point and Interval Estimators of Monotone Functions by Rearrangement. *Biometrika*, page asp030.

Chernozhukov, V. & Hansen, C. (2006). Instrumental Quantile Regression Inference for Structural and Treatment Effect Models. *Journal of Econometrics*, 132(2):491–525.

Chernozhukov, V. & Hansen, C. (2008). Instrumental Variable Quantile Regression: A Robust Inference Approach. *Journal of Econometrics*, 142(1):379–398.

Chernozhukov, V., Imbens, G. W., & Newey, W. K. (2007). Instrumental Variable Estimation of Nonseparable Models. *Journal of Econometrics*, 139(1):4–14.

Chesher, A. (2003). Identification in Nonseparable Models. *Econometrica*, 71(5):1405–1441.

Chesher, A. (2005). Nonparametric Identification under Discrete Variation. *Econometrica*, 73(5):1525–1550.

Chesher, A. (2007). Instrumental Values. *Journal of Econometrics*, 139(1):15–34.

Chib, S., Greenberg, E., & Jeliazkov, I. (2009). Estimation of Semiparametric Models in the Presence of Endogeneity and Sample Selection. *Journal of Computational and Graphical Statistics*, 18(2):321–348.

Chiquiar, D. & Hanson, G. H. (2005). International Migration, Self-selection, and the Distribution of Wages: Evidence from México and the United States. *Journal of Political Economy*, 113(2):239–281.

Chiswick, B. R. (1978). The effect of Americanization on the Earnings of Foreign-Born Men. *The Journal of Political Economy*, 86(5):897–921.

Chiswick, B. R. (2000). Are Immigrants Favorably Self-Selected? An Economic Analysis. Discussion Paper 131, IZA.

Chiswick, B. R., Le, A. T., & Miller, P. W. (2008). How Immigrants Fare across the Earnings Distribution in Australia and the United States. *Industrial & Labor Relations Review*, 61(3):353–373.

Chiswick, B. R. & Miller, P. W. (2002). Immigrant Earnings: Language Skills, Linguistic Concentrations and the Business Cycle. *Journal of Population Economics*, 15(1):31–57.

Chiswick, B. R. & Miller, P. W. (2015). International Migration and the Economics of Language. *Handbook of the Economics of International Migration.*, 1:211–269.

Christmann, A. & Hable, R. (2012). Consistency of Support Vector Machines using Additive Kernels for Additive Models. *Computational Statistics & Data Analysis*, 56(4):854–873.

Clark, X., Hatton, T. J., & Williamson, J. G. (2002). Where Do US Immigrants Come From, and Why? Technical report, National Bureau of Economic Research.

Clark, X., Hatton, T. J., & Williamson, J. G. (2007). Explaining U.S. Immigration, 1971–1998. *Review of Economics and Statistics*, 89(2):359–373.

Cleophas, T. J. & Zwinderman, A. H. (2013). *Machine Learning in Medicine*. Springer.

Cobas, A. & Fernández, A. (2014). Cuban Migration to the United States and the Educational Self-Selection Problem. *International Journal of Cuban Studies*, 6(1):41–54.

Cobas-Valdés, A., Fernández-Macho, J., & Fernández-Sainz, A. (2016). Earnings Distribution of Cuban Immigrants in the USA: Evidence from Quantile Regression with Sample Selection. *Applied Economics*, pages 1–16.

Colby, S. L. & Ortman, J. M. (2015). Projections of the Size and Composition of the U.S. Population: 2014 to 2060. *Current Population Reports*, pages 25–1143.

Cortes, C. & Vapnik, V. (1995). Support-Vector Networks. *Machine Learning*, 20(3):273–297.

Cortes, K. E. (2004). Are Refugees Different from Economic Immigrants? Some Empirical Evidence on the Heterogeneity of Immigrant Groups in the United States. *Review of Economics and Statistics*, 86(2):465–480.

Craig, S. G. & Ng, P. T. (2001). Using Quantile Smoothing Splines to Identify Employment Subcenters in a Multicentric Urban Area. *Journal of Urban Economics*, 49(1):100–120.

Cribari-Neto, F. (2004). Asymptotic Inference under Heteroskedasticity of Unknown Form. *Computational Statistics & Data Analysis*, 45(2):215–233.

Cuecuecha, A. (2005). The Immigration of Educated Mexicans: The Role of Informal Social Insurance and Migration Costs. *Instituto Tecnológico Autónomo de México. Mimeograph.*

Cuecuecha, A. (2010). Las Características Educativas de los Emigrantes Mexicanos a Estados Unidos. *EconoQuantum*, 7(1):9–42.

Currie, I. D. & Durbán, M. (2002). Flexible Smoothing with P-Splines: A Unified Approach. *Statistical Modelling*, 2(4):333–349.

Cutler, A. (2010). Remembering Leo Breiman. *The Annals of Applied Statistics*, 4(4):1621–1633.

Das, M., Newey, W. K., & Vella, F. (2003). Nonparametric Estimation of Sample Selection Models. *The Review of Economic Studies*, 70(1):33–58.

Das, P. & Ghosal, S. (2017). Bayesian quantile regression using random B-spline series prior. *Computational Statistics & Data Analysis*, 109:121–143.

Davino, C., Furno, M., & Vistocco, D. (2014). *Quantile Regression: Theory and Applications.* John Wiley Sons Ltd.

De Boor, C. (1977). Package for Calculating with B-Splines. *SIAM Journal on Numerical Analysis*, 14(3):441–472.

Demirer, R., Pierdzioch, C., & Zhang, H. (2016). On the short-term predictability of stock returns: A quantile boosting approach. *Finance Research Letters.*

Denison, D. D., Hansen, M. H., Holmes, C. C., Mallick, B., & Yu, B. (2013). *Nonlinear estimation and classification.* Springer Science & Business Media.

Di Paolo, A. & Tansel, A. (2015). Returns to Foreign Language Skills in a Developing Country: The Case of Turkey. *The Journal of Development Studies*, 51(4):407–421.

Dierckx, P. (1995). *Curve and Surface Fitting with Splines.* Oxford University Press.

Dietterich, T. G. (2000a). An Experimental Comparison of Three Methods for Constructing Ensembles of Decision Trees: Bagging, Boosting, and Randomization. *Machine learning*, 40(2):139–157.

Dietterich, T. G. (2000b). Ensemble Methods in Machine Learning. In *Multiple classifier systems*, pages 1–15. Springer.

Dinse, G. E. & Lagakos, S. (1983). Regression Analysis of Tumour Prevalence Data. *Applied Statistics*, pages 236–248.

Dixon, J. & Macarov, D. (2016). *Social Welfare in Socialist Countries*. Routledge.

Donald, S. G., Green, D. A., & Paarsch, H. J. (2000). Differences in Wage Distributions Between Canada and the United States: An Application of a Flexible Estimator of Distribution Functions in the Presence of Covariates. *The Review of Economic Studies*, 67(4):609–633.

Donato, K. M., Hiskey, J., Durand, J., & Massey, D. S. (2010). Migration in the Americas: Mexico and Latin America in Comparative Context. *The ANNALS of the American Academy of Political and Social Science*, 630(1):6–17.

Draper, N. R. & Smith, H. (2014). *Applied Regression Analysis*. John Wiley & Sons.

Drucker, H., Burges, C. J. C., Kaufman, L., Smola, A. J., & Vapnik, V. (1997). Support Vector Regression Machines. In Mozer, M. C., Jordan, M. I., & Petsche, T., editors, *Advances in Neural Information Processing Systems 9*, pages 155–161. MIT Press.

Drucker, H., Schapire, R., & Simard, P. (1993). Boosting Performance in Neural Networks. *International Journal of Pattern Recognition and Artificial Intelligence*, 7(04):705–719.

Durand, J., Massey, D. S., & Zenteno, R. M. (2001). Mexican Immigration to the United States: Continuities and Changes. *Latin American Research Review*, 36(1):107–127.

Durbán, M., Lee, D.-J., & Ugarte, M. D. (2008). *Splines con Penalizaciones: Teoría y Aplicaciones*. Universidad Pública de Navarra= Nafarroako Unibertsitate Publikoa.

Eckstein, S. (2009). *The Immigrant Divide: How Cuban Americans Changed the US and Their Homeland*. Routledge. Taylor & Francis Group.

Efron, B. (1979). Bootstrap Methods: Another Look at the Jackknife. *The Annals of Statistics*, 7(1):1–26.

Efron, B., Hastie, T., Johnstone, I., Tibshirani, R., et al. (2004). Least Angle Regression. *The Annals of Statistics*, 32(2):407–499.

Eicker, F. et al. (1963). Asymptotic Normality and Consistency of the Least Squares Estimators for Families of Linear Regressions. *The Annals of Mathematical Statistics*, 34(2):447–456.

Eilers, P. H. & Marx, B. D. (1996). Flexible smoothing with B-splines and penalties. *Statistical science*, 11(2):89–102.

Elith, J. & Leathwick, J. R. (2009). Species Distribution Models: Ecological Explanation and Prediction Across Space and Time. *Annual Review of Ecology, Evolution, and Systematics*, 40(1):677.

Elliott, R. J. & Lindley, J. K. (2008). Immigrant Wage Differentials, Ethnicity and Occupational Segregation. *Journal of the Royal Statistical Society: Series A*, 171(3):645–671.

Engle, R. F., Granger, C. W., Rice, J., & Weiss, A. (1986). Semiparametric Estimates of the Relation Between Weather and Electricity Sales. *Journal of the American statistical Association*, 81(394):310–320.

Eubank, R. L. (1999). *Nonparametric Regression and Spline Smoothing*. CRC press.

Fagen, R. R. & Brody, R. A. (1964). Cubans in Exile: A Demographic Analysis. *Social Problems*, 11(4):389–401.

Fahrmeir, L. & Kneib, T. (2009). Propriety of Posteriors in Structured Additive Regression Models: Theory and Empirical Evidence. *Journal of Statistical Planning and Inference*, 139(3):843–859.

Fahrmeir, L., Kneib, T., & Lang, S. (2004). Penalized Structured Additive Regression for Space-Time Data: A Bayesian Perspective. *Statistica Sinica*, 14(3):731–761.

Fahrmeir, L., Kneib, T., Lang, S., & Marx, B. (2013). *Regression: Models, Methods and Applications*. Springer.

Fahrmeir, L. & Lang, S. (2001a). Bayesian Inference for Generalized Additive Mixed Models Based on Markov Random Field Priors. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 50(2):201–220.

Fahrmeir, L. & Lang, S. (2001b). Bayesian Semiparametric Regression Analysis of Multicategorical Time-Space Data. *Annals of the Institute of Statistical Mathematics*, 53(1):11–30.

Fahrmeir, L., Wolff, J., & Bender, S. (2003). Semiparametric bayesian time-space analysis of unemployment duration. *Journal of the german Statistical Society*, 87:281–307.

Fairlie, R. W., Zissimopoulos, J., & Krashinsky, H. (2010). The International Asian Business Success Story? A Comparison of Chinese, Indian and other Asian Businesses in the United States, Canada and United Kingdom. In *International Differences in Entrepreneurship*, pages 179–208. University of Chicago Press.

Fan, J., Gijbels, I., Hu, T.-C., & Huang, L.-S. (1996). A Study of Variable Bandwidth Selection for Local Polynomial Regression. *Statistica Sinica*, pages 113–127.

Fan, J. & Li, R. (2001). Variable Selection Via Nonconcave Penalized Likelihood and its Oracle Properties. *Journal of the American Statistical Association*, 96(456):1348–1360.

Fan, J. & Zhang, W. (2008). Statistical Methods with Varying Coefficient Models. *Statistics and its Interface*, 1(1):179–195.

Fan, W., Stolfo, S. J., Zhang, J., & Chan, P. K. (1999). AdaCost: Misclassification Cost-Sensitive Boosting. In *Proceedings of the Sixteenth International Conference on Machine Learning*, volume 99 of *ICML '99*, pages 97–105, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Faraway, J. J. (2014). *Linear Models with R*. CRC Press.

Faraway, J. J. (2016). *Extending the Linear Model with R: Generalized Linear, Mixed Effects and Nonparametric Regression Models*. CRC press.

Farebrother, R. W. (1999). The Methods of Boscovich and Mayer. In *Fitting Linear Relationships: A History of the Calculus of Observations 1750-1900*, pages 9–22. Springer New York.

Feng, G., Huang, G.-B., Lin, Q., & Gay, R. (2009). Error Minimized Extreme Learning Machine with Growth of Hidden Nodes and Incremental Learning. *IEEE Transactions on Neural Networks*, 20(8):1352–1357.

Fenske, N., Burns, J., Hothorn, T., & Rehfuess, E. A. (2013a). Understanding Child Stunting in India: A Comprehensive Analysis of Socio-Economic, Nutritional and Environmental Determinants Using Additive Quantile Regression. *PloS ONE*, 8(11):e78692.

Fenske, N., Fahrmeir, L., Hothorn, T., Rzehak, P., & Hohle, M. (2013b). Boosting Structured Additive Quantile Regression for Longitudinal Childhood Obesity Data. *The International Journal of Biostatistics*, 9(1):1–18.

Fenske, N., Kneib, T., & Hothorn, T. (2012). Identifying Risk Factors for Severe Childhood Malnutrition by Boosting Additive Quantile Regression. *Journal of the American Statistical Association*, 106(494):494–510.

Fox, B. L. & Byker, E. J. (2015). Searching for Equity in Education: A Critical Ethnographic Exploration in CUBA. *Journal of Ethnographic & Qualitative Research*, 9(3).

Freund, Y. (1995). Boosting a Weak Learning Algorithm by Majority. *Information and Computation*, 121(2):256–285.

Freund, Y. & Schapire, R. E. (1996). Experiments with a New Boosting Algorithm. In *In Proceedings of the Thirteenth International Conference on Machine Learning*, volume 96, pages 148–156. Morgan Kaufmann Publishers Inc., San Francisco, CA.

Freund, Y. & Schapire, R. E. (1997). A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting. *Journal of Computer and System Sciences*, 55(1):119–139.

Friedberg, R. M. (1992). The Labor Market Assimilation of Immigrants in the United States: The Role of Age at Arrival. In *Mimeograph*. Brown University, Providence R.I.

Friedman, J., Hastie, T., & Tibshirani, R. (2001). *The Elements of Statistical Learning*. Springer Series in Statistics Springer, Berlin.

Friedman, J., Hastie, T., Tibshirani, R., et al. (2000). Additive Logistic Regression: A Statistical View of Boosting (With Discussion and a Rejoinder by the Authors). *The Annals of Statistics*, 28(2):337–407.

Friedman, J. H. (2001). Greedy Function Approximation: A gradient Boosting Machine. *The Annals of Statistics*, 29(5):1189–1232.

Friedman, J. H. (2002). Stochastic Gradient Boosting. *Computational Statistics & Data Analysis*, 38(4):367–378.

Friedman, J. H. & Silverman, B. W. (1989). Flexible Parsimonious Smoothing and Additive Modeling. *Technometrics*, 31(1):3–21.

Friedmann, J. (1991). Multivariate Adaptive Regression Splines. *The Annals of Statistics*, 19(1):1–67.

Fung, G. & Mangasarian, O. L. (2001). Proximal Support Vector Machine Classifiers. In Provost, F. & Srikant, R., editors, *Proceedings KDD-2001: Knowledge Discovery and Data Mining, August 26-29, 2001, San Francisco, CA*, pages 77–86, New York. Asscociation for Computing Machinery.

Furno, M. (2007). Parameter Instability in Quantile Regression. *Statistical Modelling*, 7(4):345–362.

Gaglianone, W. P., Lima, L. R., Linton, O., & Smith, D. R. (2011). Evaluating Value-at-Risk Models via Quantile Regression. *Journal of Business & Economic Statistics*, 29(1):150–160.

Gaglianone, W. P., Lima, L. R., Linton, O., & Smith, D. R. (2012). Evaluating Value-at-Risk Models via Quantile Regression. *Journal of Business & Economic Statistics*.

Gagliardini, P. & Scaillet, O. (2012). Nonparametric Instrumental Variable Estimation of Structural Quantile Effects. *Econometrica*, 80(4):1533–1562.

Galar, M., Fernández, A., Barrenechea, E., Bustince, H., & Herrera, F. (2012). A Review on Ensembles for the Class Imbalance Problem: Bagging, Boosting and Hybrid-Based Approaches. *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on*, 42(4):463–484.

Galton, F. (1886). Regression towards Mediocrity in Hereditary Stature. *The Journal of the Anthropological Institute of Great Britain and Ireland*, 15:246–263.

Galvao, A. F. & Montes-Rojas, G. V. (2010). Penalized Quantile Regression for Dynamic Panel Data. *Journal of Statistical Planning and Inference*, 140(11):3476–3497.

Gasperini, L. (2000). The Cuban Education System: Lessons and Dilemmas.

Gastil, R. D. (1987). *Freedom in the World*. Greenwood Press.

Gelbach, J. B. (2016). When Do Covariates Matter? And Which Ones, and How Much? *Journal of Labor Economics*, 34(2):509–543.

Green, P. J. & Silverman, B. W. (1993). *Nonparametric Regression and Generalized Linear Models: A Roughness Penalty Approach*. CRC Press.

Greene, W. H. (1994). Accounting for Excess Zeros and Sample Selection in Poisson and Negative Binomial Regression Models.

Greene, W. H. (2003). *Econometric Analysis*. Prentice Hall.

Grenier, G. J. (2015). The Cuban-American Transition: Demographic Changes Drive Ideological Changes. In *Proceedings of the Latin American Studies Association Conference*.

Grogger, J. & Hanson, G. H. (2011). Income Maximization and the Selection and Sorting of International Migrants. *Journal of Development Economics*, 95(1):42–57.

Gronau, R. (1974). The Effect of Children on the Housewife's Value of Time. In *Economics of the family: Marriage, children, and human capital*, pages 457–490. University of Chicago Press.

Groot, W. & Van Den Brink, H. M. (2000). Overeducation in the Labor Market: A meta-Analysis. *Economics of Education Review*, 19(2):149–158.

Hamermesh, D. S. & Trejo, S. J. (2013). How do Immigrants Spend their Time? The Process of Assimilation. *Journal of Population Economics*, 26(2):507–530.

Hao, L. & Naiman, D. Q. (2007). *Quantile Regression*. Sage Publications.

Härdle, W. & Liang, H. (2007). Partially Linear Models. In *Statistical Methods for Biostatistics and Related Fields*, pages 87–103. Springer-Verlag Berlin Heidelberg GmnH.

Harris, J. R. & Todaro, M. P. (1970). Migration, Unemployment and Development: A Two-Sector Analysis. *The American Economic Review*, 60(1):126–142.

Harris, P., Fotheringham, A. S., & Juggins, S. (2010). Robust Geographically Weighted Regression: A Technique for Quantifying Spatial Relationships Between Freshwater Acidification Critical Loads and Catchment Attributes. *Annals of the Association of American Geographers*, 100(2):286–306.

Hasebe, T. & Vijverberg, W. P. (2012). A Flexible Sample Selection Model: A GTL-Copula Approach. Discussion Paper 7003, IZA.

Hastie, T. & Tibshirani, R. (1993). Varying-Coefficient Models. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 757–796.

Hastie, T., Tibshirani, R., Friedman, J., & Franklin, J. (2005). The Elements of Statistical Learning: Data Mining, Inference and Prediction. *The Mathematical Intelligencer*, 27(2):83–85.

Hastie, T. J. & Tibshirani, R. J. (1990). *Generalized Additive Models*. CRC Press.

Haupt, H., Lösel, F., & Stemmler, M. (2014). Quantile Regression Analysis and Other Alternatives to Ordinary Least Squares: A Methodological Comparison on Corporal Punishment. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences*, 10(3):81–91.

Heckman, J. (1974). Shadow Prices, Market Wages, and Labor Supply. *Econometrica*, 42(4):679–694.

Heckman, J. (1976). The Common Structure of Statistical Models of Truncation, Sample Selection and Limited Dependent Variables and a Simple Estimator for such Models. *Annals of Economic and Social Measurement*, 5(4):475–492.

Heckman, J. (1979). Sample Selection Bias as a Specification Error. *Econometrica*, 47(1):153–161.

Heckman, N. E. (1986). Spline Smoothing in a Partly Linear Model. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 244–248.

Hendricks, W. & Koenker, R. (1992). Hierarchical Spline Models for Conditional Quantiles and the Demand for Electricity. *Journal of the American statistical Association*, 87(417):58–68.

Hofner, B., Boccuto, L., & Göker, M. (2015). Controlling False Discoveries in High-Dimensional Situations: Boosting with Stability Selection. *BMC bioinformatics*, 16(1):144.

Hofner, B., Hothorn, T., Kneib, T., & Schmid, M. (2011). A Framework for Unbiased Model Selection Based on Boosting. *Journal of Computational and Graphical Statistics*, 20(4):956–971.

Hofner, B., Mayr, A., Robinzonov, N., & Schmid, M. (2014). Model-Based Boosting in R: A Hands-on Tutorial using the R Package Mboost. *Computational Statistics*, 29(1-2):3–35.

Honda, T. (2004). Quantile Regression in Varying Coefficient Models. *Journal of Statistical Planning and Inference*, 121(1):113–125.

Horowitz, J. L. (1998). Bootstrap Methods for Median Regression Models. *Econometrica*, pages 1327–1351.

Horowitz, J. L. & Lee, S. (2007). Nonparametric Instrumental Variables Estimation of a Quantile Regression Model. *Econometrica*, 75(4):1191–1208.

Hothorn, T., Buhlmann, P., Kneib, T., Schmid, M., & Hofner, B. (2010). Model-Based Boosting 2.0. *The Journal of Machine Learning Research*, 11:2109–2113.

Hothorn, T., Buhlmann, P., Kneib, T., Schmid, M., & Hofner, B. (2015). Mboost: Model-Based Boosting. R package version 2.4-2.

Huang, G.-B. (2015). What Are Extreme Learning Machines? Filling the Gap Between Frank Rosenblattâs Dream and John Von Neumannâs Puzzle. *Cognitive Computation*, 7(3):263–278.

Huang, G.-B., Zhou, H., Ding, X., & Zhang, R. (2012). Extreme Learning Machine for Regression and Multiclass Classification. *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on*, 42(2):513–529.

Huang, G.-B., Zhu, Q.-Y., Mao, K., Siew, C.-K., Saratchandran, P., & Sundararajan, N. (2006a). Can Threshold Networks Be Trained Directly? *IEEE Transactions on Circuits and Systems Part 2: Express Briefs*, 53(3):187–191.

Huang, G.-B., Zhu, Q.-Y., & Siew, C.-K. (2006b). Extreme Learning Machine: Theory and Applications. *Neurocomputing*, 70(1):489–501.

Huang, J. Z., Wu, C. O., & Zhou, L. (2004). Polynomial Spline Estimation and Inference for Varying Coefficient Models with Longitudinal Data. *Statistica Sinica*, pages 763–788.

Huber, M. & Melly, B. (2015). A Test of the Conditional Independence Assumption in Sample Selection Models. *Journal of Applied Econometrics*, 30(7):1144–1168.

Huber, P. J. (1967). The Behavior of Maximum Likelihood Estimates under Nonstandard Conditions. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, volume I, pages 221–233. University of California Press.

Hunt, P. (2012). From the Bottom to the Top: A More Complete Picture of the Immigrant-Native Wage Gap in Britain. *IZA Journal of Migration*, 1(1):1–18.

Ibarraran, P. & Lubotsky, D. (2007). Mexican Immigration and Self-Selection: New Evidence from the 2000 Mexican Census. In *Mexican immigration to the United States*, pages 159–192. University of Chicago Press.

Ichimura, H. (1993). Semiparametric Least Squares (SLS) and Weighted SLS Estimation of Single–Index Models. *Journal of Econometrics*, 58(1):71–120.

Imbens, G. W. & Newey, W. K. (2009). Identification and Estimation of Triangular Simultaneous Equations Models without Additivity. *Econometrica*, 77(5):1481–1512.

Joshi, M. V., Kumar, V., & Agarwal, R. C. (2001). Evaluating Boosting Algorithms to Classify Rare Classes: Comparison and Improvements. In *Data Mining, 2001. ICDM 2001, Proceedings IEEE International Conference on*, pages 257–264. IEEE.

Kaestner, R. & Malamud, O. (2014). Self-Selection and International Migration: New Evidence from Mexico. *Review of Economics and Statistics*, 96(1):78–91.

Kandala, N.-B., Fahrmeir, L., Klasen, S., & Priebe, J. (2009). Geo-Additive Models of Childhood Undernutrition in Three Sub-Saharan African Countries. *Population, Space and Place*, 15(5):461–473.

Kandala, N.-B., Madungu, T. P., Emina, J. B., Nzita, K. P., & Cappuccio, F. P. (2011). Malnutrition Among Children under the Age of Five in the Democratic Republic of Congo (DRC): Does Geographic Location Matter? *BMC Public Health*, 11(1):1–15.

Karemera, d., Iwuagwu Oguledo, V., & Davis, B. (2000). A Gravity Model Analysis of International Migration to North America. *Applied Economics*, 32(13):1745–1755.

Kaushal, N., Lu, Y., Denier, N., Wang, J. S.-H., & Trejo, S. J. (2016). Immigrant Employment and Earnings Growth in Canada and the USA: Evidence from Longitudinal Data. *Journal of Population Economics*, 29(4):1249–1277.

Kearns, M. & Valiant, L. (1994). Cryptographic Limitations on Learning Boolean Formulae and Finite Automata. *Journal of the ACM (JACM)*, 41(1):67–95.

Khan, A. H. (1997). Post-Migration Investment in Education by Immigrants in the United States. *The Quarterly Review of Economics and Finance*, 37:285–313.

Kim, M.-O. (2007). Quantile Regression with Varying Coefficients. *The Annals of Statistics*, pages 92–108.

Kim, T.-H. & White, H. (2003). Estimation, Inference, and Specification Testing for Possibly Misspecified Quantile Regression. *Advances in Econometrics*, 17:107–132.

Kleiber, C. & Zeileis, A. (2008). *Applied Econometrics with R*. Springer Science & Business Media.

Klein, R. W. & Spady, R. H. (1993). An Efficient Semiparametric Estimator for Binary Response Models. *Econometrica*, 61(2):387–421.

Kleven, H. J., Landais, C., Saez, E., & Schultz, E. (2014). Migration and Wage Effects of Taxing Top Earners: Evidence from the Foreigners' Tax Scheme in Denmark. *Quarterly Journal of Economics*, 129(1).

Kneib, T. (2013). Beyond Mean Regression. *Statistical Modelling*, 13(4):275–303.

Kneib, T., Hothorn, T., & Tutz, G. (2009). Variable Selection and Model Choice in Geoadditive Regression Models. *Biometrics*, 65(2):626–634.

Koenker, R. (2004). Quantile regression for longitudinal data. *Journal of Multivariate Analysis*, 91(1):74–89.

Koenker, R. (2005). *Quantile regression*. Cambridge University Press: Cambridge, UK.

Koenker, R. (2011). Additive models for quantile regression: Model selection and confidence bandaids. *Brazilian Journal of Probability and Statistics*, 25(3):239–262.

Koenker, R. & Bassett, G. (1978). Regression Quantiles. *Econometrica: Journal of the Econometric Society*, pages 33–50.

Koenker, R. & Bassett, G. (1982). Robust Tests for Heteroscedasticity Based on Regression Quantiles. *Econometrica: Journal of the Econometric Society*, pages 43–61.

Koenker, R. & Hallock, K. (2001). Quantile Regression: An Introduction. *Journal of Economic Perspectives*, 15(4):43–56.

Koenker, R. & Machado, J. A. (1999). Goodness of Fit and Related Inference Processes for Quantile Regression. *Journal of the american statistical association*, 94(448):1296–1310.

Koenker, R. & Mizera, I. (2004). Penalized Triograms: Total Variation Regularization for Bivariate Smoothing. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 66(1):145–163.

Koenker, R., Ng, P., & Portnoy, S. (1994). Quantile Smoothing Splines. *Biometrika*, 81(4):673–680.

Kottas, A. & Gelfand, A. E. (2001). Bayesian Semiparametric Median Regression Modeling. *Journal of the American Statistical Association*, 96(456):1458–1468.

Kottas, Athanasios and Krnjajic, Milovan (2009). Bayesian Semiparametric Modelling in Quantile Regression. *Scandinavian Journal of Statistics*, 36(2):297–319.

Kriegler, B. & Berk, R. (2010). Small Area Estimation of the Homeless in Los Angeles: An Application of Cost-Sensitive Stochastic Gradient Boosting. *The Annals of Applied Statistics*, 4(3):1234–1255.

Kuhlenkasper, T. & Steinhardt, M. F. (2012). Who Leaves and When: Selective outmigration of immigrants from Germany. *Hamburg Institute of International Economics (HWWI) Research Paper*, (128).

Kuhn, M. & Johnson, K. (2013). *Applied Predictive Modeling*. Springer.

Lang, S. & Brezger, A. (2004). Bayesian P-splines. *Journal of Computational and Graphical Statistics*, 13(1):183–212.

Lang, S., Umlauf, N., Wechselberger, P., Harttgen, K., & Kneib, T. (2014). Multilevel Structured Additive Regression. *Statistics and Computing*, 24(2):223–238.

Langford, J., Oliveira, R., & Zadrozny, B. (2012). Predicting Conditional Quantiles Via Reduction to Classification. In *Proceedings of Uncertainty in Artificical Intelligence*, pages 257â–264.

Leathwick, J., Elith, J., Francis, M., Hastie, T., & Taylor, P. (2006). Variation in Demersal Fish Species Richness in the Oceans Surrounding New Zealand: An Analysis Using Boosted Regression Trees. *Marine Ecology Progress Series*, 321:267–281.

Lee, D. S. (2009). Training, Wages and Sample Selection: Estimating Sharp Bounds on Treatment Effects. *Review of Economics Studies*, 76(3):1071–1102.

Lee, S. (2007). Endogeneity in Quantile Regression Models: A Control Function Approach. *Journal of Econometrics*, 141(2):1131–1158.

Lee, Y. J. & Mangasarian, O. L. (2001). RSVM: Reduced Support Vector Machines Classifier. In *Proceedings of the 1st SIAM International Conference on Data Mining. Chicago, IL, 2001, April. 5-7*, volume 1, pages 325–361. SIAM.

Lemieux, T. (2006). Increasing Residual Wage Inequality: Composition Effects, Noisy Data, or Rising Demand for Skill? *The American Economic Review*, pages 461–498.

Lewis, W. A. (1954). Economic Development with Unlimited Supplies of Labour. *The Manchester School*, 22(2):139–191.

Li, M. B., Huang, G. B., Saratchandran, P., & Sundararajan, N. (2005). Fully Complex Extreme Learning Machine. *Neurocomputing*, 68:306–314.

Li, Q. & Sweetman, A. (2014). The Quality of Immigrant Source Country Educational Outcomes: Do They Matter in the Receiving Country? *Labour Economics*, 26:81–93.

Li, Q., Xi, R., Lin, N., et al. (2010). Bayesian Regularized Quantile Regression. *Bayesian Analysis*, 5(3):533–556.

Li, R. & Liang, H. (2008). Variable Selection in Semiparametric Regression Modeling. *Annals of statistics*, 36(1):261.

Liang, N. Y., Huang, G.-B., Saratchandran, P., & Sundararajan, N. (2006). A Fast and Accurate Online Sequential Learning Algorithm for Feedforward Networks. *Neural Networks, IEEE Transactions on*, 17(6):1411–1423.

Lin, H., Song, P. X., & Zhou, Q. M. (2007). Varying-Coefficient Marginal Models and Applications in Longitudinal Data Analysis. *Sankhyā: The Indian Journal of Statistics*, 69(3):581–614.

Long, J. E. (1980). The Effect of Americanization on Earnings: Some Evidence for Women. *Journal of Political Economy*, 88(3):620–629.

López, G. (2015). Hispanics of Cuban Origin in the United States, 2013. *Washington, DC, Pew Research Center*.

Lou, Y., Bien, J., Caruana, R., & Gehrke, J. (2015). Sparse Partially Linear Additive Models. *Journal of Computational and Graphical Statistics*, (just-accepted):1–31.

Lowell, L., Pederzini, C., Passel, J., Escobar, A., & Martin, S. (2008). The Demography of Mexico/US Migration. *Mexico US Migration Management: A Binational Approach, Lanham: Lexington Books*, pages 1–31.

Ma, L. & Koenker, R. (2006). Quantile Regression Methods for Recursive Structural Equation Models. *Journal of Econometrics*, 134(2):471–506.

Ma, S. & Song, P. X.-K. (2015). Varying Index Coefficient Models. *Journal of the American Statistical Association*, 110(509):341–356.

Machado, J. & Silva, S. (2000). Glejser's Test Revisited. *Journal of Econometrics*, 97(1):189–202.

Machado, J. A. & Mata, J. (2005). Counterfactual Decomposition of Changes in Wage Distributions Using Quantile Regression. *Journal of Applied Econometrics*, 20(4):445–465.

Machado, J. A. & Silva, J. (2013). Quantile Regression and Heteroskedasticity. Technical report, Discussion Paper, University of Essex, Department of Economics.

Manski, C. F. & Sims, C. (1994). The Selection Problem. In *Advances in Econometrics, Sixth World Congress*, volume 1, pages 143–70.

Marra, G. & Radice, R. (2013). A Penalized Likelihood Estimation Approach to Semiparametric Sample Selection Binary Response Modeling. *Electronic Journal of Statistics*, 7:1432–1455.

Martin, J. A., Hamilton, B. E., Osterman, M., Curtin, S. C., & Matthews, T. (2015). Births: Final data for 2013. *National Vital Statistics Reports*, 64(1):1–65.

Martínez-Silva, I., Lustres-Pérez, V., Lorenzo-Arribas, A., Roca-Pardiñas, J., & Cadarso-Suárez, C. (2013). Flexible Quantile Regression Models: Application to the Study of the Purple Sea Urchin. *SORT-Statistics and Operations Research Transactions*, 37(1):81–94.

Martins, P. S. & Pereira, P. T. (2004). Does Education Reduce Wage Inequality? Quantile Regression Evidence from 16 Countries. *Labour Economics*, 11(3):355–371.

Marx, B. D. & Eilers, P. H. (1998). Direct generalized additive modeling with penalized likelihood. *Computational Statistics & Data Analysis*, 28(2):193–209.

Mason, L., Baxter, J., Bartlett, P. L., Frean, M., et al. (2000). Functional Gradient Techniques for Combining Hypotheses. *Advances in Large Margin Classifiers*, pages 221–246.

Matos, A. D. D. & Liebig, T. (2014). The Qualifications of Immigrants and their Value in the Labour Market. pages 187–228.

Mayda, A. M. (2010). International Migration: A Panel Data Analysis of the Determinants of Bilateral Flows. *Journal of Population Economics*, 23(4):1249–1274.

Mayr, A., Hofner, B., & Schmid, M. (2016). Boosting the discriminatory power of sparse survival models via optimization of the concordance index and stability selection. *BMC Bioinformatics*, 17(1).

Mayr, A., Hofner, B., Schmid, M., et al. (2012a). The Importance of Knowing When to Stop. *Methods of Information in Medicine*, 51(2):178–186.

Mayr, A., Hothorn, T., & Fenske, N. (2012b). Prediction Intervals for Future BMI Values of Individual Children: A Non–Parametric Approach by Quantile Boosting. *BMC Medical Research Methodology*, 12(1):1.

McCullagh, P. & Nelder, J. A. (1989). *Generalized Linear Models*. 2nd edition. Chapman-Hall, London.

McKenzie, D. & Rapoport, H. (2010). Self- Selection Patterns in Mexico-US Migration: The Role of Migration Networks. *The Review of Economics and Statistics*, 92(4):811–821.

Meinshausen, N. (2006). Quantile Regression Forests. *The Journal of Machine Learning Research*, 7:983–999.

Meinshausen, N. & Bühlmann, P. (2010). Stability Selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(4):417–473.

Miller, T., Kim, A. B., & Holmes, K. (2015). 2015 Index of Economic Freedom. *The Heritage Foundation*.

Mincer, J. A. (1974). Age and Experience Profiles of Earnings. In *Schooling, Experience and Earnings*, pages 64–82. NBER.

Mitchell, J., Pate, R., Beets, M., & Nader, P. (2013). Time Spent in Sedentary Behavior and Changes in Childhood BMI: A Longitudinal Study from Ages 9 to 15 Years. *International journal of obesity*, 37(1):54–60.

Moraga, J. F.-H. (2011). New Evidence on Emigrant Selection. *The Review of Economics and Statistics*, 93(1):72–96.

Motel, S. & Patten, E. (2012). Hispanic Origin Profiles, 2010. *Pew Hispanic Center*.

Mullahy, J. (1986). Specification and Testing of Some Modified Count Data Models. *Journal of econometrics*, 33(3):341–365.

Nair, N. U., Sankaran, P., & Balakrishnan, N. (2013). *Quantile-Based Reliability Analysis*. Springer.

Natekin, A. & Knoll, A. (2013). Gradient Boosting Machines: A Tutorial. *Frontiers in neurorobotics*, 7.

Nelder, J. & Wedderburn, R. (1972). Generalized Linear Models. *Journal of Royal Statistical Society Series A*, 135:370–384.

Neocleous, T. & Portnoy, S. (2008). On Monotonicity of Regression Quantile Functions. *Statistics & Probability Letters*, 78(10):1226–1229.

Newey, W. K. (1991). Uniform Convergence in Probability and Stochastic Equicontinuity. *Econometrica*, 59(4):1161–1167.

Newey, W. K. (2009). Two-Step Series Estimation of Sample Selection Models. *The Econometrics Journal*, 12(s1):S217–S229.

Newey, W. K. & Powell, J. L. (1987). Asymmetric Least Squares Estimation and Testing. *Econometrica: Journal of the Econometric Society*, pages 819–847.

Nychka, D., Gray, G., Haaland, P., Martin, D., & O'connell, M. (1995). A Nonparametric Regression Approach to Syringe Grading for Quality Improvement. *Journal of the American Statistical Association*, 90(432):1171–1178.

Oh, H. S., Lee, T. C., & Nychka, D. W. (2011). Fast Nonparametric Quantile Regression with Arbitrary Smoothing Methods. *Journal of Computational and Graphical Statistics*, 20:510–526.

ONE (2010). *Demographic Yearbook of Cuba*. Oficina Nacional de Estadśtica e Información.

Orrenius, P. M. & Zavodny, M. (2005). Self-Selection among Undocumented Immigrants from Mexico. *Journal of Development Economics*, 78(1):215–240.

O'Sullivan, F. (1988). Nonparametric Estimation of Relative Risk Using Splines and Cross-Validation. *SIAM Journal on Scientific and Statistical Computing*, 9(3):531–542.

Pérez, L. (1986). Immigrant Economic Adjustment and Family Organization: The Cuban Success Story Reexamined. *International Migration Review*, pages 4–20.

Pérez de La Riva, J. (2000). Los Culíes Chinos en Cuba. *La Habana: Editorial Ciencias Sociales*.

Petscher, Y. & Logan, J. A. R. (2013). Quantile Regression in the Study of Developmental Sciences. *Child Development*, 85(3):861–881.

Phelan, J., Cuffney, T., Patterson, L., Eddy, M., Dykes, R., Pearsall, S., Goudreau, C., Mead, J., & Tarver, F. (2017). Fish and Invertebrate Flow-Biology Relationships to Support the Determination of Ecological Flows for North Carolina. *JAWRA Journal of the American Water Resources Association*, 53(1):42–55.

Picchio, M. & Mussida, C. (2011). Gender Wage Gap: A Semi–Parametric Approach with Sample Selection Correction. *Labour Economics*, 18(5):564–578.

Pierdzioch, C., Risse, M., & Rohloff, S. (2016). A quantile-boosting approach to forecasting gold returns. *The North American Journal of Economics and Finance*, 35:38–55.

Pigini, C. (2015). Bivariate Non–Normality in the Sample Selection Model. *Journal of Econometric Methods*, 4(1):123–144.

Pirttilä, J. (2004). Is International Labour Mobility a Threat to the Welfare State? Evidence from Finland in the 1990s. *Finnish Economic Papers*, 17(1):18–34.

Portes, A. (2010). Migration and Social Change: Some Conceptual Reflections. *Journal of Ethnic and Migration Studies*, 36(10):1537–1563.

Portes, A., Ansley, F., & Shefner, J. (2009). *The New Latin Nation: Immigration and the Hispanic Population of the United States*, pages 3–34. Univ. of Tennessee Press Knoxville.

Portes, A. & Puhrmann, A. (2015). A Bifurcated Enclave: The Economic Evolution of the Cuban and Cuban American Population of Metropolitan Miami. *Cuban Studies*, 43:40–63.

Powell, J. L. (1984). Least Absolute Deviations Estimation for the Censored Regression Model. *Journal of Econometrics*, 25(3):303–325.

Pumar, E. S. (2013). Poverty and the Effects on Aversive Social Control. *Cuba in Transition*, 23:364–368.

Ratha, D., Yi, S., & Yousefi, S. R. (2015). Migration and development. *Routledge Handbook of Immigration and Refugee Studies*, 1(3):260.

Reich, B. J., Bondell, H. D., & Wang, H. J. (2010). Flexible Bayesian Quantile Regression for Independent and Clustered Data. *Biostatistics*, 11(2):337–352.

Ridgeway, G. (1999). The State of Boosting. *Computing Science and Statistics*, pages 172–181.

Ridgeway, G. (2007). Generalized boosted models: A guide to the gbm package. *Available at http://cran. open-source-solution. org/web/packages/gbm/vignettes/gbm. pdf.*

Ridgeway, G. (2010). GBM: Generalized Boosted Regression Models. R Package Version 1.6-3.1.

Rigby, R. A. & Stasinopoulos, D. M. (2005). Generalized Additive Models for Location, Scale and Shape. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 54(3):507–554.

Rigby, R. A. & Stasinopoulos, D. M. (2006). Using the Box-Cox t Distribution in GAMLSS to Model Skewness and Kurtosis. *Statistical Modelling*, 6(3):209–229.

Rigby, R.A and Stasinopoulos, D.M (2001). The GAMLSS Project: A Flexible Approach to Statistical Modelling. In *New Trends in Sstatistical Modelling: Proceedings of the 16th International Workshop on Statistical Modelling*, pages 249–256.

Rong, H.-J., Huang, G.-B., Sundararajan, N., & Saratchandran, P. (2009). Online Sequential Fuzzy Extreme Learning Machine for Function Approximation and Classification Problems. *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on*, 39(4):1067–1072.

Roy, A. D. (1951). Some Thoughts on the Distribution of Earnings. *Oxford economic papers*, 3(2):135–146.

Rue, H., Martino, S., & Chopin, N. (2009). Approximate Bayesian Inference for Latent Gaussian Models by Using Integrated Nested Laplace Approximations. *Journal of the royal statistical society: Series b (statistical methodology)*, 71(2):319–392.

Ruggles, S., King, M. L., Levison, D., McCaa, R., & Sobek, M. (2011a). In *Integrated Public Use Microdata Series (IPUMS), Version 5.0, [Machine-readable database]*. University of Minnesota.

Ruggles, S., King, M. L., Levison, D., McCaa, R., & Sobek, M. (2011b). In *Integrated Public Use Microdata Series, International: Version 6.4 [dataset]*. University of Minnesota.

Ruggles, S., King, M. L., Levison, D., McCaa, R., & Sobek, M. (2013). *Integrated Public Use Microdata Series (IPUMS), Version 5.0*. University of Minnesota.

Ruppert, D., Wand, M. P., & Carroll, R. J. (2003). *Semiparametric Regression*. Number 12. Cambridge university press.

Ryan, C. L. & Bauman, K. (2016). Educational Attainment in the United States: 2015. Current population reports, US Census Bureau, Washington, DC.

Sainz-Cano, H., Marrero-Peniche, G., & Ménendez-Pérez, D. (2015). Los Cubanos en el Rompecabezas Estadounidense/Cubans in the US puzzle. *Mundi Migratios*, 3(1):1–47.

Schapire, R. E. (1990). The Strength of Weak Learnability. *Machine learning*, 5(2):197–227.

Schapire, R. E. (2003). The Boosting Approach to Machine Learning: An Overview. In *Nonlinear estimation and classification*, pages 149–171. Springer.

Schapire, R. E. & Singer, Y. (1999). Improved Boosting Algorithms Using Confidence-Rated Predictions. *Machine learning*, 37(3):297–336.

Scheipl, F., Kneib, T., & Fahrmeir, L. (2013). Penalized Likelihood and Bayesian Function Selection in Regression Models. *AStA Advances in Statistical Analysis*, 97(4):349–385.

Schmalensee, R. & Stoker, T. M. (1999). Household Gasoline Demand in the United States. *Econometrica*, 67(3):645–662.

Schmid, M. & Hothorn, T. (2008). Boosting additive models using component-wise p-splines. *Computational Statistics & Data Analysis*, 53(2):298–311.

Schnabel, S. K. & Eilers, P. H. (2009). Optimal Expectile Smoothing. *Computational Statistics & Data Analysis*, 53(12):4168–4177.

Schnabel, S. K. & Eilers, P. H. (2013). Simultaneous Estimation of Quantile Curves Using Quantile Sheets. *AStA Advances in Statistical Analysis*, 97(1):77–87.

Schwiebert, J. (2013). Sieve Maximum Likelihood Estimation of a Copula–Based Sample Selection Model. *IZA Discussion Paper*.

Seber, G. A. & Lee, A. J. (2012). *Linear Regression Analysis*, volume 936. John Wiley & Sons.

Shadish, W. R., Zuur, A. F., & Sullivan, K. J. (2014). Using Generalized Additive (Mixed) Models to Analyze Single Case Designs. *Journal of school psychology*, 52(2):149–178.

Shah, R. D. & Samworth, R. J. (2012). Variable selection with error control: another look at stability selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 75(1):55–80.

Sjaastad, L. A. (1962). The Costs and Returns of Human Migration. *Journal of political Economy*, 70(5, Part 2):80–93.

Skop, E. H. (2001). Race and Place in the Adaptation of Mariel Exiles. *International Migration Review*, 35(2):449–471.

Smith, P. L. (1982). Curve Fitting and Modeling with Splines Using Statistical Variable Selection Techniques. *NASA Report 166034. NASA, Langley Research Center, Hampton, VA*.

Sobotka, F. & Kneib, T. (2012). Geoadditive Expectile Regression. *Computational Statistics & Data Analysis*, 56(4):755–767.

Speckman, P. (1988). Kernel Smoothing in Partial Linear Models. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 413–436.

Stone, C. J., Hansen, M. H., Kooperberg, C., Truong, Y. K., et al. (1997). Polynomial Splines and Their Tensor Products in Extended Linear Modeling: 1994 Wald Memorial Lecture. *The Annals of Statistics*, 25(4):1371–1470.

Sun, X., Peng, L., Huang, Y., & Lai, H. J. (2016). Generalizing Quantile Regression for Counting Processes With Applications to Recurrent Events. *Journal of the American Statistical Association*, 111(513):145–156.

Sun, Y., Kamel, M. S., Wong, A. K., & Wang, Y. (2007). Cost–Sensitive Boosting for Classification of Imbalanced Data. *Pattern Recognition*, 40(12):3358–3378.

Suykens, J. A. & Vandewalle, J. (1999). Least Squares Support Vector Machine Classifiers. *Neural processing letters*, 9(3):293–300.

Takeuchi, I., Le, Q. V., Sears, T. D., & Smola, A. J. (2006). Nonparametric Quantile Estimation. *The Journal of Machine Learning Research*, 7:1231–1264.

Taylor, J. W. (2000). A Quantile Regression Neural Network Approach to Estimating the Conditional Density of Multiperiod Returns. *Journal of Forecasting*, 19(4):299–311.

Tibshirani, R. (1996). Regression Shrinkage and Selection Via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288.

Tokdar, S. T., Kadane, J. B., et al. (2012). Simultaneous Linear Quantile Regression: A Semiparametric Bayesian Approach. *Bayesian Analysis*, 7(1):51–72.

Torreira, R. & Buajasán, J. (2000). Operación Peter Pan: Un caso de guerra psicológica contra Cuba. *La Habana: Editora Política*.

Umlauf, N., Adler, D., Kneib, T., Lang, S., & Zeileis, A. (2015). Structured Additive Regression Models: An R Interface to BayesX. *Journal of Statistical Software*, 63(1):1–46.

UN (2015). Trends in International Migrant Stock: the 2015 Revision. In *Department of Economic and Social Affairs Population Division*.

US Bureau of Labor Statistics (2015). Wage and hour division. Technical report.

US Census Bureau (2010). *Geographical Mobility and Migration Main*. U.S. Department of Commerce.

US Census Bureau (2015). In *Geographical Mobility and Migration Main*.

US Custom and Border Protection (2016). In *Statistics and Accomplishments*.

Valiant, L. G. (1984). A Theory of the Learnable. *Communications of the ACM*, 27(11):1134–1142.

Vapnik, V. (2013). *The Nature of Statistical Learning Theory*. Springer.

Vapnik, V. & Chervonenkis, A. (1964). A Note on One Class of Perceptrons. *Automation and Remote control*, 25(1):821–837.

Vapnik, V. & Lerner, A. (1963). Generalized Portrait Method for Pattern Recognition. *Automation and Remote Control*, 24(6):774–780.

Vovk, V., Papadopoulos, H., & Gammerman, A. (2015). *Measures of Complexity*. Springer.

Wainer, H. (2000). The Centercept: An Estimable and Meaningful Regression Parameter. *Psychological Science*, 11(5):434–436.

Waldmann, E., Kneib, T., Yue, Y. R., Lang, S., & Flexeder, C. (2013). Bayesian Semiparametric Additive Quantile Regression. *Statistical Modelling*, 13(3):223–252.

Waltrup, L. S. & Kauermann, G. (2015). Smooth Expectiles for Panel Data Using Penalized Splines. *Statistics and Computing*, pages 1–12.

Waltrup, L. S., Sobotka, F., Kneib, T., & Kauermann, G. (2015). Expectile and Quantile Regression: David and Goliath? *Statistical Modelling*, 15(5):433–456.

Wand, M. P. (1999). On the Optimal Amount of Smoothing in Penalised Spline Regression. *Biometrika*, 86(4):936–940.

Wang, H., Li, G., & Jiang, G. (2007). Robust Regression Shrinkage and Consistent Variable Selection through the LAD-Lasso. *Journal of Business & Economic Statistics*, 25(3):347–355.

Wang, H. J., Zhu, Z., & Zhou, J. (2009). Quantile Regression in Partially Linear Varying Coefficient Models. *The Annals of Statistics*, pages 3841–3866.

Wang, Y., Feng, X.-N., & Song, X.-Y. (2016). Bayesian Quantile Structural Equation Models. *Structural Equation Modeling: A Multidisciplinary Journal*, 23(2):246–258.

Wasem, R. E. (2009). Cuban migration to the United States: Policy and Trends. CRS Report for Congress 7-5700, Congressional Research Service.

Wheeler, D. C. (2014). Geographically Weighted Regression. In *Handbook of Regional Science*, pages 1435–1459. Springer.

White, H. (1980). A Heteroskedasticity–Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity. *Econometrica: Journal of the Econometric Society*, pages 817–838.

Wiesenfarth, M. & Kneib, T. (2010). Bayesian Geoadditive Sample Selection Models. *Journal of the Royal Statistical Society: Series C*, 59(3):381–404.

Wong, C. M., Ting, L. L. O., et al. (2016). A quantile regression approach to the multiple period value at risk estimation. *Journal of Economics and Management*, 12(1):1–35.

Wood, S. N. (2003). Thin Plate Regression Splines. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 65(1):95–114.

Wood, S. N. (2006). *Generalized Additive Models: An Introduction with R*. CRC press.

Wooldridge, J. M. (2015). *Introductory Econometrics: A Modern Approach*. Nelson Education.

Wu, X., Kumar, V., Quinlan, J. R., Ghosh, J., Yang, Q., Motoda, H., McLachlan, G. J., Ng, A., Liu, B., Philip, S. Y., et al. (2008). Top 10 Algorithms in Data Mining. *Knowledge and information systems*, 14(1):1–37.

Wyatt, J. C. & Altman, D. G. (1995). Commentary: Prognostic models: clinically useful or quickly forgotten? *BMJ*, 311(7019):1539–1541.

Xiao, Z., Guo, H., & Lam, M. S. (2015). Quantile Regression and Value at Risk. In *Handbook of Financial Econometrics and Statistics*, pages 1143–1167. Springer.

Yang, Y., He, X., et al. (2012). Bayesian Empirical Likelihood for Quantile Regression. *The Annals of Statistics*, 40(2):1102–1131.

Yao, Q. & Tong, H. (1996). Asymmetric Least Squares Regression Estimation: A Nonparametric Approach. *Journal of Nonparametric Statistics*, 6(2-3):273–292.

Yashiv, E. (2008). Positive or Negative? Migrant Workers' Self Selection Revisited. Working paper, Centre for Economic Performance, London School of Economics.

Yu, K. & Moyeed, R. A. (2001). Bayesian Quantile Regression. *Statistics & Probability Letters*, 54(4):437–447.

Zavodny, M. (2003). Race, Wages and Assimilation among Cuban Immigrants. *Population Research and Policy Review*, 22(3):201–219.

Zeileis, A. (2004). Econometric Computing with HC and HAC Covariance Matrix Estimators. *Journal of Statitical Software*, 11(10):1–17.

Zeileis, A., Kleiber, C., & Jackman, S. (2008). Regression Models for Count Data in R. *Journal of Statistical Software*, 27(8):1–25.

Zhang, H. H., Cheng, G., & Liu, Y. (2012). Linear or Nonlinear? Automatic Structure Discovery for Partially Linear Models. *Journal of the American Statistical Association*.

Zhang, W., Lee, S.-Y., & Song, X. (2002). Local Polynomial Fitting in Semivarying Coefficient Model. *Journal of Multivariate Analysis*, 82(1):166–188.

Zheng, S. (2012). QBoost: Predicting Quantiles with Boosting for Regression and Binary Classification. *Expert Systems with Applications*, 39(2):1687–1697.

Zou, H. (2006). The Adaptive Lasso and its Oracle Properties. *Journal of the American statistical association*, 101(476):1418–1429.