# PRÁCTICAS para el APRENDIZAJE de la ECONOMETRÍA

Pilar González Casimiro
Susan Orbe Mandaluniz

ARGITALPEN ZERBITZUA
SERVICIO EDITORIAL

eman ta zabal zazu
Universidad del País Vasco
Euskal Herriko Unibertsitatea

www.argitalpenak.ehu.es
ISBN: 978-84-9860-605-8

# Prácticas para el Aprendizaje de la ECONOMETRÍA

Pilar González Casimiro
Susan Orbe Mandaluniz

Departamento de Economía Aplicada III
(Econometría y Estadística)

eman ta zabal zazu

Universidad del País Vasco
Euskal Herriko Unibertsitatea



ISBN: 978-84-9860-605-8
LG/D.L.: BI-336-2012

# Contenido

# PRESENTACIÓN

En este libro, las autoras han querido recoger y organizar parte del material docente que han elaborado en los últimos años para la la asignatura de Introducción a la Econometría impartida en tercer curso de la licenciatura en Economía en la Facultad de Ciencias Económicas y Empresariales de la UPV/EHU. Esta labor ha sido consecuencia directa de la participación de las autoras en los programas de innovación docente impulsados desde el Vicerrectorado de Calidad e Innovación Docente de la UPV/EHU con el objetivo de facilitar la adaptación de las asignaturas al crédito ECTS e impulsar la puesta en práctica de nuevas metodologías docentes, como el Aprendizaje Basado en Proyectos o el Aprendizaje Basado en Problemas. En este sentido queremos reconocer especialmente la labor realizada por nuestra profesora y colega, la catedrática de Econometría Inmaculada Gallastegui Zulaica, pionera en este área de innovación docente tanto en el departamento como en la Facultad. Sin su ánimo y su empuje quizás este trabajo no se hubiera materializado.

La implantación de las nuevas metodologías docentes ha traido consigo la aparición de nuevas modalidades docentes: clases magistrales, dedicadas a la transmisión de conocimientos teóricos a grupos grandes, clases prácticas de aula para abordar los aspectos más prácticos de la materia, y seminarios en grupos pequeños para facilitar la comunicación entre los docentes y el alumnado. Algunos de los seminarios se pueden realizar en los laboratorios de informática para adquirir las competencias de la materia referidas a la utilización del software econométrico y, en general, de las tecnologías de la información. Esta nueva forma de organizar la docencia ha planteado la necesidad de generar nuevo material docente para poder trabajar las distintas competencias de la materia en cada una de las modalidades docentes.

El objetivo de esta publicación es proporcionar material docente que sea de utilidad tanto para el estudio autonómo del estudiante como para la preparación por parte del docente de las clases prácticas y los seminarios de una asignatura de introducción a la econometría. De hecho, el material aquí presentado recoge muchas de las propuestas que las autoras han utilizado, en diferentes versiones, para trabajar en las clases prácticas y seminarios, así como para las pruebas finales de evaluación, tanto teóricas como prácticas.

Esta publicación está dirigida, por lo tanto, a estudiantes de asignaturas introductorias de econometría tanto de las licenciaturas, actualmente en proceso de extinción, como de los grados del ámbito de la economía y la empresa, es decir, grado en Economía, en Administración y Dirección de Empresas, en Marketing, en Finanzas y seguros, en Fiscalidad y Administración Pública, ... Dado el carácter eminentemente práctico de este libro, es preciso completementar su utilización con la consulta y estudio de los libros más generales que se incluyen en cualquier bibligrafía de la guía docente de una asignatura de introducción a la econometría.

Hoy en día, la realización de un trabajo econométrico empírico requiere la utilización del software econométrico apropiado. Existen en el mercado muchos paquetes estadístico-econométricos diferentes que cubren de sobra los requisitos de un programa introductorio de econometría. En concreto, las autoras utilizan el programa Gretl porque es un software libre, muy sencillo de utilizar y además existen versiones en muchos idiomas, entre ellos, castellano, euskera e inglés. Se puede descargar de la siguiente página:

http://gretl.sourceforge.net

Además del programa se pueden cargar los datos utilizados como ejemplos de aplicaciones econométricas en muchos libros de texto: Wooldridge (2003), Gujartati (2003), Ramanathan (2002), entre otros.

El libro está dividido en cuatro partes. En la primera se recogen las ideas fundamentales sobre qué es la econometría y los principios básicos de la especificación, estimación e inferencia del Modelo de Regresión Lineal General (MRLG). Estas simples notas no son suficientes para adquirir las competencias correspondientes a un curso de introducción a la econometría. La intención de las autoras es que sirvan, simplemente, para establecer la nomenclatura que se va a utilizar después en el resto del libro.

Cuadro 1: Ejercicios propuestos

| Ejercicio | Fichero de datos |
|---|---|
| E4 | helados.gdt |
| E7 | ordenadores.gdt |
| E8 | consumo.gdt |
| E9 | merluza.gdt |
| E10 | lalimpia.gdt |
| E11 | casas_rurales.gdt |
| E12 | leasing.gdt |
| E13 | nihon.gdt |
| E15 | electricidad.gdt |

La segunda parte presenta una colección de quince ejercicios ordenados de menor a

Cuadro 2: Prácticas de autoevaluación

| Práctica | Fichero | Fichero adaptado |
|---|---|---|
| P1 | practicaP1.gdt | hprice.gdt [6] |
| P2 | practicaP2.gdt | wage2.gdt [6] |
| P3 | practicaP3.gdt | mroz87.gdt [1] |
| P4 | practicaP4.gdt | kielmc.gdt [6] |
| P5 | practicaP5.gdt | table_9.3.gdt [2] |
| P6 | practicaP6.gdt | table_7.9.gdt [2] |
| P7 | practicaP7.gdt | data7_5.gdt [5] |
| P8 | practicaP8.gdt | data7_2.gdt [5] |
| P9 | practicaP9.gdt | kielmc.gdt [6] |
| P10 | practicaP10.gdt | pizza.gdt [3] |
| P11 | practicaP11.gdt | data7_19.gdt [5] |
| P12 | practicaP12.gdt | table_21.1.gdt [2] |
| P13 | practicaP13.gdt | data7_2.gdt [5] |
| P14 | practicaP14.gdt | tuna.gdt [3] |
| P15 | practicaP15.gdt | data9_5.gdt [5] |

mayor dificultad. Así, los ejercicios E1 y E2 están dedicados a la especificación y estimación del Modelo de Regresión Lineal Simple. Los ejercicios E3 a E10 se dedican a la especificación, estimación e inferencia del Modelo de Regresión Lineal General con variables cuantitativas. Algunos de estos ejercicios tratan aspectos específicos del MRLG como el E4, dedicado al tema del cambio de escala, el E5 al planteamiento de hipótesis, el E8 al error de especificación y el E10 a la forma funcional. Los ejercicios E11 a E15, incluyen también variables explicativas cualitativas, considerando también la posibilidad de introducir interacciónes entre variables explicativas cuantitativas y cualitativas.

Como se puede observar, muchos de los ejercicios están divididos en varias partes o plantean trabajar bajo diferentes escenarios. Esta estructura facilita su utilización en clases prácticas y/o seminarios. La idea es dividir al alumnado en grupos de 3-4 estudiantes, y proponer a los distintos grupos que, de forma autónoma, realicen una parte del ejercicio o trabajen uno de los escenarios propuestos. Luego, en el aula, se ponen en común los resultados obtenidos por cada grupo para su contraste y debate. De esta manera, partiendo de estudios parciales se puede desarrollar en común un aspecto de interés del MRLG.

Algunos de los ejercicios están pensados para realizarlos “a mano”, es decir, con calculadora, pero otros vienen acompañados de un fichero de datos en formato Gretl para poder resolverlos. El cuadro 1 recoge los ficheros de datos correspondientes a algunos ejercicios propuestos.

La mayoría de los datos de estos ficheros son artificiales. Otros, como es el caso de los ficheros ordenadores.gdt y leasing.gdt, provienen de bases de datos construidas por estudiantes de tercer curso de la licenciatura en Economía para sus proyectos de la asignatura de Introducción a la Econometría. Por último, el fichero nihon.gdt contiene datos del fichero wagedisc.gdt [4] y el fichero casas_rurales.gdt es una adptación del fichero data4_1.gdt [5] y se pueden encontrar en su versión original en Gretl.

La tercera parte recoge quince prácticas que pueden ser de utilidad para que el alumnado pueda estudiar de forma autónoma, a su propio ritmo, ya que, en la cuarta parte, se incluyen soluciones bastante detalladas de las mismas.

Los ficheros de datos practicaP1 a practica P15 contienen los datos correspondientes a cada una de las prácticas. En la mayoría de los casos no son necesarios, ya que los enunciados de las prácticas contienen toda la información necesaria para su resolución. Esto es debido a que distintas versiones de estas pruebas de autoevaluación han sido objeto de pruebas de examen escrito y práctico de la asignatura Introducción a la Econometría (véase el cuadro 2). Sin embargo, se proporcionan los ficheros de datos utilizados para que los estudiantes puedan reproducir los resultados. Estos ficheros son, salvo el caso del correspondiente a la Práctica 3, adaptaciones de ficheros de libros de Econometría muy utilizados en la docencia de esta materia y de cuyos datos se puede disponer en Gretl.

Los ficheros con los datos correspondientes a los ejercicios y a las prácticas se pueden descargar de las páginas web personales de las autoras.

# Parte I

# Introducción a la Econometría: Modelo de Regresión Lineal General

# Capítulo 1

# ¿Qué es la econometría?

La Econometría se ocupa de formular relaciones entre variables económicas, cuantificarlas y valorar los resultados obtenidos, bien contrastando teorías económicas bien evaluando e implementando políticas económicas públicas o de empresa. La Econometría se dedica además al estudio y desarrollo de los métodos precisos para llevar a cabo estas tareas.

Probablemente las aplicaciones más conocidas de la Econometría, a nivel general, son las predicciones de variables macroeconómicas de interés: PIB, inflación, tipos de interés, etc. Pero la importancia del uso de la Econometría va más allá de la Economía siendo un instrumento utilizado, en la actualidad, no sólo en muchos campos de los negocios (finanzas, marketing, gerencia, etc.) sino por muchos investigadores sociales (historia, ciencias políticas, sociología, criminología, sicología, antropología, agricultura económica, etc.). No hay que olvidar que la Economía es una ciencia social por lo que cualquier avance o método de investigación utilizado por los economistas, entre ellos la Econometría, es útil para otras ciencias sociales.

El siguiente ejemplo puede servir de ayuda para clarificar qué es la Econometría.

> Ejemplo. En épocas de tensiones inflacionistas, el director del Banco Central Europeo (BCE) tiene que decidir qué hacer para enfriar la economía, tratando de rebajar la inflación manteniendo un crecimiento económico estable. La teoría económica explica cuál es el comportamiento de determinadas variables en la economía, de forma que el director del BCE sabe que para conseguir su objetivo ha de subir el tipo de interés. Esta subida de interés afecta a las empresas que buscan capital para su expansión y a los consumidores que desean comprar bienes duraderos. Así, tipos de interés más altos hacen que los consumidores y las empresas se retraigan a la hora de hacer sus inversiones con lo que la demanda agregada disminuye y se desacelera la inflación. Pero este conocimiento no es suficiente para tomar decisiones. El director del BCE necesita saber exactamente ¿cuánto ha de subir el tipo de interés para conseguir su objetivo, un cuartillo, dos? La respuesta dependerá de la capacidad

de respuesta de las empresas e individuos a los aumentos de los tipos de interés y a los efectos de una reducción del PIB. Estas respuestas, elasticidades y/o multiplicadores son, en general, parámetros desconocidos y han de ser estimados a partir de los datos económicos disponibles.

En este ejemplo, se puede observar qué objetivos pretende cubrir la econometría y cómo se relaciona con otras ciencias. En primer lugar, actúa la teoría económica que informa sobre cómo unas variables económicas están relacionadas con otras. En este sentido la Econometría está en contacto con la Teoría Económica que es la que se ocupa de estudiar las relaciones existentes entre las variables económicas. En economía, las teorías sobre relaciones entre variables económicas se expresan formulando modelos económicos expresados como funciones matemáticas. Ahora bien, el siguiente paso es siempre cuantificar, y éste no es un objetivo de la Teoría Económica; sin embargo, *cuantificar las relaciones entre las variables*, sí es uno de los objetivos de la Econometría. Para lograr este objetivo necesitamos disponer de información cuantificada, es decir, de datos. Para realizar un trabajo empírico econométrico es fundamental contar con información factual relativa a las variables cuyas relaciones se trata de estudiar. La econometría trata de la estimación de estos parámetros desconocidos de los modelos (elasticidades, multiplicadores, etc.) en base a los datos económicos. Para ello es necesario contar con métodos de tratamiento de datos que nos permitan cuantificar las relaciones de manera adecuada así como valorar los resultados de acuerdo con criterios reconocidos. Para realizar esta tarea, la Econometría utiliza elementos que toma prestados de la Estadística.

En resumen, para contestar las preguntas que se hace el director del BCE, se puede utilizar la econometría: construyendo modelos que relacionen las variables económicas de interés con ayuda de la teoría económica o del sentido común, recogiendo datos sobre estas variables relevantes y aplicando métodos estadístico-econométricos para estimar los parámetros desconocidos y obtener conclusiones (contrastar teorías, predecir, etc.).

> Econometría: medición de la economía con el propósito de contrastar y desarrollar la teoría económca. (Sloan, 1949)
>
> Econometría: ciencia social que aplica la Teoría Económica, las Matemáticas y la Inferencia Estadística al análisis de los fenómenos económico. (Goldberger, 1964)

# Capítulo 2

# Modelo econométrico

El modelo econométrico es uno de los pilares fundamentales del trabajo econométrico. Ahora bien, ¿qué es un modelo econométrico?, ¿qué relación tiene y en qué se diferencia de un modelo económico?

Supongamos que la dirección de una conocida marca de coches quiere saber si ha tenido éxito una determinada campaña publicitaria que se ha llevado a cabo para relanzar la venta de un determinado modelo. Para poder contestar a esta cuestión habrá que conocer cuál es la relación que existe entre las ventas y los gastos en publicidad.

En primer lugar, se habrá de formular un modelo económico proporcionado por la teoría económica o la lógica económica que recoja los principales factores que influyen sobre las ventas del modelo de coche considerado. Dado que se trata de un modelo de demanda, se considera que las variables explicativas que determinan las ventas son: el precio del coche, $P$ , el precio de los coches de la misma línea de otras marcas, $P^c$ , los gastos en publicidad, $GP$, y las condiciones de mercado (época de bonanza económica o no), $CM$. El modelo económico vendría dado por:

$$V = f(P, P^c, GP, CM) \qquad \frac{dV}{dP} < 0 \quad \frac{dV}{dP^c} > 0 \quad \frac{dV}{dGP} > 0 \quad \frac{dV}{dCM} > 0. \tag{2.1}$$

Este modelo postula que la cantidad de coches vendida, $V$ , es una función inversa del precio del coche, $P$, y una función creciente del precio de los coches de la misma línea de otras marcas, $P^c$, de los gastos en publicidad, $GP$, y de las condiciones de mercado, $CM$.

El modelo propuesto podría servir para representar la relación entre ventas, precios y gastos en publicidad de cualquier empresa, mientras que lo que le interesa a la dirección de una empresa en particular es cómo responden sus ventas a su precio, sus campañas publicitarias, etc. Por lo tanto, el siguiente paso que hay que dar es cuantificar el modelo económico general (2.1) con los datos de la empresa concreta. Para ello es necesario contar con datos tanto para la variable a explicar, las ventas, como para las variables explicativas, precios, precios de la competencia, gastos en publicidad y condiciones de mercado. Supongamos que la dirección de la empresa cuenta con datos semanales sobre sus ventas, precios, precios de la competencia y gastos en publicidad. ¿Qué ocurre con la variable *condiciones de mercado*? En principio, no es numérica, pero para poder cuantificar el modelo hay que hacerla numérica. Una de las soluciones a este problema pasa por basarse en que las condiciones de mercado dependerán de la fase del ciclo económico, favorable o no, y medir el ciclo económico a través del Índice de Producción Industrial. Es decir, se propone utilizar los datos del Índice de Producción Industrial como proxy para medir la variable *condiciones de mercado*.

De momento se cuenta con el modelo económico que formula la relación entre variables económicas, y, como mucho, la dirección de esa relación y, con los datos. Es necesario incorporar más información al modelo económico para poder cuantificar la relación propuesta.

En primer lugar, hay que especificar una forma funcional, es decir, hay que hacer alguna hipótesis sobre cómo es la relación entre la variable a explicar y las variables explicativas. Se suele suponer una relación lineal:

$$V = \beta_0 + \beta_1 P + \beta_2 P^c + \beta_3 GP + \beta_4 CM \tag{2.2}$$

donde $\beta_0, \beta_1, \beta_2, \beta_3$ y $\beta_4$ son los parámetros que miden la relación de las distintas variables explicativas con las ventas. Estos parámetros desconocidos se suponen constantes para toda la muestra de datos utilizada.

Antes de tratar de estimar estos parámetros desconocidos, hay que profundizar en la especificación del modelo: se ha de tener en cuenta que las relaciones económicas no son exactas, no responden a una fórmula matemática como la que expresa el modelo económico (2.2). Los modelos económicos que se proponen no pretenden explicar o predecir el comportamiento de cada individuo, sino describir el comportamiento promedio o sistemático de un conjunto de individuos, empresas, etc. Para recoger este hecho es preciso incluir un término de perturbación para cada observación que ligue el modelo económico con los datos para todas las observaciones.

Por lo tanto, el modelo económico (2.2) no es exacto y se tiene que incluir una parte no predecible, $u$, que recoge todos los factores que afectan a las ventas y que no están incluidos explícitamente en el modelo, además de recoger la incertidumbre intrínseca a toda actividad económica o social. De forma que el modelo econométrico queda como sigue:

$$V_t = \beta_0 + \beta_1 P_t + \beta_2 P_t^c + \beta_3 GP_t + \beta_4 CM_t + u_t. \tag{2.3}$$

La distinción fundamental entre modelo económico (2.2) y modelo econométrico (2.3) es la presencia de esta perturbación. Si el modelo económico ha recogido todas las influencias "importantes y sistemáticas" que existen sobre las ventas, en el nuevo elemento se recogerán todos los efectos no sistemáticos que serán, en general, más erráticos, aleatorios. Se supone, por lo tanto, que $u_t$ es una variable aleatoria cuya distribución de probabilidad habrá que especificar al mismo tiempo que el resto del modelo.

Así, el modelo econométrico consta de dos partes,

$$V = \text{Parte sistemática} + \text{Parte aleatoria}.$$

La parte sistemática es la que establece la relación estable o en promedio entre las variables de interés basado en la teoría económica o en el razonamiento económico, más un supuesto sobre la forma funcional. La parte aleatoria, impredecible, es la perturbación, $u$, que recoge el resto de los efectos que influyen en las ventas y no tiene un comportamiento sistemático.

Utilizando este modelo econométrico, que es un modelo probabilístico, se pueden estimar los parámetros desconocidos en base a los datos disponibles, contrastar hipótesis de interés sobre la relación entre la variable a explicar y las variables explicativas y predecir, utilizando métodos tomados de la estadística.

Una pieza fundamental en el trabajo empírico econométrico son los datos de que se dispone. Vamos a dedicar un breve espacio a la discusión de la naturaleza de los datos económicos.

Cuando se va a realizar una investigación se puede contar con dos tipos de datos:

1. *Datos experimentales:* se recogen normalmente en laboratorios bajo situaciones controladas, siendo el resultado de experimentos controlados. Estos experimentos se caracterizan porque:

   - Las condiciones que rodean el experimento están bajo el control del investigador.
   - Por lo tanto, el experimento puede ser reproducido y repetido por otros investigadores.
   - El experimento puede repetirse tantas veces como se quiera bajo diferentes condiciones o valores de las variables de control.

   Por ejemplo, supongamos que se quiere analizar el efecto del fertilizante en la cantidad de cosecha recogida. Se pueden controlar los factores que afectan la cosecha y experimentar con diferentes cantidades de fertilizante. El resultado del experimento es una variable aleatoria porque no se conoce su valor hasta que no se realiza el experimento. Aún en los experimentos controlados, los resultados serán diferentes si se repite bajo las mismas condiciones. Esto se denomina variación muestral y es debida a múltiples factores: factores que no se controlan, errores de medida, etc. De todos modos la cosecha depende de otros factores que no se controlan como la climatología, con lo que aunque se repita el experimento en las mismas condiciones tendremos diferentes resultados.

2. *Datos no experimentales:* son el resultado de experimentos no controlados.

   Las variables económicas (PIB, tipo de interés, etc.) toman valores que, generalmente, no son el resultado de experimentos controlados sino que simplemente se observan. Los factores que afectan a estas variables no suelen estar bajo el control de un grupo o una persona, y, en la mayoría de las ocasiones no se pueden repetir. De todas formas se considera que son variables aleatorias porque no se conoce su valor hasta que no se observa.

Los datos no experimentales (o datos observacionales) se llaman así para resaltar el hecho de que el investigador es simplemente un recolector pasivo de datos. Son muy habituales en Ciencias Sociales por lo que si se quiere hacer un experimento es imposible, o muy caro, o moralmente incorrecto, llevar a cabo experimentos controlados.

La Econometría trabaja con datos económicos que normalmente son no experimentales. En consecuencia, toma prestado de la estadística matemática todos los métodos que son útiles para su trabajo, pero también ha tenido que diseñar técnicas nuevas para poder trabajar con las complejidades de los datos económicos y contrastar las teorías económicas.

Otra característica de interés de los datos económicos es que normalmente no se recogen para los propósitos de un análisis econométrico concreto sino por cualquier otra razón. Así puede ocurrir que los datos disponibles no se refieran a las variables de interés del investigador, o que en las encuestas no se pregunten las cosas como le interesa exactamente al investigador. De forma que uno de los retos de la investigación económica es obtener datos que sean consistentes con los valores de las variables teóricas del modelo económico y que sean útiles para analizar el problema económico al que el investigador se enfrenta.

Por último, los datos económicos se pueden presentar de diversas formas:

1. *Datos de sección cruzada.* Consisten en una muestra de individuos, familias, empresas, ciudades, etc. tomadas en un momento determinado de tiempo. Por lo tanto, estos datos presentan una característica de interés: en muchos casos se puede suponer que los datos de sección cruzada se obtienen por muestreo aleatorio de una población.

2. *Datos de series temporales.* Son observaciones sobre una o varias variables a lo largo del tiempo. El tiempo es una característica fundamental en los datos de series temporales, porque sabemos que el pasado influye en el futuro, que hay retardos en el comportamiento de los agentes económicos. En este caso, el orden cronológico de los datos trae en sí mismo información.

   Una característica fundamental de los datos de series temporales es que raramente van a ser independientes entre sí a lo largo del tiempo. Las series temporales económicas tienen, en general, una fuerte dependencia de su pasado. Se han desarrollado nuevos métodos econométricos para explotar esta dependencia en las series temporales económicas y para recoger las especiales características de las mismas.

3. *Datos de panel.* Tienen características de sección cruzada y series temporales: se cuenta con una serie temporal para cada individuo del conjunto de datos.

# Capítulo 3

# Etapas de un trabajo empírico

Los métodos econométricos son de relevancia en cualquier campo de la economía aplicada cuando se quiere contrastar una teoría económica o una determinada hipótesis sobre una relación económica de interés para tomar decisiones de política económica, empresarial, etc... El análisis empírico utiliza los datos y los métodos econométricos para contrastar una teoría o estimar una relación. Las etapas en la realización de un trabajo empírico son las siguientes:

1. Especificación del modelo econométrico.
   1.1. Selección de las variables explicativas.
   1.2. Forma funcional.
   1.3. Distribución del término de perturbación.
2. Estimación de los parámetros.
3. Valoración de los resultados obtenidos.
   3.1. Contrastación de las hipótesis de interés.
   3.2. Predección.

En los siguientes secciones se presentan unas nociones básicas de cada una de estas etapas.

## 3.1. Especificación.

Para proceder a la especificación de un modelo econométrico, en primer lugar, hay que formular cuidadosamente el problema de interés, para después construir un modelo económico, o sea un conjunto de ecuaciones matemáticas que describan las relaciones económicas entre la variable a explicar y las variables explicativas. Si la teoría económica no proporciona un modelo para el problema bajo estudio, habrá que utilizar la lógica económica y el sentido común para elegir las variables relevantes en cada caso, formular relaciones entre ellas, etc.

La especificación de un modelo econométrico supone hacer hipótesis sobre la forma o relación funcional existente entre las variables y la función de distribución de la perturbación $u$.

El modelo econométrico que se va a utilizar es el Modelo de Regresión Lineal General:

$$Y_i \,=\, \beta_1 \,+\, \beta_2\, X_{2i} \,+\, \beta_3\, X_{3i} \,+\, \ldots \,+\, \beta_k\, X_{ki} \,+\, u_i \qquad i = 1, \ldots, N \tag{3.1}$$

donde

$Y$ es la variable a explicar, variable dependiente o variable endógena,
$X_j, j = 2, \ldots, k$ son las variables explicativas o variables exógenas,
$\beta = (\beta_1, \beta_2, \beta_3, \ldots, \beta_k)'$ es el vector de coeficientes de la regresión,
$u$ es el término de perturbación aleatoria que sigue una distribuión con media cero y varianza constante y
$N$ es el tamaño muestral.

Este modelo se ha especificado con datos de sección cruzada pero la misma especificación se puede utilizar con datos de series temporales.

Como el modelo econométrico (3.1) se cumple para toda la muestra, se puede escribir en forma matricial como sigue:

$$Y \,=\, X\beta \,+\, u$$

donde:

$$Y = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_N \end{bmatrix} \quad X = \begin{bmatrix} 1 & X_{21} & X_{31} & \ldots & X_{k1} \\ 1 & X_{22} & X_{32} & \ldots & X_{k2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & X_{2N} & X_{3N} & \ldots & X_{kN} \end{bmatrix} \quad \beta = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{bmatrix} \quad u = \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_N \end{bmatrix}$$

La matriz $X$ es de orden $(N \times k)$ y recibe el nombre de matriz de datos.

HIPÓTESIS BÁSICAS del modelo de regresión lineal general:

Dado que se trata de un curso introductorio de Econometría, vamos a trabajar con el siguiente conjunto de hipótesis básicas ideales.

HB1. El modelo (3.1) refleja el comportamiento de la variable dependiente $Y_i$ durante toda la muestra, $i = 1, 2, \ldots, N$. Esta hipótesis implica:

- Correcta selección de las variables explicativas en el sentido de que están incluidas en el modelo todas las variables relevantes para la determinación de la variable endógena. Además, todas las variables incluidas en el modelo son relevantes.
- El modelo es lineal en los coeficientes.
- Los coeficientes del modelo se mantienen constantes durante toda la muestra.

HB2. Las variables explicativas $X_{ji}, j = 2, \ldots, k$ son no estocásticas, es decir, se consideran fijas en muestras repetidas.

HB3. Ausencia de colinealidad perfecta: $rg(X) = k$.

No puede existir una combinación exacta entre las columnas de la matriz de datos $X$. Esto implica que no puede haber una relación lineal exacta entre los datos de las variables explicativas, ni una variable explicativa que sea constante.

HB4. La esperanza matemática de cada una de las perturbaciones es cero:

$$E(u_i) = 0, \ \forall i.$$

HB5. Homocedasticidad: la varianza de todas las perturbaciones es la misma:

$$E(u_i)^2 = \sigma_u^2, \ \forall i.$$

HB6. Ausencia de autocorrelación: no existe correlación entre las perturbaciones de las distintas observaciones:

$$E(u_i\, u_j) = 0, \ \forall i \neq j.$$

HB7. Las perturbaciones siguen distribuciones normales, independientemente distribuidas.

Las hipótesis básicas HB4-HB7 se pueden resumir como sigue: las perturbaciones siguen distribuciones normales, independientemente distribuidas, con media cero y varianza constante:

$$u_i \sim NID(0, \sigma_u^2)$$

## 3.2. Estimación.

### Criterio de estimación MCO.

Si se contara con una estimación del vector de coeficientes $\beta$, se podría estimar para cada observación un valor de la variable dependiente, dados los valores de las variables explicativas:

$$\widehat{Y}_i = \hat{\beta}_1 + \hat{\beta}_2 X_{2i} + \hat{\beta}_3 X_{3i} + \ldots + \hat{\beta}_k X_{ki}.$$

En esta estimación se cometería un error, que se denomina residuo:

$$\hat{u}_i = Y_i - \widehat{Y}_i = Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_{2i} - \hat{\beta}_3 X_{3i} - \ldots - \hat{\beta}_k X_{ki}.$$

Uno de los criterios de estimación de los coeficientes del modelo es el criterio de estimación de Mínimos Cuadrados Ordinarios (MCO) que se basa en estimar el vector $\hat{\beta}$ de forma que minimice la suma de cuadrados de los residuos:

$$\sum_{i=1}^{N} \hat{u}_i^2 = \sum_{i=1}^{N} (Y_i - \widehat{Y}_i)^2 = \sum_{i=1}^{N} (Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_{2i} - \hat{\beta}_3 X_{3i} - \ldots - \hat{\beta}_k X_{ki})^2.$$

De las condiciones de primer orden de este problema de minimización se obtienen las ecuaciones normales:

$$\begin{aligned}
\sum_{i=1}^{N} Y_i &= N\hat{\beta}_1 + \hat{\beta}_2 \sum_{i=1}^{N} X_{2i} + \hat{\beta}_2 \sum_{i=1}^{N} X_{3i} + \ldots + \hat{\beta}_k \sum_{i=1}^{N} X_{ki} \\
\sum_{i=1}^{N} Y_i X_{2i} &= \hat{\beta}_1 \sum_{i=1}^{N} X_{2i} + \hat{\beta}_2 \sum_{i=1}^{N} X_{2i}^2 + \hat{\beta}_2 \sum_{i=1}^{N} X_{3i} X_{2i} + \ldots + \hat{\beta}_k \sum_{i=1}^{N} X_{ki} X_{2i} \\
\sum_{i=1}^{N} Y_i X_{3i} &= \hat{\beta}_1 \sum_{i=1}^{N} X_{3i} + \hat{\beta}_2 \sum_{i=1}^{N} X_{2i} X_{3i} + \hat{\beta}_2 \sum_{i=1}^{N} X_{3i}^2 + \ldots + \hat{\beta}_k \sum_{i=1}^{N} X_{ki} X_{3i} \\
\ldots & \quad \ldots \\
\sum_{i=1}^{N} Y_i X_{ki} &= \hat{\beta}_1 \sum_{i=1}^{N} X_{ki} + \hat{\beta}_2 \sum_{i=1}^{N} X_{2i} X_{ki} + \hat{\beta}_2 \sum_{i=1}^{N} X_{3i} X_{ki} + \ldots + \hat{\beta}_k \sum_{i=1}^{N} X_{ki}^2
\end{aligned}$$

Las ecuaciones normales conforman un sistema de $k$ ecuaciones con $k$ incógnitas. Como no existe ninguna combinación lineal exacta entre las variables explicativas, existe una solución única a este sistema de ecuaciones que es el estimador MCO de los coeficientes del modelo:

$$\hat{\beta}_{MCO} = (\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3, \ldots, \hat{\beta}_k)'_{MCO}$$

Este sistema se puede escribir fácilmente en términos matriciales. Así, las ecuaciones normales toman la forma:

$$X'Y = (X'X)\,\hat{\beta}_{MCO}$$

Como $rg(X) = k$, la solución del sistema es:

$$\hat{\beta}_{MCO} = (X'X)^{-1}\,X'Y$$

Una vez estimados los coeficientes, la varianza de las perturbaciones se puede estimar mediante la expresión:

$$\hat{\sigma}_u^2 = \frac{\sum \hat{u}_i^2}{N-k} = \frac{\hat{u}'\hat{u}}{N-k}$$

## Propiedades de los estimadores MCO.

Los estimadores MCO del vector $\beta$ tienen las siguientes propiedades si se cumplen las hipótesis básicas HB1-HB6, como demuestra el teorema de Gauss-Markov.

- Lineales.
- Insesgados.
- De varianza mínima dentro de la clase estimadores, lineales e insesgados. Esta varianza viene dada por:

  $$V(\hat{\beta}) = \sigma_u^2\,(X'X)^{-1}$$

  Para estimar esta varianza basta con sustituir la varianza de las perturbaciones por su estimador.

Además se puede demostrar que, si se cumplen las hipótesis básicas HB1-HB6, el estimador de la varianza de las perturbaciones propuesto es insesgado.

## Recta o función de regresión muestral.

El modelo estimado es:

$$\widehat{Y}_i = \hat{\beta}_1 + \hat{\beta}_2\,X_{2i} + \hat{\beta}_3\,X_{3i} + \ldots + \hat{\beta}_k\,X_{ki} \qquad i = 1, 2, \ldots, N.$$

Del sistema de ecuaciones normales, se obtienen, entre otras, las siguientes propiedades de la estimación MCO:

- La suma de los residuos MCO es cero: $\sum \hat{u}_i = 0$.
- Los residuos MCO son ortogonales a las variables explictivas: $\sum X_{ji}\,\hat{u}_i = 0,\ j = 2, \ldots, k$.

- Se cumple la siguiente descomposición de la variabilidad:

$$\sum_{i=1}^{N}(Y_i - \bar{Y})^2 = \sum_{i=1}^{N}(\widehat{Y}_i - \bar{Y})^2 + \sum_{i=1}^{N}\hat{u}_i^2 \tag{3.2}$$

  donde:

  SCT $= \sum_{i=1}^{N}(Y_i - \bar{Y})^2$: variabilidad total de la variable dependiente en la muestra.

  SCE $= \sum_{i=1}^{N}(\widehat{Y}_i - \bar{Y})^2$: variabilidad total de la variable dependiente explicada por el modelo de regresión lineal.

  SCR $= \sum_{i=1}^{N}\hat{u}_i^2$: suma de cuadrados de los residuos o variabilidad de la variable dependiente no explicada en la regresión.

### Medida de bondad de ajuste.

El coeficiente de determinación, definido como sigue, proporciona una medida de la capacidad explicativa del modelo, es decir, de la bondad de ajuste:

$$R^2 = \frac{\sum_{i=1}^{N}(\widehat{Y}_i - \bar{Y})^2}{\sum_{i=1}^{N}(Y_i - \bar{Y})^2} = 1 - \frac{\sum_{i=1}^{N}\hat{u}_i^2}{\sum_{i=1}^{N}(Y_i - \bar{Y})^2}.$$

Dada la descomposición de la varianza (3.2), el coeficiente de determinación, $R^2$, está comprendido entre 0 y 1. Este coeficiente mide la proporción de la variabilidad total de la variable endógena, $Y$, en la muestra que viene explicada por la variabilidad conjunta de las variables explicativas.

## 3.3. Valoración de los resultados.

Se han de llevar a cabo diagnósticos para comprobar si las hipótesis que hemos hecho concuerdan con los datos, si el modelo planteado es correcto para representar el problema económico de interés: ¿hemos elegido bien la forma funcional? ¿hemos introducido en el modelo todas las variables relevantes? Para ello se utilizan técnicas de inferencia estadística.

Si el resultado es positivo, el modelo econométrico es útil para analizar el problema económico y tomar decisiones al respecto. Si el resultado es negativo, se ha cometido algún error en el diseño del modelo y habrá que corregirlo.

## Contrastes de hipótesis de interés.

Consideremos el siguiente modelo de regresión lineal general:

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + \ldots + \beta_k X_{ki} + u_i \quad u_i \sim NID(0, \sigma_u^2) \quad i = 1, \ldots, N. \quad (3.3)$$

### A. Contraste sobre un sólo coeficiente.

a. Planteamiento del contraste:

$$\begin{array}{l} H_0 : \beta_i = \beta_{i0} \\ H_a : \beta_i \neq \beta_{i0} \end{array}$$

b. Estadístico de contraste. Se puede demostrar que:

$$t = \frac{\hat{\beta}_i - \beta_{i0}}{\hat{\sigma}_{\hat{\beta}_i}} \overset{H_0}{\sim} t(N-k)$$

donde la desviación típica estimada del estimador $\hat{\beta}_i$ viene dada por la expresión $\hat{\sigma}_{\hat{\beta}_i} = \hat{\sigma}_u \, a_{ii}$, siendo $a_{ii}$ el elemento i-ésimo de la diagonal de la matriz $(X'X)^{-1}$.

c. Regla de decisión. Se rechaza la hipótesis nula con un nivel de significación del $\alpha \,\%$ si:

$$|t| = \left| \frac{\hat{\beta}_i - \beta_{i0}}{\hat{\sigma}_{\hat{\beta}_i}} \right| > t(N-K)_{\alpha/2}$$

donde $t(N-K)_{\alpha/2}$ es la ordenada de la distribución $t$ de Student que deja a la derecha una probabilidad del $\alpha_2/2\,\%$.

### B. Contraste sobre un subconjunto de coeficientes.

Supongamos que la hipótesis que se desea contrastar supone varias restricciones sobre el conjunto de parámetros. Por ejemplo:

$$\beta_2 = \beta_3 \qquad \text{y} \qquad \beta_{k-1} = \beta_k \qquad (3.4)$$

a. Planteamiento del contraste:

$$\begin{array}{llll} H_0 : & \beta_2 = \beta_3 & \text{y} & \beta_{k-1} = \beta_k \\ H_a : & \beta_2 \neq \beta_3 & \text{y/o} & \beta_{k-1} \neq \beta_k \end{array}$$

b. Estadístico de contraste. Se puede demostrar que:

$$F = \frac{(SSR_R - SSR_{NR})/q}{SSR_N R/(N-k)} \overset{H_0}{\sim} \mathcal{F}(N-k) \qquad (3.5)$$

donde:

$q$: número de restricciones, en el ejemplo, 2.

$SSR_{NR}$ : suma de cuadrados de residuos del modelo no restringido, el modelo (3.3).

$SSR_R$ : suma de cuadrados de residuos del modelo restringido, que resulta de incluir la restricción (3.4) en el modelo (3.3):

$$\begin{aligned} Y_i &= \beta_1 + \beta_2 X_{2i} + \beta_2 X_{3i} + \ldots + \beta_{k-1} X_{(k-1)i} + \beta_{k-1} X_{ki} + u_i \\ Y_i &= \beta_1 + \beta_2 (X_{2i} + X_{3i}) + \ldots + \beta_{k-1} (X_{(k-1)i} + X_{ki}) + u_i \end{aligned}$$

c. Regla de decisión. Se rechaza la hipótesis nula con un nivel de significación del $\alpha\,\%$ si:

$$F > \mathcal{F}(N-k)_\alpha$$

donde $\mathcal{F}(N-k)_\alpha$ es la ordenada de la distribución $\mathcal{F}$ de Snedecor que deja a la derecha una probabilidad del $\alpha\,\%$.

C. Contraste de significación conjunta del modelo de regresión.

a. Planteamiento del contraste:

$$\begin{array}{l} H_0 : \beta_2 = \beta_3 = \ldots = \beta_k = 0 \\ H_a : \beta_2 \neq 0 \text{ y/o } \beta_3 \neq 0 \ldots \text{ y/o } \beta_k \neq 0 \end{array}$$

b. Estadístico de contraste. El estadístico de contraste (3.5) toma la forma siguiente:

$$F = \frac{R^2/(k-1)}{(1-R^2)/(N-k} \overset{H_0}{\sim} \mathcal{F}(N-k)$$

donde $R^2$ es el coeficiente de determinación del modelo (3.3) y $k-1$ es el número de variables explicativas, es decir, de parámetros involucrados en la hipótesis nula.

c. Regla de decisión. Se rechaza la hipótesis nula con un nivel de significación del $\alpha\,\%$ si:

$$F > \mathcal{F}(N-k)_\alpha$$

donde $\mathcal{F}(N-k)_\alpha$ es la ordenada de la distribución $\mathcal{F}$ de Snedecor que deja a la derecha una probabilidad del $\alpha\,\%$.

## Predicción.

Supongamos que se quiere predecir el valor de la variable endógena para una observación $p$ que no corresponde a la muestra, $Y_p$.

Un supuesto crucial que hay que realizar si se quiere utilizar el modelo de regresión lineal para predecir valores de la variable dependiente $Y$ que no pertenecen a la muestra es que la relación entre la variable dependiente $Y$ y las variables explicativas $X_j$ se cumpla también fuera de la muestra. Entonces, dados los valores de las variables explicativas para la observación $p$, $X_{2p}, \ldots, X_{kp}$, se tiene que:

**Predicción por punto.** Viene dada por la función de regresión muestral:

$$\widehat{Y}_p = \hat{\beta}_1 + \hat{\beta}_2 X_{2p} + \hat{\beta}_3 X_{3p} + \ldots + \hat{\beta}_k X_{kp} = X'_p\hat{\beta}.$$

**Predicción por intervalo.** Se obtiene a partir de la distribución del error de predicción:

$$\hat{u}_p = Y_p - \widehat{Y}_p = X'_p\beta + u_p - X'_p\hat{\beta} = X'_p(\beta - \hat{\beta}).$$

Se puede demostrar que:

$$\hat{u}_p \sim N(0, \sigma_p^2) \qquad \text{donde} \qquad \sigma_p^2 = \sigma_u^2 \left(1 + X'_p (X'X)^{-1} X_p\right).$$

El intervalo de confianza $(1-\alpha)$ para $Y_p$ viene dado por:

$$IC(Y_p)_{(1-\alpha)\%} = \left[\widehat{Y}_p \pm t(N-k)_{\alpha/2} \;\hat{\sigma}_u \sqrt{1 + X'_p (X'X)^{-1} X_p}\;\right].$$

Nótese que este intervalo es aleatorio ya que depende de los estimadores de $\beta$ y $\sigma_u^2$ que son variables aleatorias. Su interpretación es la siguiente: si se dispusiera de 100 muestras y se calcularan 100 intervalos de confianza diferentes, $(1-\alpha)\,\%$ de ellos contendrían el verdadero valor $Y_p$.

# Parte II

# Ejercicios Propuestos

## Ejercicio E1. Modelo de Regresión Lineal Simple (MRLS).

Un analista inmobiliario desea estudiar a qué fue debida la fuerte subida experimentada por los precios de las viviendas a principios de siglo antes de la crisis financiera del año 2008. Cree que la evolución del precio de la vivienda puede depender en gran medida del tipo de interés real, por lo que plantea el siguiente modelo de regresión lineal simple:

$$P_t = \beta_1 + \beta_2 I_t + u_t \qquad t = 2002, 2003, \ldots, 2007 \tag{E1.1}$$

donde $P$ es el precio del metro cuadrado de vivienda libre medido en cientos de euros e $I$ es el tipo de interés real en porcentaje. Los datos para ambas variables desde el año 2002 hasta el 2007 están recogidos en la siguiente tabla:

| Año | 2002 | 2003 | 2004 | 2005 | 2006 | 2007 |
|---|---|---|---|---|---|---|
| $P_t$ | 12,6 | 13,8 | 16,2 | 18,2 | 19,9 | 20,8 |
| $I_t$ | 4,3 | 4,3 | 3,8 | 4,0 | 4,5 | 5,0 |

1. Calcula los siguientes estadísticos muestrales: media y varianza de cada una de las variables y coeficiente de correlación entre el precio y el tipo de interés. Comenta los resultados.

2. Dibuja la nube de puntos $(P_t, I_t)$ y el punto de medias $(\overline{P}, \overline{I})$.

3. En el modelo de regresión lineal (E1.1), ¿cuál es la variable endógena?, ¿cuál es la variable exógena?, ¿qué elementos son aleatorios?

4. Escribe la recta de regresión poblacional e interpreta sus coeficientes.

5. Según el criterio Mínimo Cuadrático Ordinario, ¿cuál es la función objetivo que hay que minimizar para estimar los coeficientes del modelo (E1.1)?

6. Deriva el sistema de ecuaciones normales del problema de optimización anterior y sustituye los valores muestrales.

7. Estima por Mínimos Cuadrados Ordinarios (MCO) los coeficientes del modelo (E1.1).

8. Escribe la recta de regresión muestral e interpreta los coeficientes estimados.

9. En el caso de que se conociera la verdadera recta de regresión poblacional, ¿debería coincidir con la recta de regresión muestral?

10. Dibuja en el mismo gráfico del apartado 2 la recta de regresión ajustada. ¿Pasa por el punto de medias $(\overline{P}, \overline{I})$? ¿Por qué?

11. Calcula los residuos MCO y represéntalos gráficamente.

12. Calcula el coeficiente de determinación e interpreta el resultado obtenido.

13. Estima la varianza de las perturbaciones del modelo.

14. Estima la varianza del estimador $\hat{\beta}_2$. Si se conociera la verdadera varianza de $\hat{\beta}_2$, ¿debería coincidir con la estimada?

15. Comprueba que en el modelo de regresión lineal simple (E1.1) se cumple que el coeficiente de determinación es igual al cuadrado del coeficiente de correlación entre el precio y el tipo de interés.

16. Estima por MCO el siguiente modelo de regresión lineal simple:

$$I_t = \gamma_1 + \gamma_2 \, P_t + u_t \qquad t = 2002, \ldots, 2007. \tag{E1.2}$$

Interpreta los coeficentes estimados y el coeficiente de determinación de este modelo y compáralos con los obtenidos para el modelo (E1.1). ¿Son iguales? ¿Por qué?

## Ejercicio E2. MRLS bajo distintos escenarios.

Se desea analizar el comportamiento de las ventas de perfume en función de los gastos en publicidad. Para ello se dispone de datos para cinco empresas del sector sobre las variables: ventas ($V$, en millones de cajas) y gastos en publicidad ($GP$, en miles de euros).

| Empresa | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| $V_i$ | 4 | 7 | 3 | 9 | 17 |
| $GP_i$ | 2 | 3 | 1 | 5 | 9 |

Sea el siguiente modelo de regresión lineal simple:

$$V_i = \beta_1 + \beta_2 \, GP_i + u_i \qquad i = 1, 2, \ldots, 5. \tag{E2.1}$$

Considera los siguientes supuestos sobre los coeficientes del modelo (E2.1):

**Escenario 1.** $\beta_1 \neq 0$ $\beta_2 \neq 0$
**Escenario 2.** $\beta_1 = 0$ $\beta_2 \neq 0$
**Escenario 3.** $\beta_1 \neq 0$ $\beta_2 = 0$

Contesta a las siguientes preguntas para cada uno de estos escenarios:

1. ¿Qué significan en este escenario los supuestos sobre los coeficientes del modelo (E2.1)?

2. Escribe el modelo resultante de imponer estos supuestos sobre los coeficientes.

3. Estima por MCO los coeficientes del modelo derivado en el apartado anterior.
4. Escribe la recta de regresión muestral e interpreta los coeficientes estimados.
5. Dibuja la recta de regresión ajustada. ¿Pasa por el punto de medias $(\overline{V}, \overline{GP})$?
6. Comprueba si se cumple que:

$$\sum_{i=1}^{5} \hat{u}_i = 0 \qquad \text{y} \qquad \sum_{i=1}^{5} \hat{u}_i \, GP_i = 0.$$

   Comenta el resultado obtenido.

7. Comprueba si se cumple la descomposición de la suma de cuadrados:

$$SCT = SCE + SCR$$

   donde $SCT$ es la suma de cuadrados total, $SCE$ es la suma de cuadrados explicada por la regresión y $SCR$ es la suma de cuadrados de los residuos. Comenta el resultado obtenido.

8. Calcula el coeficiente de determinación e interpreta la bondad de ajuste obtenida.

Escribe un informe que recoja las conclusiones que obtienes al comparar los resultados obtenidos en los tres escenarios propuestos.

## **Ejercicio E3.** Modelo de Regresión Lineal General (MRLG).

Se desea explicar el comportamiento de las ventas de automóviles en función del precio y de los gastos en publicidad de las empresas. Para ello se dispone de datos para seis empresas sobre las variables: ventas ($V$, en millones de euros), precios de los automóviles ($P$, en miles de euros) y gastos en publicidad ($G$, en decenas de miles de euros).

| Empresa | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| $V_i$ | 10 | 8 | 7 | 6 | 13 | 6 |
| $P_i$ | 10 | 11 | 13 | 12 | 7 | 9 |
| $G_i$ | 12 | 10 | 9 | 9 | 12 | 11 |

### **PRIMERA PARTE.** Modelo de Regresión Lineal Simple.

Se propone, en primer lugar, el siguiente modelo de regresión lineal simple:

$$V_i = \gamma_1 + \gamma_2 \, P_i + u_i \qquad i = 1, 2, \ldots, 6. \tag{E3.1}$$

1. Según el criterio Mínimo Cuadrático Ordinario, ¿cuál es la función objetivo que hay que minimizar para estimar los coeficientes del modelo (E3.1)?
2. Obtén el sistema de ecuaciones normales correspondiente y sustituye los valores muestrales.
3. Escribe las matrices $X$, $X'X$ y $X'Y$ con los datos disponibles.
4. Estima los coeficientes del modelo (E3.1) por MCO. Escribe la recta de regresión muestral e interpreta los coeficientes estimados.
5. Estima la varianza de las perturbaciones del modelo $\sigma_u^2$.
6. Estima la matriz de varianzas y covarianzas de los estimadores MCO.
7. Calcula el coeficiente de determinación e interprétalo.

**SEGUNDA PARTE.** Modelo de Regresión Lineal General. Estimación.

Se propone ahora el siguiente modelo de regresión lineal general:

$$V_i = \beta_1 + \beta_2 P_i + \beta_3 G_i + v_i \qquad i = 1, 2, \ldots, 6. \tag{E3.2}$$

1. ¿Cuál es la función objetivo que hay que minimizar para estimar por MCO los coeficientes del modelo (E3.2)?
2. Obtén el sistema de ecuaciones normales correspondiente y sustituye los valores muestrales.
3. Escribe las matrices $X$, $X'X$ y $X'Y$ con los datos disponibles.
4. Estima por MCO los coeficientes del modelo (E3.2). Escribe la recta de regresión muestral e interpreta los coeficientes estimados.
5. Considera la primera empresa de la muestra, ¿cuáles son las ventas estimadas según este modelo?, ¿cuál es el residuo correspondiente?
6. ¿Coinciden los valores estimados para $\gamma_1$ y $\beta_1$? ¿Y los valores estimados para $\gamma_2$ y $\beta_2$?, ¿por qué? Ilustra tus conclusiones basándote en las ecuaciones normales.
7. Calcula la suma de cuadrados de los residuos y compárala con la obtenida para el modelo (E3.1). ¿Era de esperar el resultado obtenido?
8. Estima la varianza de las perturbaciones $\sigma_v^2$. ¿Coinciden las estimaciones $\hat{\sigma}_u^2$ y $\hat{\sigma}_v^2$? ¿Por qué?

9. Estima la matriz de varianzas y covarianzas de los estimadores MCO del modelo (E3.2).

10. Calcula el coeficiente de determinación y compáralo con el obtenido para el modelo (E3.1). ¿Era de esperar el resultado obtenido?

11. Considera los modelos (E3.1) y (E3.2). Con la información con la que cuentas hasta ahora, ¿puedes decidir qué modelo es el mejor de los dos? ¿Por qué?

## Ejercicio E4. MRLG: cambio de escala.

Una empresa tiene la concesión de los puestos de helados de la playa de un pueblo turístico. Para determinar los factores que influyen en las ventas de helado en los meses de verano, cuenta con observaciones para los meses de julio, agosto y septiembre de los últimos 6 años sobre las siguientes variables: ventas de helados ($V$), temperatura ($T$) y precio medio del helado ($P$).

Se especifica el siguiente modelo de regresión lineal para la determinación de las ventas:

$$V_t = \beta_1 + \beta_2 T_t + \beta_3 P_t + u_t \qquad t = 1, 2, \ldots, 18. \tag{E4.1}$$

**PRIMERA PARTE.** Estimación.

1. ¿En qué unidades está medida la variable endógena del modelo (E4.1)? ¿Y las variables exógenas?

2. Estima por Mínimos Cuadrados Ordinarios los coeficientes del modelo (E4.1).

3. Escribe la recta de regresión muestral e interpreta los coeficientes estimados.

4. ¿Cuál es la suma de cuadrados de los residuos?

5. Interpreta la bondad de ajuste obtenida.

6. Estima la varianza de las perturbaciones.

7. Estima la matriz de varianzas y covarianzas de los estimadores MCO.

**SEGUNDA PARTE.** ¿Qué ocurre si cambian las unidades de medida de las variables?

Considera los siguientes escenarios:

**Escenario 1**. Las ventas se miden en miles de euros y el resto de las variables permanecen en las unidades de medida originales.

**Escenario 2**. El precio se mide en miles de euros y el resto de las variables permanecen en las unidades de medida originales.

**Escenario 3**. Las ventas y el precio se miden en miles de euros y la temperatura se mide en grados Farenheit.

**Escenario 4**. La temperatura se mide en grados Celsius y las ventas y el precio se miden en euros.

Para cada uno de estos escenarios contesta a las siguientes preguntas:

1. Calcula la/s transformación/es de la/s variable/s que sean oportunas.
2. Estima con los nuevos datos los coeficientes del modelo propuesto.
3. Interpreta los coeficientes estimados.
4. ¿Cuál es la suma de cuadrados de los residuos?
5. Interpreta la bondad de ajuste obtenida.
6. Estima la varianza de las perturbaciones.
7. Estima la matriz de varianzas y covarianzas de los estimadores MCO.
8. Comenta las diferencias que observas en los resultados respecto a los obtenidos en la Primera Parte del ejercicio.

## Ejercicio E5. MRLG: planteamiento de las hipótesis.

Una cadena local de supermercados vende tres marcas de zumo envasado: A, B y C. La empresa productora del zumo A ha llevado a cabo una importante campaña de publicidad tanto en los propios establecimientos de la cadena, como en la televisión y en el periódico locales.

Consideremos el siguiente modelo de regresión lineal para la determinación de las ventas:

$$\begin{aligned} V_t^A &= \beta_1 + \beta_2 P_t^A + \beta_3 P_t^B + \beta_4 P_t^C + \beta_5 GP_t^e + \beta_6 GP_t^{tv} + \beta_7 GP_t^p + \\ &+ \beta_8 GP_t^B + \beta_9 GP_t^C + u_t \qquad\qquad t = 1, 2, \ldots, 40 \end{aligned}$$

donde:

$V^A$ : ventas del zumo A en los establecimientos de la cadena en miles de euros.

$P^A$ : precio medio en euros del zumo A en los establecimientos de la cadena.

$P^B$ : precio medio en euros del zumo B en los establecimientos de la cadena.

$P^C$ : precio medio en euros del zumo C en los establecimientos de la cadena.

$GP^e$ : gasto en publicidad del zumo A en los establecimientos de la cadena en miles de euros.

$GP^{tv}$ : gasto en publicidad del zumo A en TV en miles de euros.

$GP^p$ : gasto en publicidad del zumo A en el periódico local en miles de euros.

$GP^B$ : gasto total en publicidad del zumo B en los distintos medios en miles de euros.

$GP^C$ : gasto total en publicidad del zumo C en los distintos medios en miles de euros.

Se dispone de datos para los últimos cuarenta meses sobre las variables incluidas en el modelo de regresión. El objetivo del productor del zumo A es estimar el modelo propuesto para poder contrastar las siguientes hipótesis de trabajo:

1. El precio de la propia marca influye en sus ventas.

2. Los gastos en publicidad de la marca A influyen en sus ventas.

3. Todas las variables explicativas incluidas en el modelo son conjuntamente significativas.

4. Variaciones en los precios de las marcas competidoras no afectan a la ventas del zumo A.

5. Si se incrementa en dos mil euros los gastos en publicidad en la televisión local, el efecto sobre las ventas es el mismo que si se incrementa en mil euros los gastos en publicidad en el periódico local.

6. El efecto de aumentar mil euros los gastos en publicidad en los establecimientos es igual al efecto de aumentar en la misma cantidad los gastos en publicidad en la televisión local y al doble de aumentar en mil euros los gastos en publicidad en el periódico local.

7. El efecto de incrementar 1 euro el precio medio de la marca del zumo B es igual pero de signo contrario al efecto de aumentar 1 euro el precio de la propia marca A y, además, el efecto de incrementar 1 euro el precio medio de la marca C es la mitad del de aumentar en la misma cantidad el precio medio de la marca B.

8. El efecto sobre las ventas de aumentar los gastos en publicidad de la marca B en tres mil euros es el mismo que el de incrementar en la misma cuantía los gastos en publicidad de la publicidad de la marca C.

9. El efecto de aumentar mil euros los gastos de publicidad tanto de la marca B como de la marca C se puede compensar con un incremento de mil euros en los gastos en publicidad de la marca A en los establecimientos.

Para cada una de las hipótesis planteadas, escribe la hipótesis nula y la alternativa del contraste, así como el modelo restringido correspondiente a cada caso.

## Ejercicio E6. MRLG: contrastes de hipótesis.

Se propone el siguiente modelo de regresión lineal general para analizar el grado de ocupación hotelera ($OC$, en porcentaje) en función de la calidad de los servicios prestados ($C$, índice sobre 100), el precio medio por habitación ($P$, en euros) y la puntuación obtenida por el hotel en la Guía de Hoteles del año correspondiente ($GH$, puntos de 1 a 100):

$$OC_i = \beta_1 + \beta_2 C_i + \beta_3 P_i + \beta_4 GH_i + u_i \qquad i = 1, 2, \ldots, N. \tag{E6.1}$$

Los resultados de la estimación por MCO de los coeficientes de este modelo con los datos de una muestra de 20 hoteles han sido los siguientes:

$$\widehat{OC}_i = 66431{,}018 - 100{,}143\, C_i - 3649{,}113\, P_i + 0{,}8882\, GH_i \qquad R^2 = 0{,}9876.$$

Contrasta la significatividad individual de cada una de las variables explicativas del modelo así como su significatividad conjunta. Interpreta los resultados obtenidos.

Para ayudarte en esta tarea, cuentas además con la siguiente información:

$$\widehat{OC}_i = 98116{,}978 + 4038{,}84\, C_i + 10798{,}901\, P_i \qquad R^2 = 0{,}7335. \tag{E6.2}$$

$$\widehat{OC}_i = 60071{,}404 + 192{,}823\, C_i + 0{,}8742\, GH_i \qquad R^2 = 0{,}9776. \tag{E6.3}$$

$$\widehat{OC}_i = 67310{,}275 - 3503{,}14\, P_i + 0{,}8717\, GH_i \qquad R^2 = 0{,}9874. \tag{E6.4}$$

## Ejercicio E7. MRLG: estimación y contraste de hipótesis.

Se desea determinar el precio de los ordenadores portátiles en función de sus características. Para ello, se toma una muestra de 94 ordenadores sobre las siguientes variables: precio (en euros), potencia de la CPU (en megaherzios), resolución horizontal de la pantalla (en puntos), resolucion vertical de la pantalla (en puntos), peso (en kilogramos) y memoria ram (en gigaherzios).

**PRIMERA PARTE.** Estimación.

1. Especifica un modelo de regresión lineal general para determinar los precios de los ordenadores en función de las características señaladas.

2. Interpreta los coeficientes del modelo propuesto.

3. Estima por MCO el modelo propuesto y escribe la función de regresión muestral.

4. ¿Qué se puede decir sobre la bondad de ajuste del modelo?

**SEGUNDA PARTE.** Contraste de hipótesis.

1. Contrasta la significatividad conjunta de las variables explicativas.

2. Construye intervalos de confianza para los coeficientes que acompañan a las variables peso y memoria ram. Interpreta los resultados obtenidos.

3. Contrasta la significatividad individual de las variables explicativas peso y memoria ram. Compara estos resultados con los obtenidos en el apartado anterior.

4. ¿Existe evidencia en la muestra de que la potencia de la CPU influya positivamente en el precio?

5. Contrasta individual y conjuntamente si la resolución de la pantalla, tanto horizontal como vertical, influye en el precio. Comenta los resultados obtenidos.

6. ¿Existe evidencia en la muestra a favor de la hipótesis de que el efecto de la resolución horizontal de la pantalla es igual al efecto de la resolución vertical pero de signo contrario?

7. Contrasta si el efecto sobre el precio del ordenador de aumentar 1 gigaherzio la memoria ram es el mismo que el efecto de disminuir el peso del ordenador en medio kilogramo.

8. ¿Se está dispuesto a pagar al menos 10 euros más por cada 100 gramos menos que pese el ordenador pórtatil?

## Ejercicio E8. MRLG: error de especificación.

Se dispone de datos para 30 pueblos sobre las variables: consumo familiar ($C$, en miles de euros), riqueza media familiar ($R$, en miles de euros) y nivel de precios ($P$, en euros).

| Obs. | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| $R_i$ | 12 | 17 | 15 | 19 | 18 | 13 | 10 | 16 | 20 | 22 |
| $P_i$ | 150 | 143 | 127 | 169 | 138 | 152 | 129 | 130 | 147 | 153 |
| $C_i$ | 72,4 | 90,2 | 104,6 | 54,4 | 89,9 | 71,0 | 98,9 | 102,2 | 82,3 | 75,5 |
| **Obs.** | **11** | **12** | **13** | **14** | **15** | **16** | **17** | **18** | **19** | **20** |
| $R_i$ | 21 | 23 | 11 | 12 | 20 | 19 | 17 | 15 | 15 | 18 |
| $P_i$ | 131 | 156 | 144 | 149 | 140 | 142 | 146 | 158 | 154 | 156 |
| $C_i$ | 101,2 | 73,2 | 82,7 | 77,5 | 90,0 | 82,6 | 85,7 | 67,1 | 72,0 | 70,1 |
| **Obs.** | **21** | **22** | **23** | **24** | **25** | **26** | **27** | **28** | **29** | **30** |
| $R_i$ | 18 | 16 | 24 | 26 | 23 | 12 | 14 | 14 | 16 | 17 |
| $P_i$ | 156 | 151 | 162 | 149 | 147 | 155 | 142 | 145 | 152 | 147 |
| $C_i$ | 70,1 | 76,8 | 65,3 | 86,0 | 85,9 | 70,6 | 89,5 | 83,5 | 73,2 | 78,9 |

1. Estima por MCO los parámetros del siguiente modelo de regresión lineal simple y contrasta la significatividad de la variable explicativa:
$$C_i = \delta_1 + \delta_2 R_i + u_i \qquad i = 1, 2, \ldots, 30. \tag{E8.1}$$
2. Estima por MCO los parámetros del siguiente modelo de regresión lineal general:
$$C_i = \beta_1 + \beta_2 R_i + \beta_3 P_i + v_i \qquad i = 1, 2, \ldots, 30. \tag{E8.2}$$
3. ¿Por qué no coinciden los valores estimados de $\delta_2$ y $\beta_2$ en los modelos (E8.1) y (E8.2)?
4. En el modelo (E8.2):
   4.1. ¿Son individualmente significativas las variables explicativas? ¿Por qué?
   4.2. Contrasta la siguiente hipótesis
$$H_0: \beta_2 = \beta_3 = 0$$
$$H_a: \beta_2 \neq 0 \;\; y/o \;\; \beta_3 \neq 0.$$
   4.3. Interpreta los resultados obtenidos.
5. Si tuvieras que elegir entre los dos modelos estimados en este ejercicio, ¿con cuál te quedarías?, ¿por qué? Comenta las propiedades de los estimadores del modelo que has elegido y del que no has elegido.

## Ejercicio E9. MRLG: contraste de hipótesis y predicción.

Un investigador desea analizar la cantidad demandada de merluza en función de su propio precio y del precio del pollo, proponiendo el siguiente modelo:

$$Q_i = \beta_1 + \beta_2 \, PM_i + \beta_3 \, PP_i + u_i \qquad i = 1, 2, \ldots, N. \tag{E9.1}$$

donde $Q$ representa la cantidad demandada de merluza en cientos de kilos, $PM$ el precio de la merluza en euros y $PP$ el precio del pollo en euros. Se dispone de los datos siguientes sobre estas variables, correspondientes a veinte mercados diferentes.

| Obs. | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $Q_i$ | 3 | 1 | 8 | 3 | 5 | 6 | 9 | 8 | 7 | 6 | 8 | 7 | 4 | 6 | 7 | 9 | 8 | 6 | 5 | 7 |
| $PM_i$ | 6 | 5 | 6 | 5 | 4 | 4 | 3 | 4 | 5 | 7 | 6 | 5 | 8 | 7 | 6 | 7 | 6 | 4 | 5 | 5 |
| $PP_i$ | 3 | 1 | 5 | 2 | 3 | 2 | 4 | 3 | 4 | 5 | 5 | 4 | 3 | 4 | 5 | 6 | 5 | 6 | 3 | 3 |

**PRIMERA PARTE.** Especificación y estimación.

1. ¿Qué hipótesis se deberían establecer sobre los diferentes elementos de la regresión para que los estimadores MCO sean lineales e insesgados? ¿Y para que sean de mínima varianza?

2. Estima el modelo propuesto bajo las hipótesis consideradas en el apartado anterior.

3. Escribe la recta de regresión muestral e interpreta los coeficientes estimados.

4. ¿Cuáles serían las ventas estimadas de merluza si la merluza costara 6 euros el kilo y el precio del pollo fuera de 2 euros?

5. Si el precio de la merluza aumenta en dos euros manteniéndose el resto de los factores constantes, ¿cuál sería la variación estimada en la venta de la merluza?

6. Si el precio de la merluza aumenta en dos euros y el del pollo baja en 1 euro, ¿cuál sería la variación estimada en la venta de la merluza?

7. Propón un estimador insesgado de la varianza de las perturbaciones y calcúlalo con la información facilitada.

8. Estima la matriz de varianzas y covarianzas de los estimadores MCO.

9. En caso de conocer la verdadera matriz de varianzas y covarianzas, ¿debería ésta coincidir con la estimada?, ¿por qué?

10. Analiza la bondad de ajuste del modelo.

**SEGUNDA PARTE.** Contraste de hipótesis.

1. Contrasta la hipótesis de que si se aumenta en un euro el precio de la merluza disminuye la cantidad demandada de merluza en 600 kilos.
2. Establece un intervalo de confianza para el coeficiente $\beta_2$ e interpreta el resultado. Relaciona este resultado con el obtenido en el apartado anterior.
3. Contrasta la hipótesis de que los precios, tanto de la merluza como del pollo, influyen en la cantidad demandada de merluza en sentido opuesto pero en la misma cuantía.
4. Dada la respuesta del apartado anterior, ¿existe algún método para estimar el modelo obteniendo menores varianzas de los estimadores?, ¿cuál?

**TERCERA PARTE.** Predicción.

1. Obtén una predicción por punto y por intervalo de la cantidad demandada de merluza cuando ambos precios toman el valor de 10 euros.
2. Un tendero dice que si el precio de la merluza fuera de 5 euros y el del pollo de 2 euros no se vendería nada de merluza. ¿Consideras que está justificada esta opinión?

## Ejercicio E10. ¿Qué factores determinan la evolución de las ventas?

La empresa "La Limpia, S.A.", produce Brillante, una marca de detergente líquido. La compañía desea contar con un modelo para predecir la demanda de botellas de detergente con el objeto de planificar su producción y estimar las necesidades de materias primas y de almacenamiento. El experto de la empresa en este campo cree que las ventas de Brillante ($V$, en miles de botellas) vienen determinadas, entre otros, por los siguientes factores:

$P$ : precio de venta de fábrica en euros.

$PS$ : precio medio de los productos sustitutivos existentes en el mercado en euros.

$GP$ : gastos en publicidad en miles de euros.

$ICE$ : Indice de Crecimiento Económico en porcentaje.

**PRIMERA PARTE.** Especificación y estimación.

El experto de la empresa especifica el siguiente modelo de regresión en el que las ventas se determinan en función de los factores antes mencionados:

$$V_t = \beta_1 + \beta_2 P_t + \beta_3 PC_t + \beta_4 GP_t + \beta_5 ICE_t + u_t \quad t = 2009{:}1, .., 2011{:}6. \qquad \text{(E10.1)}$$

Para estimar este modelo dispone de datos mensuales de las variables mencionadas desde enero de 2009 hasta el mes de junio de 2011.

1. Estima la función de regresión poblacional.
2. Interpreta los coeficientes estimados que acompañan a las variables gastos en publicidad e ICE. ¿Tienen los signos esperados?
3. Contrasta si los gastos en publicidad influyen en las ventas.
4. Representa gráficamente los residuos del modelo. Analizando su evolución, ¿crees que se cumplen, en este modelo, los hipótesis básicas sobre las perturbaciones?, ¿por qué?

**SEGUNDA PARTE.** Reespecificación del modelo.

Visto el patrón de comportamiento de los residuos MCO del modelo estimado en la Primera Parte, el experto de la empresa propone un segundo modelo de regresión para las ventas:

$$V_t = \beta_1 + \beta_2 P_t + \beta_3 PC_t + \beta_4 GP_t + \beta_5 ICE_t + \beta_6 ICE_t^2 + u_t \quad t = 2009{:}1, .., 2011{:}6. \tag{E10.2}$$

1. ¿En qué se diferencia el modelo (E10.2) del modelo (E10.1)? ¿Qué efecto se pretende recoger en este nuevo modelo?
2. ¿Cumple el modelo (E10.2) la hipótesis básica de linealidad del modelo de regresión? ¿Por qué?
3. Estima por MCO los coeficientes del modelo (E10.2).
4. Escribe la recta de regresión muestral e interpreta los coeficientes estimados. ¿Tienen los signos esperados?
5. Representa gráficamente los residuos de este modelo y compáralos con los del modelo (E10.1).
6. ¿Son las variables explicativas del modelo conjuntamente significativas?
7. ¿Cuál es el incremento medio estimado en las ventas cuando el ICE crece en un punto manteniéndose el resto de las variables fijas? ¿Es constante a lo largo del tiempo?
8. En base a los resultados, ¿se puede concluir que la variable ICE influye en las ventas?

9. ¿Existe evidencia en la muestra sobre la significatividad individual de las variables relativas a los precios?

10. Contrasta si los gastos de publicidad influyen positivamente en las ventas del producto.

11. Analizados los resultados obtenidos en la estimación del modelo (E10.2), ¿qué problema tenía el modelo especificado en la Primera Parte del ejercicio? ¿Cuáles son las propiedades de los estimadores MCO en el modelo (E10.1)? ¿Qué modelo de los dos propuestos en el ejercicio eligirías para determinar el comportamiento de las ventas?, ¿por qué?

12. Estima el valor futuro de las ventas en el mes de julio del año 2011 bajo el supuesto de que los valores de las variables explicativas serán los siguientes:

| | $P$ | $PC$ | $GP$ | ICE |
|---|---|---|---|---|
| Julio 2011 | 3,65 euros | 4,10 euros | 7000 euros | 31 % |

## Ejercicio E11. ¿Construyo una piscina en mi casa rural?

Se desea conocer los determinantes del precio de alquiler de las casas rurales que se encuentran en un determinado valle. Para ello se cuenta con una muestra de 14 casas rurales de la zona que contiene información sobre el precio de alquiler de la casa rural durante los fines de semana ($P$, medido en euros), la superficie ($S$, medido en metros cuadrados), el número de dormitorios ($D$) y el número de aseos ($A$).

1. Con la información disponible, especifica y estima un modelo para determinar el precio de alquiler en función de las características mencionadas.

2. Interpreta los coeficientes estimados. ¿Tienen el signo esperado?

3. Contrasta la significatividad conjunta de las vaiables explicativas.

4. Contrasta la significatividad individual de las variables explicativas.

5. Dados los resultados de los contrastes anteriores, ¿tenemos algún problema muestral?, ¿se cumplen todas las hipótesis básicas del modelo de regresión lineal general?

6. Si se sabe que las casas rurales 3, 6, 7, 8 y 13 de la muestra tienen piscina:

   6.1. Explica cómo introducirías este nuevo factor en el modelo de regresión.

   6.2. Construye una variable que recoja el hecho de tener o no piscina.

6.3. Estima un modelo que incluya esta nueva variable e interpreta el coeficiente estimado que le acompaña.

6.4. ¿Influye positivamente la existencia de piscina en el precio de alquiler?

7. Contrasta la significatividad individual de las variables explicativas y compara el resultado con el obtenido anteriormente. ¿A qué se debe la diferencia?

8. El dueño de una nueva casa rural construida en el mismo valle (con 1512 metros cuadrados, dos habitaciones, un aseo y sin piscina) quiere alquilarla por 279 euros, ¿crees que este precio es razonable para la zona en cuestión?, ¿y si tuviera piscina?

## Ejercicio E12. ¿Es más caro un coche con motor diesel?

Una empresa de leasing quiere renovar por completo su flota automovilística. Antes de comprar los coches en cuestión quiere saber cuáles son los factores determinantes del precio de un coche para lo cual toma una muestra de 209 coches recogiendo dos de sus principales características, peso y aceleración.

**PRIMERA PARTE.** Especificación y estimación. Forma Funcional.

1. Analiza gráficamente la relación entre el precio de un coche y cada uno de los factores que determinan su precio, peso y aceleración. En base a los resultados de este análisis, propón un modelo que permita analizar el precio del coche en función de las características mencionadas.

2. Interpreta los coeficientes del modelo propuesto y comenta los signos esperados.

3. Estima por MCO el modelo y escribe la función de regresión muestral.

4. ¿Cuáles son las características del tercer coche de la muestra?

5. ¿En cuánto estimas el precio del primer coche de la muestra?, ¿cuál es su precio real?, ¿cómo se llama la diferencia entre estos dos precios y a qué se debe?

6. Si el precio de un coche aumenta en 100 kgr. manteniendo la aceleración constante, ¿en cuánto estimas que varía el precio medio del coche?

7. ¿Son las variables peso y aceleración individualmente significativas?, ¿y conjuntamente?

**SEGUNDA PARTE.** Variables explicativas cualitativas.

Se considera también que el tipo de motor del coche (diesel o gasolina) y la gama (baja, media, alta) pueden influir en su precio. Esta información para los 209 coches de la muestra está incluida en el fichero de datos correspondiente.

1. Propón un modelo de regresión lineal general para determinar el precio de los coches que incorpore de manera adecuada toda la información de que dispones. Explica detalladamente cómo lo has especificado.

2. Según el modelo propuesto en el apartado anterior:

   2.1. ¿Cuál es el precio medio para un coche diesel de gama alta?

   2.2. ¿Cuál es el precio medio para un coche gasolina de gama media, 1200 kilogramos de peso y una aceleración de 100 km/h a los 8 segundos?

   2.3. ¿Cuál es la diferencia entre el precio medio de un coche diesel de gama baja y uno de gama alta? ¿Y a igualdad de características?

   2.4. ¿Cuál es la diferencia entre el precio medio de un coche diesel y uno gasolina, manteniendo el resto de las características iguales?

3. Estima por MCO el modelo propuesto e interpreta los coeficientes estimados. ¿Tienen los signos esperados?

4. ¿Son las variables explicativas individualmente significativas?

5. ¿Cuál es la diferencia estimada en el precio entre un coche de gama baja y uno de gama alta si el resto de variables se mantienen constantes? ¿Es significativa esta diferencia?

6. ¿Cuál es la diferencia estimada en el precio entre un coche de gama media y uno de gama alta si el resto de las características son iguales? ¿Es significativa esta diferencia?

7. En base a los resultados de los contrastes de los dos apartados anteriores, ¿qué efecto tiene la gama sobre el precio del coche? Especifica un modelo que recoja esta relación y estímalo por MCO.

8. En base al modelo del apartado anterior, estima el precio del primer coche de la muestra. ¿Coincide con el obtenido en el apartado 5 de la Primera Parte? ¿Por qué?

9. Cuando se estaba recogiendo la muestra, el director de un concesionario comentó al entrevistador que un coche, por el único hecho de ser diesel, cuesta en promedio 3000 euros más que el mismo modelo en versión gasolina. Dados los resultados obtenidos, ¿qué se puede decir sobre la opinión del experto?

## Ejercicio E13. A distinto turno de trabajo, distinto salario.

El gerente de una empresa quiere analizar el salario mensual ($S$, en euros) de los trabajadores que están contratados en función de la experiencia ($E$, en años), el puesto que ocupan (técnico o administrativo) y el turno en el que trabajan (mañana, tarde y noche).

1. Propón un modelo que tenga en cuenta todas estas variables para determinar el salario medio de los trabajadores y que sea capaz de contemplar las siguientes situaciones:
   1.1. El puesto que ocupan los trabajadores puede afectar de forma diferente al salario medio percibido.
   1.2. Los trabajadores de distintos turnos pueden no percibir el mismo salario medio.
   1.3. El efecto de la experiencia sobre el salario medio de los trabajadores puede no ser el mismo para todos los puestos.
   1.4. Es probable que los trabajadores de los turnos de la mañana y de la tarde que ocupan el mismo puesto ganen en media el mismo salario.
2. Escribe la hipótesis nula y alternativa adecuadas para contrastar cada una de las situaciones planteadas anteriormente.
3. Escribe el modelo restringido que correspondería a cada situación.

## Ejercicio E14. ¿Existe discriminación laboral en la empresa NIHON.SA?

Se desea estimar el salario ($S$, en miles de euros) de los trabajadores de la empresa NIHON.SA que ocupan un determinado puesto en función de su nivel de formación ($F$, en años de educación) y su experiencia ($E$, en años que llevan ocupando el puesto). El modelo de regresión lineal propuesto para la determinación del salario es:

$$S_i = \beta_1 + \beta_2 F_i + \beta_3 E_i + u_i \qquad i = 1, 2, \ldots, 100. \tag{E14.1}$$

Una muestra de 100 trabajadores de la empresa proporciona los siguientes datos:

$$\sum S_i = 2489,017 \quad \sum F_i = 288 \quad \sum E_i = 119 \quad \sum F_i^2 = 2090 \quad \sum S_i^2 = 64870,3$$

$$\sum E_i^2 = 695 \quad \sum F_i E_i = 329 \quad \sum S_i F_i = 7869,098 \quad \sum S_i E_i = 3921,297$$

$$(X'X)^{-1} = \begin{bmatrix} 0,019276 & -0,0023 & -0,0022 \\ -0,0023 & 0,00079 & -0,00001967 \\ -0,0022 & -0,00001967 & 0,00180753 \end{bmatrix}$$

**PRIMERA PARTE.** Estimación.

1. Interpreta los coeficientes del modelo de regresión lineal (E14.1).
2. Según el criterio Mínimo Cuadrático Ordinario, ¿cuál es la función objetivo que hay que minimizar para estimar los coeficientes del modelo (E14.1)?
3. Obtén el sistema de ecuaciones normales correspondiente y sustituye los valores muestrales.
4. Escribe las matrices $X'X$ y $X'Y$ con los datos muestrales disponibles.
5. Estima por MCO los coeficientes del modelo y escribe la recta de regresión muestral.
6. En base al modelo estimado, contesta a las siguientes preguntas:
    6.1. El empleado A lleva trabajando en la empresa dos años más que el empleado B, ¿cuál es la diferencia estimada de salario entre ambos?
    6.2. El empleado C lleva trabajando en la empresa tres años menos que el trabajador D pero cuenta con 5 años más de formación, ¿cuál es la diferencia estimada de salario entre ambos?
    6.3. Si un trabajador se ha estado formando durante doce años y lleva tres años trabajando en la empresa, ¿cuál es su salario estimado?
7. Calcula e interpreta una medida de la bondad de ajuste del modelo.

**SEGUNDA PARTE.** Inferencia.

Basándote en los resultados de la estimación del modelo (E14.1):

1. Construye intervalos de confianza del 95 % para los coeficientes $\beta_2$ y $\beta_3$.
2. Contrasta la significatividad individual de las variables explicativas. Relaciona las conclusiones obtenidas con los resultados del apartado anterior.
3. Contrasta la significatividad conjunta de las variables explicativas.
4. El gerente de la empresa supone que el efecto de la formación sobre el salario es igual al de la experiencia.
    4.1. ¿Qué restricción impone la hipótesis del gerente sobre los coeficientes del modelo de regresión lineal?
    4.2. Estima los coeficientes del modelo sujeto a la restricción anterior.

4.3. ¿Cómo se denomina el estimador que has utilizado en el apartado anterior? ¿Cuáles son sus propiedades?

4.4. Contrasta la hipótesis planteada por el gerente.

4.5. De acuerdo a los resultados obtenidos en el apartado 4.4, ¿qué modelo propones para la determinación del salario? ¿por qué?

5. Un empleado de la empresa NIHON.SA afirma que su salario asciende a cuarenta y cinco mil euros anuales. Sabiendo que lleva un año trabajando en la empresa y que su formación ha durado diez años, ¿parece razonable su comentario?

6. Comenta el gráfico de los residuos de este modelo. ¿Crees que se podría incumplir alguna hipótesis básica del modelo?

**TERCERA PARTE.** Discriminación salarial: alternativa I.

El comité de empresa de NIHON.SA sostiene que existe discriminación salarial por razones de género. Para investigar este tema desea utilizar el modelo de determinación del salario desarrollado hasta el momento, pero incluyendo una variable explicativa más: el género. Esta variable cualitativa se puede incluir mediante la variable ficticia $G_i$ que toma el valor 1 si el empleado i-ésimo es una mujer y cero en caso contrario.

1. Calcula algunos valores muestrales de interés: salario medio total, salario medio para las mujeres y salario medio para los hombres, sabiendo que los primeros 31 individuos de la muestra son hombres.

2. Estima por MCO un modelo de regresión lineal simple del salario sobre el género.

   2.1. Compara los resultados de esta estimación con los obtenidos en el apartado 1.

   2.2. ¿Es la variable género significativa?

   2.3. ¿Puedes concluir que hay discriminación salarial por razones de género basándote únicamente en este resultado?

3. Escribe un modelo de regresión lineal que determine el salario en función de la formación, la experiencia y el género en términos lineales.

4. Interpreta los coeficientes del modelo propuesto en el apartado anterior.

5. Estima el modelo propuesto por MCO.

   5.1. ¿Cuál es el salario estimado para una mujer?

   5.2. ¿Cuál es el salario estimado para un hombre con cinco años de experiencia y 10 años de formación?

5.3. ¿Cuál es la diferencia salarial estimada entre un hombre y una mujer a igualdad de condiciones?

5.4. ¿Cuál es el efecto estimado, ceteris paribus, de la experiencia sobre el salario?, ¿y para una mujer?, ¿y para un hombre?

6. Interpreta los resultados obtenidos y redacta un informe para el comité de empresa que incluya tus conclusiones en lo que al tema de la discriminación salarial se refiere.

**CUARTA PARTE.** Discriminación salarial: alternativa II.

El comité de empresa estudia el informe presentado y cree que es incompleto porque piensa que la discriminación salarial se acentúa para mayores niveles de experiencia. Para poder contrastar esta sospecha, propone estimar el siguiente modelo:

$$S_i = \beta_1 + \beta_2 F_i + \beta_3 E_i + \beta_4 G_i + \beta_5 (E_i \times G_i) + u_i \qquad i = 1, 2, \ldots, 100. \quad \text{(E14.2)}$$

1. Explica detalladamente la especificación de este modelo: ¿qué variables afectan al salario?, ¿de qué forma?, ¿está contemplada o no la posibilidad de discriminación salarial?, ¿cómo?

2. Estima el modelo (E14.2) por MCO e interpreta los coeficientes estimados.

   2.1. ¿Cuál es el salario estimado para una mujer?

   2.2. ¿Cuál es el salario estimado para un hombre con cinco años de experiencia y 10 años de formación?

   2.3. ¿Cuál es la diferencia salarial estimada entre un hombre y una mujer a igualdad de condiciones?

   2.4. ¿Cuál es el efecto estimado, ceteris paribus, de la experiencia sobre el salario?, ¿y para una mujer?, ¿y para un hombre?

3. En base a los resultados obtenidos en la estimación del modelo (E14.2), ¿puedes concluir que existe discriminación salarial por género?

4. Explica al comité de empresa las conclusiones finales a las que has llegado respecto a la existencia de discriminación salarial por género en la empresa NIHON.SA.

## Ejercicio E15. ¿Es estacional el consumo de electricidad?

Un técnico de una empresa eléctrica quiere analizar la evolución temporal de la tendencia de la demanda de electricidad en una determinada región. Para este análisis dispone de datos trimestrales correspondientes al periodo 2007-2010.

**PRIMERA PARTE.** Especificación y estimación.

1. Propón un modelo que permita analizar la evolución temporal de la tendencia de la demanda de electricidad.
2. Escribe la matriz $X$ indicando todos sus valores.
3. Escribe las matrices $X'X$ y $X'Y$ indicando sus elementos en términos de sumatorios.
4. Estima por MCO el modelo de regresión propuesto en el apartado 1 y escribe la función de regresión muestral.
5. Obtén el gráfico de los residuos contra el tiempo y coméntalo.
6. Dado el gráfico de los residuos y la frecuencia temporal de los datos, ¿cómo especificarías un modelo de regresión lineal alternativo para la demanda de electricidad?
7. Estima el modelo que has propuesto en el apartado anterior y comenta el gráfico de los residuos.
8. Según el modelo estimado en el apartado 7, ¿cuál es la demanda estimada para la última observación de la muestra?

**SEGUNDA PARTE.** Contraste de hipótesis y predicción.

1. ¿Cuántas variables explicativas tiene el modelo que has especificado en el apartado 7 de la Primera Parte? Contrasta la significatividad individual y conjunta de las mismas.
2. El técnico de la empresa cree que el consumo medio de los segundos y terceros trimestres del año son iguales:

   2.1. Escribe el modelo restringido correspondiente a la restricción anterior y estímalo.

   2.2. ¿Existe evidencia empírica de que esta restricción se cumple?

3. ¿En cuánto se estima la demanda media de electricidad para los dos primeros trimestres del año 2011?

# Parte III

# Prácticas de Autoevaluación

# PRÁCTICA P1.

La inmobiliaria WOOHOUSE quiere analizar los factores que influyen en el precio de las viviendas que tiene a su disposición con el objetivo de fijar una política de precios. Para ello dispone de datos sobre las siguientes variables:

$P$: precio de la vivienda medido en miles de dólares.

$S$: superficie de la vivienda medida en pies cuadrados.

$H$: número de habitaciones de la vivienda.

El siguiente cuadro recoge un extracto de la muestra disponible.

| Obs. | 1 | 2 | 3 | 4 | 5 | ...... | 11 | 12 | 13 | 14 | 15 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| $P_i$ | 219 | 225 | 255 | 306 | 230 | ...... | 195 | 425 | 240 | 253 | 111 |
| $S_i$ | 1185 | 1421 | 1478 | 2205 | 1171 | ...... | 1448 | 1502 | 1536 | 1536 | 1535 |
| $H_i$ | 2 | 2 | 2 | 2 | 3 | ...... | 3 | 3 | 3 | 3 | 4 |

En base a la muestra se obtiene la siguiente información:

$$\sum_{i=1}^{N} P_i = 3600 \qquad \sum_{i=1}^{N} H_i = 42 \qquad \sum_{i=1}^{N} S_i = 21909 \qquad \sum_{i=1}^{N} P_i S_i = 5322245 \qquad \sum_{i=1}^{N} S_i^2 = 32780493$$

$$\sum_{i=1}^{N}(S_i - \bar{S})^2 = 780207,6 \qquad \sum_{i=1}^{N}(P_i - \bar{P})^2 = 62766,03 \qquad \sum_{i=1}^{N}(S_i - \bar{S})(P_i - \bar{P}) = 64085$$

### PRIMERA PARTE.

1. ¿Cuál es el tamaño muestral?, ¿cuánto cuesta la cuarta vivienda?, ¿cuántas habitaciones posee?

2. Considera el siguiente modelo de regresión lineal simple:

$$P_i = \alpha + \beta\, S_i + u_i \qquad i = 1, \ldots, N. \tag{P1.1}$$

   ¿Cuál es la variable explicada del modelo?, ¿cuál es la variable explicativa del modelo?, ¿qué elementos son aleatorios?

3. Escribe la recta de regresión poblacional.

4. Interpreta el coeficiente que acompaña a la variable superficie en el modelo (P1.1).

5. Escribe la función objetivo que hay que minimizar para estimar los coeficientes del modelo (P1.1) por Mínimos Cuadrados Ordinarios. Deriva las ecuaciones normales y sustituye los momentos muestrales por sus valores respectivos.

6. Estima el modelo (P1.1) por MCO y escribe la recta de regresión muestral.

7. Obtén el precio estimado para la primera vivienda de la muestra y el correspondiente residuo.

8. Construye un intervalo de confianza del 95 % para el coeficiente $\beta$. ¿Se puede concluir que la variable superficie es estadísticamente significativa?

## SEGUNDA PARTE.

En una segunda fase el gerente de la inmobiliaria piensa que el número de habitaciones también puede ser un factor importante a la hora de determinar el precio de la vivienda. Los resultados de la estimación de un modelo de regresión que incluye también esta variable explicativa son los siguientes:

$$\underset{(\hat{\sigma}_{\hat{\beta}})}{\widehat{P}_i} = \underset{(158{,}068)}{238,808} + \underset{(0{,}07653)}{0,06593}\, S_i - \underset{(32{,}228)}{33,968}\, H_i \qquad \sum_{i=1}^{N} \hat{u}_i^2 = 52630,2. \tag{P1.2}$$

1. ¿Cúal es el orden de la matriz de datos? Escribe las cinco primeras filas de dicha matriz.

2. Escribe la expresión del estimador MCO empleado en términos de sumatorios.

$$\hat{\beta}_{MCO} = \left[ \qquad\qquad\qquad\qquad \right]^{-1} \left[ \qquad\qquad \right]$$

3. Interpreta los coeficientes estimados. ¿Tienen el signo esperado?

4. Estima el precio de un piso de 1250 pies cudrados y 3 habitaciones.

5. ¿En cuánto estimas que aumentaría el precio de un piso si contara con una habitación más manteniéndose el resto de los factores constantes?

6. ¿En cuánto estimas que aumentaría el precio de un piso si contara con una habitación más y 120 pies cuadrados más?

7. Escribe la expresión del coeficiente de determinación, calcúlalo e interpreta su valor.

8. Propón un estimador insesgado para la varianza de las perturbaciones y calcúlalo.

9. ¿Son las variables explicativas conjuntamente significativas?

10. ¿Son las variables explicativas individualmente significativas?

### TERCERA PARTE.

La variable precio está medida en miles de dólares y se desea medirla en dólares. Utilizando el fichero de datos practicaP1.gdt:

1. Transforma los datos de la variable precio de miles de dólares a dólares y estima el modelo (P1.2) con estos nuevos datos. Escribe la recta de regresión muestral.

2. Interpreta los coeficientes estimados.

3. ¿Cuál es la suma de los residuos al cuadrado? Estima la varianza de las perturbaciones.

4. Estima la matriz de covarianzas de los estimadores MCO.

5. Calcula una medida de la bondad de ajuste e interpreta el resultado.

6. Contrasta la significatividad individual y conjunta de las variables explicativas.

7. Compara los resultados obtenidos con los precios medidos en miles de dólares y con los precios medidos en dólares.

### CUARTA PARTE.

Supongamos que la variable precio sigue medida en miles de dólares pero se desea medir la superficie en metros cuadrados en vez de en pies cuadrados. Utilizando el fichero de datos practicaP1.gdt:

1. Transforma los datos de la variable superficie de pies cuadrados a metros cuadrados y estima el modelo (P1.2) con estos nuevos datos. Escribe la recta de regresión muestral.

2. Interpreta los coeficientes estimados.

3. ¿Cuál es la suma de los residuos al cuadrado? Estima la varianza de las perturbaciones.

4. Estima la matriz de covarianzas de los estimadores MCO.

5. Calcula una medida de la bondad de ajuste e interpreta el resultado.

6. Contrasta la significatividad de las variables explicativas.

7. Compara los resultados obtenidos con la superficie medida en pies cuadrados y en metros cuadrados.

# PRÁCTICA P2.

El Ministerio de Educación de un país quiere analizar cómo depende el salario de un trabajador del nivel de educación de toda la familia. Para ello dispone de una muestra de 718 individuos para los que ha recogido información sobre el salario percibido, su nivel de educación, así como los niveles de educación de su padre y de su madre:

$S$: salario medio mensual del individuo en euros.

$ED$: nivel de educación del individuo en años.

$FED$: nivel de educación del padre del individuo en años.

$MED$: nivel de educación de la madre del individuo en años.

El siguiente cuadro recoge las primeras y las últimas observaciones de la muestra disponible.

| Obs. | $S_i$ | $ED_i$ | $FED_i$ | $MED_i$ |
|---|---|---|---|---|
| 1 | 1350 | 12 | 12 | 12 |
| 2 | 1469 | 12 | 8 | 8 |
| 3 | 1573 | 12 | 10 | 10 |
| 4 | 1392 | 12 | 8 | 7 |
| 5 | 1800 | 12 | 12 | 12 |
| 6 | 2449 | 12 | 12 | 9 |
| $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ |
| 713 | 2262 | 17 | 12 | 13 |
| 714 | 1057 | 14 | 13 | 16 |
| 715 | 2239 | 17 | 8 | 8 |
| 716 | 1325 | 12 | 9 | 8 |
| 717 | 1662 | 10 | 9 | 12 |
| 718 | 2262 | 17 | 12 | 12 |

### PRIMERA PARTE.

El experto contratado por el Ministerio para realizar el estudio comienza estimando la relación existente entre el salario percibido y el nivel de educación del propio individuo:

$$S_i = \beta_1 + \beta_2\, ED_i + u_i \qquad i = 1, \ldots, N. \tag{P2.1}$$

Empleando la siguiente información muestral,

$$\sum_{i=1}^{718} S_i = 1204591 \qquad \sum_{i=1}^{718} ED_i = 9814 \qquad \sum_{i=1}^{718} ED_i^2 = 137732$$

$$\sum_{i=1}^{718} S_i^2 = 2140967727 \qquad \sum_{i=1}^{718} S_i\, ED_i = 11821856074.$$

1. Escribe la función de regresión muestral del modelo propuesto.
2. Interpreta el coeficiente estimado que acompaña a la variable educación del individuo.
3. ¿Qué hipótesis básicas sobre la perturbación son necesarias para que el estimador MCO sea insesgado y de mínima varianza? ¿Y cuáles son necesarias para realizar inferencia?
4. En base a la información disponible, si estimaras el modelo con sólo seis observaciones, ¿sería mejor hacerlo con las seis primeras o con las seis últimas? Justifica tu respuesta.
5. Estima la función de regresión muestral con los salarios medidos en miles de euros. Interpreta el coeficiente estimado que acompaña a la variable educación del individuo.

### SEGUNDA PARTE.

En una segunda etapa el experto decide utilizar toda la información disponible, es decir, incluir las variables explicativas nivel de educación del padre y de la madre del individuo en el modelo de regresión para la determinación de los salarios. Se obtienen los siguientes resultados:

Modelo P2.2: MCO, usando las observaciones 1–718
Variable dependiente: $S$

| Variable | Coeficiente | Desv. Típica | Estadístico $t$ | Valor p |
|---|---|---|---|---|
| const | 803,840 | 92,6237 | 8,6786 | 0,0000 |
| $ED$ | 46,6073 | 7,22608 | 6,4499 | 0,0000 |
| $FED$ | 11,1165 | 5,57971 | 1,9923 | 0,0467 |
| $MED$ | 11,2765 | 6,49354 | 1,7366 | 0,0829 |

Suma cuadrados residuos 1,06e+08 $\qquad R^2$ 0,120364

1. Escribe el modelo de regresión lineal general propuesto.
2. Escribe la función de regresión muestral.
3. Interpreta los coeficientes estimados que acompañan a las variables nivel de educación del padre y nivel de educación de la madre.
4. Escribe la expresión del coeficiente de determinación e interpreta la bondad del ajuste realizado.
5. ¿Cuál es el salario medio mensual que se estima según el Modelo P2.2 para el primer individuo de la muestra?, ¿y para el último? ¿A qué se debe la diferencia?
6. ¿Se puede concluir que, ceteris paribus, por cada año adicional de educación del individuo aumenta el salario en cincuenta euros?
7. Construye un intervalo de confianza del 95 % para el coeficiente de la variable $EDU$. Relaciona este resultado con el obtenido en el apartado anterior.
8. Contrasta la significatividad individual y conjunta de las variables relacionadas con el nivel de educación del padre y de la madre. Comenta los resultados obtenidos.
9. Construye intervalos de confianza del 95 % para los coeficientes de las variables $FED$ y $MED$. Relaciona estos resultados con los del apartado anterior.

## TERCERA PARTE.

El experto piensa que los niveles de educación de los padres tienen el mismo efecto en la determinación del salario de un individuo y decide incorporar esta información al modelo.

1. Deriva el modelo de regresión que resulta de incorporar al Modelo P2.2 la hipótesis que propone el experto.
2. Dada la siguiente información:

$$\underset{(\hat{\sigma}_{\hat{\beta}})}{\widehat{S}_i} = \underset{(90{,}7939)}{804,116} + \underset{(7{,}2155)}{46,6030}\, ED_i + \underset{(3{,}02317)}{11,1884}\, (FED_i + MED_i) \quad R^2 = 0,12. \qquad \text{(P2.3)}$$

   ¿Qué puedes concluir sobre la hipótesis planteada por el experto?, ¿por qué?

3. Dada la respuesta del apartado anterior, ¿qué modelo elegirías para determinar el salario medio del individuo? Razona tu respuesta en términos de las propiedades de los estimadores empleados.

# PRÁCTICA P3.

Una empresa encarga a uno de sus técnicos que realice un estudio sobre el salario de sus empleados. Para ello recoge información de 336 empleados sobre las siguientes variables:

$S$: salario mensual actual del empleado en dólares.

$A$: edad en años por encima de la mayoría de edad (18 años).

$X$: experiencia laboral previa a la entrada en la empresa del empleado en años.

$E$: nivel de estudios máximo alcanzado: básico (B), medio (M) o superior (L).

El cuadro siguiente muestra las diez primeras observaciones de la muestra.

| Obs. | $S$ | $A_i$ | $X_i$ | $E_i$ |
|---|---|---|---|---|
| 1 | 1000 | 15 | 7 | M |
| 2 | 1130 | 36 | 14 | M |
| 3 | 1280 | 12 | 11 | M |
| 4 | 1360 | 16 | 9 | M |
| 5 | 1360 | 18 | 3 | L |
| 6 | 1440 | 36 | 9 | B |
| 7 | 1500 | 29 | 8 | L |
| 8 | 1500 | 26 | 1 | L |
| 9 | 1650 | 24 | 14 | L |
| 10 | 1800 | 19 | 6 | M |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |

## PRIMERA PARTE.

El técnico de la empresa comienza su análisis estimando un modelo en el que el salario depende sólo de la edad y la experiencia de los empleados, con los siguientes resultados:

Modelo P3.1: estimaciones MCO utilizando las 336 observaciones
Variable dependiente: $S$

| Variable | Coeficiente | Desv. Típica | Estadístico $t$ | Valor p |
|---|---|---|---|---|
| const | 3943,29 | 337,271 | 11,69 | 1,08e-026 |
| $A$ | -26,2979 | 15,8930 | -1,655 | 0,0989 |
| $X$ | 61,4109 | 14,9234 | 4,115 | 4,88e-05 |

Suma cuadrados residuos 1146540000 $R^2$ 0,049379

1. Escribe el modelo de regresión lineal general propuesto por el técnico.
2. Deriva las ecuaciones normales del problema de optimización MCO para el Modelo P3.1.
3. Interpreta el coeficiente estimado que acompaña a la variable experiencia. ¿Tiene el signo esperado?
4. ¿En cuánto se estima el salario actual del quinto individuo de la muestra?
5. ¿En cuánto se estima el salario de un recién licenciado de 21 años que acaba de ser contratado?
6. El individuo A tiene 10 años más que el individuo B y entraron a trabajar con la misma experiencia previa, ¿cuál es la diferencia de salario estimado entre ambos?
7. Escribe la expresión del coeficiente de determinación junto con su valor e interprétalo.
8. El técnico no está seguro de si la experiencia previa y la edad tienen efecto en la determinación del salario actual. ¿Existe evidencia empírica de que sea así?

### SEGUNDA PARTE.

A continuación, el técnico decide añadir la variable nivel de estudios al modelo de regresión lineal de determinación de los salarios con los siguientes resultados:

Modelo P3.2: estimaciones MCO utilizando las 336 observaciones
Variable dependiente: $S$

| Variable | Coeficiente | Desv. Típica | Estadístico $t$ | Valor p |
|---|---|---|---|---|
| const | 3028,05 | 368,389 | 8,220 | 4,68e-015 |
| $A$ | -27,8645 | 14,3069 | -1,948 | 0,0523 |
| $X$ | 63,5817 | 13,4356 | 4,732 | 3,29e-06 |
| $M$ | 508,577 | 256,450 | 1,983 | 0,0482 |
| $L$ | 2106,81 | 276,147 | 7,629 | 2,53e-013 |

Suma cuadrados residuos 921708000 $\quad R^2$ 0,235795

La variable nivel de estudios está dividida en tres niveles, básico (B), medio (M) o superior (L), por lo que para cuantificarla se han definido tres variables ficticias:

$$B_i = \begin{cases} 1 & \text{si el nivel máximo de estudios del individuo } i \in \text{ básicos} \\ 0 & \text{en otro caso} \end{cases}$$

$$M_i = \begin{cases} 1 & \text{si el nivel máximo de estudios del individuo } i \in \text{ medios} \\ 0 & \text{en otro caso} \end{cases}$$

$$L_i = \begin{cases} 1 & \text{si el nivel máximo de estudios del individuo } i \in \text{ superiores} \\ 0 & \text{en otro caso} \end{cases}$$

1. Escribe el modelo de regresión lineal general propuesto por el técnico.
2. Explica cómo se ha incluido la variable nivel de estudios en el Modelo P3.2.
3. Interpreta los coeficientes del modelo de regresión lineal propuesto.
4. Escribe los 10 primeros valores de la matriz de datos $X$.
5. ¿Cuáles son las características del primer individuo de la muestra? ¿En cuánto se estima su salario medio actual?
6. Contrasta la significatividad conjunta de las variables explicativas del modelo.
7. ¿Son las variables explicativas del modelo individualmente significativas?
8. Dados los resultados obtenidos, ¿qué modelo propondrías para la determinación de los salarios de esta empresa? Justifica tu respuesta.

# PRÁCTICA P4.

Un agente inmobiliario quiere analizar los factores que influyen en el precio de las viviendas en la ciudad de Houston. Para ello dispone de una muestra correspondiente a 321 viviendas sobre las siguientes variables:

$P$: precio de venta de la vivienda en miles de doláres U.S.A.

$A$: años de antigüedad de construcción de la vivienda.

$S$: superficie de la vivienda en metros cuadrados.

$B$: número de baños.

## PRIMERA PARTE.

En la siguiente tabla se resumen los resultados de la estimación de un modelo de regresión lineal general para determinar el precio de la vivienda en función de las variables explicativas seleccionadas:

Modelo P4.1: estimaciones MCO utilizando las 321 observaciones 1–321
Variable dependiente: $P$

| Variable | Coeficiente | Desv. Típica | Estadístico $t$ | Valor p |
|---|---|---|---|---|
| const | 8,76294 | 6,25635 | 1,4006 | 0,1623 |
| $A$ | -0,307546 | 0,0568832 | -5,4066 | 0,0000 |
| $S$ | 0,328068 | 0,0357940 | 9,1654 | 0,0000 |
| $B$ | 12,2529 | 3,20741 | 3,8202 | 0,0002 |

Suma cuadrados residuos 281066 $R^2$ 0,529875

**Matriz de covarianzas de los coeficientes**

| const | $A$ | $S$ | $B$ | |
|---|---|---|---|---|
| $39,1419$ | $-0,135865$ | $-0,0546263$ | $-9,9340$ | const |
| | $0,00323570$ | $-0,000556436$ | $0,0797150$ | $A$ |
| | | $0,00128121$ | $-0,0795505$ | $S$ |
| | | | $10,2875$ | $B$ |

1. Escribe el modelo de regresión lineal general propuesto por el agente inmobiliario.

2. Escribe la recta de regresión muestral.

3. Manteniéndose el resto de las características de la vivienda constantes, ¿en cuántos dólares se estima el valor de un cuarto de baño adicional? ¿Por qué?

4. Escribe la expresión del coeficiente de determinación. ¿Cuál es su valor? Interpreta el resultado obtenido.

5. Propón un estimador insesgado para la varianza de las perturbaciones. Calcúlalo.

6. Contrasta la significatividad conjunta de las variables explicativas.

7. Contrasta la significatividad individual de la variable antigüedad.

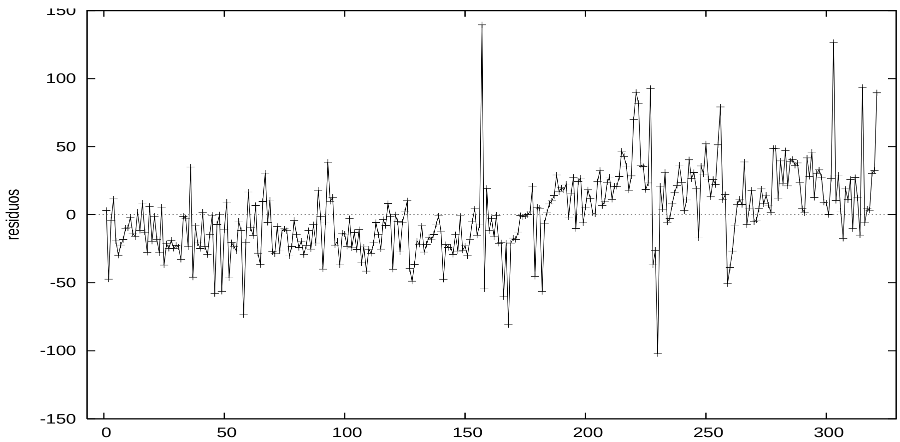
8. ¿Existe evidencia en la muestra de que, ceteris paribus, se está dispuesto a pagar 500 dólares por un metro cuadrado adicional de superficie?

9. El agente inmobiliario piensa que el efecto negativo sobre el precio de la antigüedad de una vivienda se ve compensado por el efecto de un aumento de la superficie. ¿Apoya la información disponible la hipótesis de este agente?

10. Dada la conclusión obtenida en el apartado anterior, ¿propondrías un modelo alternativo para la determinación del precio? ¿Por qué?

11. En la muestra se incluyen precios de viviendas vendidas en los años 1980 y 1981. Así, las primeras 179 observaciones corresponden a viviendas que se vendieron en 1980 y las restantes a viviendas que se vendieron en 1981. Además se sabe que en 1981 se puso en marcha una incineradora en la zona, lo que trajo consigo la construcción de una nueva autopista y, por lo tanto, una mejora en las comunicaciones.

    Teniendo en cuenta toda la información disponible, comenta el siguiente gráfico de los residuos del Modelo P4.1. ¿Crees que se podría incumplir alguna de las hipótesis básicas del MRLG?




## SEGUNDA PARTE.

Debido a la puesta en marcha de la incineradora en 1981 y a los efectos que ha podido tener sobre la mejora de las comunicaciones de la zona, el agente inmobiliario mantiene la teoría de que existen diferencias de precio entre las viviendas vendidas en el año 1980 y las vendidas en 1981.

1. Especifica un modelo de regresión lineal que tenga en cuenta la posible influencia de la puesta en marcha de la incineradora en la determinación del precio de la vivienda. ¿Cuáles son las características de la nueva variable que añadirías en el modelo para recoger este efecto? ¿Qué valor toma esta variable para la primera observación de la muestra? ¿Y para la última?

2. Denominemos por $INC$ a una variable que toma el valor 1 para las viviendas vendidas en el año 1981 y 0 para las vendidas en el año 1980. En la siguiente tabla se presentan los resultados de la estimación de un modelo de regresión que incluye esta nueva variable:

Modelo P4.2: estimaciones MCO utilizando las 321 observaciones 1–321
Variable dependiente: $P$

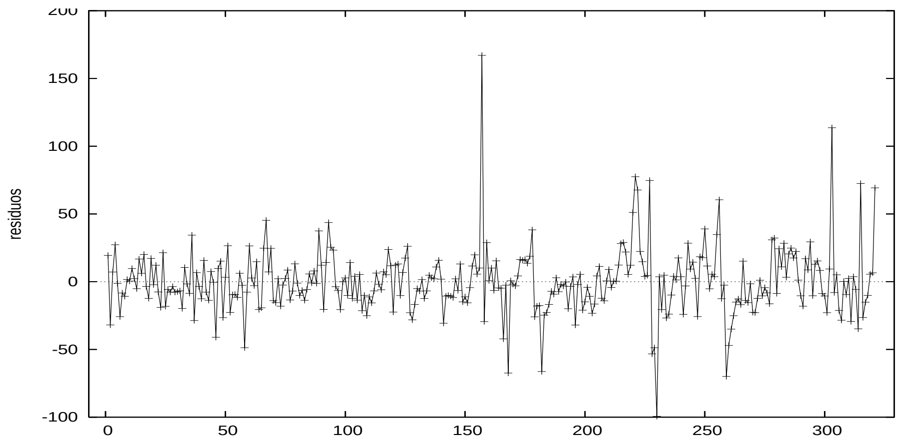
| Variable | Coeficiente | Desv. Típica | Estadístico $t$ | Valor p |
|---|---|---|---|---|
| const | -3,9605 | 5,19609 | -0,7622 | 0,4465 |
| $A$ | -0,211414 | 0,0469707 | -4,5010 | 0,0000 |
| $S$ | 0,241691 | 0,0299526 | 8,0691 | 0,0000 |
| $B$ | 17,4113 | 2,64524 | 6,5821 | 0,0000 |
| $INC$ | 35,7839 | 2,81746 | 12,7007 | 0,0000 |

Suma cuadrados residuos 186078 $R^2$ 0,688756

Interpreta el coeficiente que acompaña a la variable $INC$. ¿Tiene el signo esperado?

3. Contrasta la hipótesis del agente de que la puesta en marcha de la incineradora ha podido influir en el precio de venta.

4. Comenta el gráfico de los residuos del Modelo P4.2. Compáralo con el gráfico de los residuos del Modelo P4.1. ¿A qué crees que se debe la diferencia?




## TERCERA PARTE.

A continuación, el agente inmobiliario se cuestiona si el Modelo P4.2 recoge adecuadamente la influencia que ejerce la antigüedad de la vivienda sobre su precio de venta. La siguiente tabla muestra los resultados de la estimación de un nuevo modelo propuesto por el agente que incluye el término $sq\_A$ que representa la variable antigüedad elevada al cuadrado.

Modelo P4.3: estimaciones MCO utilizando las 321 observaciones 1–321
Variable dependiente: $P$

| Variable | Coeficiente | Desv. Típica | Estadístico $t$ | Valor p |
|---|---|---|---|---|
| const | 12,9328 | 6,82615 | 1,8946 | 0,0591 |
| $A$ | -0,698732 | 0,138937 | -5,0291 | 0,0000 |
| $sq\_A$ | 0,00328031 | 0,000882381 | 3,7176 | 0,0002 |
| $S$ | 0,239566 | 0,0293684 | 8,1572 | 0,0000 |
| $B$ | 12,0794 | 2,96337 | 4,0762 | 0,0001 |
| $INC$ | 36,3296 | 2,76589 | 13,1349 | 0,0000 |

Suma cuadrados residuos 178257 $R^2$ 0,701838

1. Escribe el modelo de regresión lineal general propuesto y estimado por el agente. ¿En qué se diferencia del Modelo P4.2? ¿Qué se pretende recoger con la inclusión del nuevo término $sq\_A$?

2. ¿Incumple este modelo alguna de las hipótesis básicas del modelo de regresión lineal general?, ¿por qué?

3. ¿Cómo contrastarías en el Modelo P4.3 la significativad de la variable antigüedad? Explica con detalle la hipótesis nula, el estadístico de contraste y la regla de decisión. Lleva a cabo el contraste utilizando el fichero de datos practicaP4.gdt.

4. Si la antigüedad de la casa aumentara en un año manteniéndose constante el resto de las características, ¿cuál sería la variación estimada en el precio medio de venta de las viviendas? ¿Y si la casa tiene una antigüedad de 5 años? ¿Y si la casa tiene una antigüedad de 50 años?

5. Con la información de que dispones, ¿estás de acuerdo con la afirmación del agente inmobiliario de que la relación entre el precio de la vivienda y su antigüedad no es lineal?

6. De acuerdo a tu respuesta en el apartado anterior, ¿qué modelo elegirías para determinar el precio de la vivienda? ¿Cuáles son los problemas que presentan los modelos que no has elegido? Razona tu respuesta y explica cuáles son las propiedades de los estimadores en los modelos que no has elegido.

7. Si, como agente inmobiliario, tuvieras que explicar a las autoridades locales tu teoría sobre la determinación de los precios de la vivienda en la zona, ¿qué aspectos destacarías de lo que has aprendido en este ejercicio y por qué?

# PRÁCTICA P5.

Una empresa encarga a su gerente el análisis de la evolución de sus ventas de aparatos refrigeradores a lo largo de los últimos ocho años. Para realizar este estudio se dispone de una muestra de datos trimestrales que abarcan el periodo comprendido desde el primer trimestre del año 2001 hasta el último trimestre del año 2008 para las siguientes variables:

$V$: ventas de refrigeradores en miles de euros.

$P$: gasto realizado en publicidad en cientos de euros.

El siguiente cuadro muestra un extracto de los datos disponibles.

| $t$ | 2001:1 | 2001:2 | 2001:3 | 2001:4 | 2002:1 | ... | 2008:2 | 2008:3 | 2008:4 |
|---|---|---|---|---|---|---|---|---|---|
| $V_t$ | 1317 | 1615 | 1662 | 1295 | 1271 | ... | 1684 | 1764 | 1328 |
| $P_t$ | 252,6 | 272,4 | 270,9 | 273,9 | 268,9 | ... | 350,3 | 369,1 | 356,4 |

Con la muestra disponible, se han estimado los siguientes modelos de regresión:

$$\underset{(\hat{\sigma}_{\hat{\beta}})}{\widehat{V}_t} = \underset{(59{,}990)}{1160} + \underset{(84{,}839)}{62,125}\ dq1_t + \underset{(84{,}839)}{307,500}\ dq2_t + \underset{(84{,}839)}{409,750}\ dq3_t \tag{P5.1}$$

$$SCR = 806142 \quad R^2 = 0,5318$$

$$\underset{(\hat{\sigma}_{\hat{\beta}})}{\widehat{V}_t} = \underset{(286{,}424)}{576,937} + \underset{(1{,}011)}{2,772}\ P_t \tag{P5.2}$$

$$SCR = 1377140 \quad R^2 = 0,2001$$

$$\underset{(\hat{\sigma}_{\hat{\beta}})}{\widehat{V}_t} = \underset{(183{,}469)}{370,164} + \underset{(0{,}623)}{2,773}\ P_t + \underset{(65{,}843)}{86,080}\ dq1_t + \underset{(65{,}793)}{328,578}\ dq2_t + \underset{(65{,}623)}{411,345}\ dq3_t \tag{P5.3}$$

$$SCR = 465085 \quad R^2 = 0,7298$$

$$\underset{(\hat{\sigma}_{\hat{\beta}})}{\widehat{V}_t} = \underset{(179{,}520)}{431,890} + \underset{(0{,}629)}{2,706}\ P_t + \underset{(57{,}583)}{285,319}\ dq2_t + \underset{(57{,}594)}{368,554}\ dq3_t \tag{P5.4}$$

$$SCR = 494526 \quad R^2 = 0,7128$$

donde $dq1$, $dq2$, $dq3$, $dq4$ son variables ficticias estacionales que se definen como sigue:

$$dq1_t = \begin{cases} 1 & \text{si } t \in 1^{\text{o}} \text{ trimestre} \\ 0 & \text{en otro caso} \end{cases} \qquad dq2_t = \begin{cases} 1 & \text{si } t \in 2^{\text{o}} \text{ trimestre} \\ 0 & \text{en otro caso} \end{cases}$$

$$dq3_t = \begin{cases} 1 & \text{si } t \in 3^{\text{o}} \text{ trimestre} \\ 0 & \text{en otro caso} \end{cases} \qquad dq4_t = \begin{cases} 1 & \text{si } t \in 4^{\text{o}} \text{ trimestre} \\ 0 & \text{en otro caso} \end{cases}$$

### PRIMERA PARTE. Modelo P5.1

1. Escribe el modelo de regresión lineal que corresponde al Modelo P5.1. ¿Cuántas variables explicativas hay en el modelo? Interpreta sus coeficientes.
2. ¿Presentan las ventas de refrigeradores comportamiento estacional?
3. ¿Cuáles son las ventas estimadas a lo largo de un año?

### SEGUNDA PARTE. Modelo P5.3

1. Escribe el modelo de regresión que corresponde al Modelo P5.3. ¿En qué se diferencia este modelo del Modelo P5.1 y del Modelo P5.2?
2. ¿Cuál es el orden de la matriz de datos del modelo? Escribe las cinco primeras filas y la última de la matriz de datos.
3. Contrasta la significatividad individual y conjunta de las variables explicativas.
4. ¿Cuál sería el incremento estimado en las ventas si aumentan los gastos en publicidad en 100 euros?
5. ¿En cuánto estimas las ventas medias para los primeros trimestres? ¿Y si el gasto en publicidad es de 1700 euros?
6. Calcula el intervalo de confianza del 95 % para el coeficiente que acompaña a la variable gasto en publicidad.
7. ¿Es posible que un aumento de un euro en la publicidad pueda aumentar las ventas medias en 50 euros?
8. Si en un trimestre se gasta en publicidad 250 euros más que en el mismo trimestre del año anterior, ¿cuál es la diferencia estimada en las ventas? ¿Y si es el primer trimestre?
9. El gerente comenta que en el primer trimestre del año 2009 no va a invertir en publicidad. Su asesor le indica que esta política podría tener como resultado no vender nada. ¿Qué opinas al respecto?

### TERCERA PARTE. Modelo P5.4

1. ¿Cuál es la restricción que relaciona el Modelo P5.4 con el Modelo P5.3? ¿Qué significa el cumplimiento de esta restricción?

2. Contrasta si la restricción del apartado anterior se cumple.

3. Estima las ventas medias del primer trimestre del año si el gasto realizado en publicidad es de 300 euros. ¿En cuánto estimarías las ventas correspondientes a los demás trimestres?

### CUARTA PARTE.

En base a todos los resultados obtenidos, ¿cuál es el modelo más adecuado para determinar las ventas de refigeradores de esta empresa? Razona tu respuesta en base a las propiedades del estimador empleado.

## PRÁCTICA P6.

El gerente de la empresa gallega "Dolfo" dedicada a la captura y venta al por menor de marisco quiere analizar la evolución de sus ventas en función de algunas variables que cree que pueden ser determinantes. Con este fin, construye una base de datos trimestrales que abarca desde el primer trimestre de 1999 al tercer trimestre de 2004 para las siguientes variables:

$V$: cantidad de marisco vendido por la empresa en miles de kilogramos.

$M$: precio medio de venta de marisco de la empresa en euros por kilogramo.

$C$: precio medio de venta de marisco de las empresas competidoras en euros por kilogramo.

$S$: sueldo medio de los trabajadores de la localidad en cientos de euros.

### PRIMERA PARTE.

La siguiente tabla proporciona los resultados obtenidos en la estimación de un modelo de regresión lineal general para explicar las ventas de marisco en función del precio de venta de la propia empresa, del precio medio de venta de las empresas competidoras y del sueldo medio.

Modelo P6.1: estimaciones MCO utilizando las 23 observaciones 1999:1–2004:3
Variable dependiente: $V$

| Variable | Coeficiente | Desv. Típica | Estadístico $t$ | Valor p |
|---|---|---|---|---|
| const | 33,7991 | 4,28653 | 7,8849 | 0,0000 |
| $M$ | -0,2223 | 0,12618 | -1,7618 | 0,0942 |
| $C$ | 0,0249 | 0,05908 | 0,4229 | 0,6771 |
| $S$ | 0,0129 | 0,00503 | 2,5770 | 0,0185 |

Suma cuadrados residuos 105,657 $R^2$ 0,911653

**Matriz de covarianzas de los coeficientes**

| const | $M$ | $C$ | $S$ | |
|---|---|---|---|---|
| 18, 3744 | −0, 439939 | −0, 100117 | 0, 0149171 | const |
| | 0, 0159228 | −0, 00121894 | −0, 000166764 | $M$ |
| | | 0, 00349152 | −0, 000266486 | $C$ |
| | | | 2, 53566e-05 | $S$ |

1. Escribe el modelo de regresión lineal general propuesto.
2. Escribe la función de regresión muestral e interpreta los coeficientes estimados.
3. Escribe la expresión del coeficiente de determinación. Calcula su valor e interprétalo.
4. Contrasta la significatividad conjunta de las variables explicativas.
5. Contrasta la significatividad individual de las variables explicativas del modelo. ¿Te parecen lógicos los resultados obtenidos?
6. Si el precio medio del marisco vendido por las empresas competidoras aumentase en un euro, manteniendo el resto de las variables explicativas fijas, ¿sería posible que la empresa Dolfo vendiera al menos 175 kilos más?
7. Si el sueldo medio de la población disminuyera en 100 euros, manteniéndose el resto de las variables explicativas constantes, ¿en cuántos kilos se estima que variaría la venta de marisco de la empresa?, ¿por qué?
8. El gerente cree que la variable renta media de los trabajadores de la localidad, $R$, puede ser un factor determinante para explicar la venta de marisco. Escribe el modelo de regresión lineal general que incluya esta variable junto con el resto de los factores.

   Teniendo en cuenta que la correlación muestral de la variable renta con la variable sueldo medio es muy alta, $\text{corr}(R, S) = r_{R,S} = 0{,}98$, ¿crees que el gerente puede encontrarse con algún problema en la estimación del modelo?, ¿cuál?, ¿cuáles serían las propiedades de los estimadores empleados?

9. Comenta el siguiente gráfico de los residuos del Modelo P6.1. ¿Crees que se podría incumplir alguna hipótesis básica del MRLG?


**SEGUNDA PARTE.**

A la vista del gráfico de los residuos del Modelo P6.1, el gerente recuerda que desde el tercer trimestre de 2000 hasta el cuarto trimestre de 2003 realizó una intensa campaña publicitaria a favor de sus productos. Con el objeto de recoger la influencia de la publicidad sobre sus ventas, define la variable $PUB$ que toma valor 1 si en el trimestre t se ha hecho publicidad y 0 en caso contrario.

La siguiente tabla muestra los resultados que se obtienen en la estimación de un modelo de regresión lineal que incluye esta nueva variable junto con las anteriores.

Modelo P6.2: estimaciones MCO utilizando las 23 observaciones 1999:1–2004:3
Variable dependiente: $V$

| Variable | Coeficiente | Desv. Típica | Estadístico $t$ | Valor p |
|---|---|---|---|---|
| const | 31,0580 | 1,47311 | 21,0834 | 0,0000 |
| $PUB$ | 4,2235 | 0,34863 | 12,1147 | 0,0000 |
| $M$ | -0,2754 | 0,04307 | -6,3954 | 0,0000 |
| $C$ | 0,0687 | 0,02038 | 3,3720 | 0,0034 |
| $S$ | 0,0103 | 0,00172 | 6,0018 | 0,0000 |

Suma cuadrados residuos 11,5427 $R^2$ 0,990348

**Matriz de covarianzas de los coeficientes**

| const | $PUB$ | $M$ | $C$ | $S$ | |
|---|---|---|---|---|---|
| 2, 17004 | −0, 0788814 | −0, 0497389 | −0, 0123623 | 0, 00176931 | const |
| | 0, 121546 | −0, 00152991 | 0, 00125932 | −7, 57169e-05 | $PUB$ |
| | | 0, 00185540 | −0, 000156414 | −1, 82774e-05 | $M$ |
| | | | 0, 000415674 | −3, 15144e-05 | $C$ |
| | | | | 2, 97118e-06 | $S$ |

1. Escribe el modelo de regresión lineal general propuesto. ¿Cuál es la diferencia con respecto al Modelo P6.1?

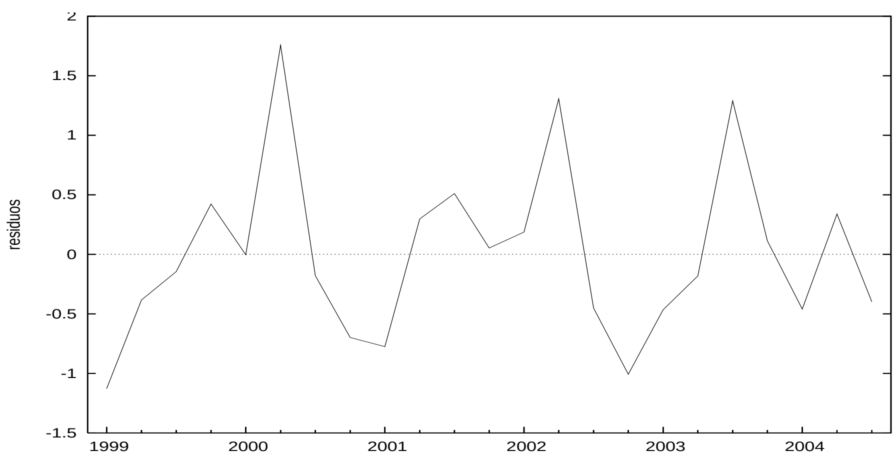
2. Interpreta el coeficiente estimado de la variable $PUB$.

3. Con la información disponible, ¿se puede concluir que la campaña publicitaria ha tenido un efecto positivo en las ventas?

4. ¿Qué puedes decir sobre la significatividad individual de la variable precio medio de venta de marisco? Compara este resultado con el obtenido en el quinto apartado de la Primera Parte. ¿A qué pueden ser debidos estos resultados?

5. ¿Existe evidencia en la muestra de que el gerente pueda contrarrestar una bajada del precio de sus competidores con una disminución de sus precios en la misma cuantía?

6. Compara el gráfico de los residuos del Modelo P6.2 con el gráfico de los residuos del del Modelo P6.1. ¿Qué conclusiones obtienes?



7. Teniendo en cuenta los resultados obtenidos hasta el momento, el gerente diseña una campaña publicitaria para el último trimestre de 2004. Dados los siguientes valores para las variables explicativas, ¿cuál es el volumen de ventas que estima conseguir?

| $M$ | $C$ | $S$ |
|---|---|---|
| 30 euros | 32 euros | 25000 euros |

## TERCERA PARTE.

Como los datos de la muestra son trimestrales y la cantidad de marisco que se vende en el último trimestre parece ser superior a la del resto del año, el gerente sospecha que la estacionalidad es un factor importante para explicar las ventas y que debe, por lo tanto, incluirse en la especificación del modelo.

1. Especifica detalladamente cómo modificarías el Modelo P6.2 para recoger el efecto estacional en las ventas de marisco.

2. Escribe las ocho primeras filas de la matriz de datos $X$ correspondiente al modelo que has especificado en el apartado anterior sustituyendo todos los valores conocidos.

3. Utilizando el fichero de datos practicaP6.gdt:

   3.1. Contrasta la hipótesis del agente de que la estacionalidad influye sobre la cantidad de marisco vendida.

   3.2. ¿Existe evidencia en la muestra de que solamente las ventas medias del último trimestre son diferentes al las del resto de los trimestres?

4. En base a los resultados obtenidos hasta el momento ¿qué modelo propondrías para determinar las ventas de marisco? Justifica tu respuesta.

# PRÁCTICA P7.

Se quieren analizar los factores determinantes de las ventas de una empresa que produce silicona para la construcción. Para ello se cuenta con datos mensuales desde enero de 1983 hasta mayo de 1990 sobre las siguientes variables:

$Q$: cantidad de silicona vendida en miles de galones por mes.

$P$: precio del galón de silicona en dólares.

$CC$: casas comenzadas a construir en miles.

$ICPP$: índice combinado de construcción pública y privada.

## PRIMERA PARTE.

Considera el siguiente modelo de regresión lineal en el que se supone que la cantidad vendida de silicona depende sólo del precio:

$$Q_t = \beta_1 + \beta_2 P_t + u_t \qquad u_t \sim NID(0, \sigma_u^2) \qquad t = 1983{:}1, 1983{:}2, \ldots, 1990{:}5. \qquad \text{(P7.1)}$$

1. En base a la siguiente información muestral, estima los coeficientes del modelo (P7.1) y escribe la función de regresión muestral.

$$\sum Q_t = 225,3035 \qquad \sum (P_t - \overline{P})^2 = 198,06658 \qquad \sum P_t = 801,1691$$

$$\sum P_t^2 = 7407,915 \qquad \sum (Q_t - \overline{Q})^2 = 217,897462 \qquad \sum Q_t^2 = 785,789900$$

$$\sum_{t=1}^{89} Q_t P_t = 1953,550 \qquad \sum_{t=1}^{89} (Q_t - \overline{Q})(P_t - \overline{P}) = -75,48641649$$

2. Interpreta el coeficiente estimado que acompaña a la variable precio. ¿Tiene el signo esperado?

3. Propón un estimador insesgado para la varianza de las perturbaciones y calcula su estimación con la información muestral disponible.

4. ¿Existe evidencia en la muestra de que la variable precio influye negativamente en la cantidad media de silicona vendida?

5. ¿Qué hipótesis básicas del modelo de regresión lineal general son necesarias para que se cumpla el Teorema de Gauss-Markov?

6. El dueño de la empresa quiere aumentar sus ventas y para ello pone en práctica una política de bajada de precios. Piensa que si en junio de 1990 fija el precio de la silicona en 9,5 dólares por galón alcanzará unas ventas ese mes de 3500 galones. En base a la información disponible, ¿crees que el dueño está en lo cierto?

## SEGUNDA PARTE.

Considerando que, además del precio, las ventas de silicona pueden depender del número de casas que se van a comenzar a construir y del índice combinado de construcción pública y privada, se estima un modelo de regresión que incluye estas variables con los siguientes resultados:

Modelo P7.2: estimaciones MCO utilizando las 89 observaciones 1983:01–1990:05
Variable dependiente: $Q$

| | Coeficiente | Desv. Típica | Estadístico $t$ | Valor p |
|---|---|---|---|---|
| const | −1,0331 | 1,64879 | −0,6266 | 0,5326 |
| $P$ | −0,326389 | 0,0959225 | −3,4026 | 0,0010 |
| $CC$ | 0,0228171 | 0,00491616 | 4,6412 | 0,0000 |
| $ICPP$ | 0,223760 | 0,00813253 | 2,7514 | 0,0073 |

Suma cuadrados residuos 140,69 $R^2$ 0,346945

**Matriz de covarianzas de los coeficientes**

| const | $P$ | $CC$ | $ICPP$ | |
|---|---|---|---|---|
| 2,71850 | −0,0767027 | −0,00448977 | −0,00917623 | const |
| | 0,00920113 | 0,000104349 | −0,00013221 | $P$ |
| | | 0,000024 | 0,0000019 | $CC$ |
| | | | 0,000066138 | $ICPP$ |

1. Escribe el modelo de regresión lineal general que corresponde al Modelo P7.2.

2. Las nuevas variables explicativas incluidas en el modelo, $CC$ y $ICCP$, ¿son individualmente significativas? ¿Y conjuntamente? ¿Observas alguna contradicción en los resultados obtenidos?

3. Dados los resultados obtenidos, ¿se puede afirmar que el estimador MCO del coeficiente $\beta_2$ en el Modelo P7.1 es el de varianza mínima? Justifica detalladamente tu propuesta.

4. Si en el próximo mes el precio bajara en un dólar, el número de casas comenzadas a construir aumentara en 500 unidades y el índice $ICPP$ se incrementara en veinte unidades, ¿en cuánto estimas que incrementaría la venta de silicona?

## TERCERA PARTE.

El dueño de la empresa estima el siguiente modelo de regresión lineal porque cree que es el apropiado para la determinación de las ventas de silicona, donde el término $sq\_P$ significa la variable precio al cuadrado:

Modelo P7.3: estimaciones MCO utilizando las 89 observaciones 1983:01–1990:05
Variable dependiente: $Q$

| | Coeficiente | Desv. Típica | Estadístico $t$ | Valor p |
|---|---|---|---|---|
| const | −0,197276 | 4,17528 | −0,0472 | 0,9624 |
| $P$ | −0,497857 | 0,792009 | −0,6286 | 0,5313 |
| $sq\_P$ | 0,00828892 | 0,0380014 | 0,2181 | 0,8279 |
| $CC$ | 0,0228098 | 0,00494405 | 4,6136 | 0,0000 |
| $ICPP$ | 0,0224988 | 0,00819782 | 2,7445 | 0,0074 |

Suma cuadrados residuos 140,6150 $R^2$ 0,347315

1. Escribe el modelo de regresión lineal teórico que corresponde al Modelo P7.3. ¿En qué se diferencia el Modelo P7.3 del Modelo P7.2? Explícalo detalladamente.
2. ¿Cumple el Modelo P7.3 las hipótesis básicas del modelo de regresión lineal general?
3. Según el Modelo P7.3, ¿cuál es el efecto marginal estimado de un incremento unitario en el precio sobre las ventas, ceteris paribus? ¿Y si el precio es de 5 dólares? ¿Y si es de 12 dólares?
4. ¿Existe evidencia en la muestra que apoye la hipótesis del dueño acerca de la forma funcional para la relación entre la cantidad vendida y el precio recogida en el Modelo P7.3?

## CUARTA PARTE.

Como los datos son mensuales y las ventas de silicona parecen presentar comportamiento estacional, se definen las siguientes variables ficticias estacionales, $dmi_t$, $i = 1, 2, \ldots, 12$:

$$dmi_t = \begin{cases} 1 & \text{si } t \in \text{mes i} \\ 0 & \text{en otro caso} \end{cases}$$

La siguiente tabla muestra los resultados de la estimación de un modelo de regresión en el que se incluyen además de las variables consideradas en los modelos anteriores, las variables ficticias definidas arriba.

Modelo P7.4: estimaciones MCO utilizando las 89 observaciones 1983:01–1990:05
Variable dependiente: $Q$

| Variable | Coeficiente | Desv. Típica | Estadístico $t$ | Valor p |
|---|---|---|---|---|
| const | -0,897081 | 2,12126 | -0,4229 | 0,6736 |
| $P$ | -0,266946 | 0,101948 | -2,6185 | 0,0107 |
| $CC$ | 0,0213323 | 0,00919383 | 2,3203 | 0,0231 |
| $ICPP$ | 0,0187688 | 0,00878573 | 2,1363 | 0,0360 |
| $dm1$ | -0,309257 | 0,655500 | -0,4718 | 0,6385 |
| $dm2$ | 0,0507874 | 0,657504 | 0,0772 | 0,9386 |
| $dm3$ | 0,00704305 | 0,723664 | 0,0097 | 0,9923 |
| $dm4$ | -0,278333 | 0,813952 | -0,3420 | 0,7334 |
| $dm5$ | 0,678127 | 0,845657 | 0,8019 | 0,4252 |
| $dm6$ | -0,916889 | 0,909544 | -1,0081 | 0,3167 |
| $dm7$ | -0,130321 | 0,835786 | -0,1559 | 0,8765 |
| $dm8$ | 1,26451 | 0,824631 | 1,5334 | 0,1294 |
| $dm9$ | 0,566385 | 0,780856 | 0,7253 | 0,4705 |
| $dm10$ | 0,251659 | 0,818026 | 0,3076 | 0,7592 |
| $dm11$ | -0,175507 | 0,699423 | -0,2509 | 0,8026 |

Suma cuadrados residuos 117,217 $R^2$ 0,455918

1. Escribe el modelo de regresión teórico que corresponde al Modelo P7.4. ¿En qué se diferencia del Modelo P7.2? Explícalo detalladamente.
2. Interpreta los coeficientes estimados $-0,897081$ y $0,251659$.
3. ¿Las ventas de silicona presentan, ceteris paribus, comportamiento estacional?
4. De todos los modelos que se han estimado hasta el momento, ¿cuál consideras que es el más apropiado para representar la determinación de las ventas de silicona? Razónalo.

# PRÁCTICA P8.

La compañía SXF de New Orleans desea estudiar cómo se determinan los salarios de sus empleados en función de las características de los trabajadores, con el objetivo de responder a las reivindaciones de los sindicatos que denuncian la existencia de discriminación salarial por razones de género.

Para ello cuenta con datos de 49 trabajadores sobre sus salarios mensuales ($S$, en dólares) y algunas de sus características principales como son:

$ED$: educación medida en número de años de educación por encima de la educación obligatoria.

$EX$: experiencia medida en número de años en la empresa.

Género: hombre o mujer.

Se especifica un modelo de regresión lineal para determinar los salarios en función de la educación, la experiencia y el género que se estima por Mínimos Cuadrados Ordinarios. Los resultados de esta estimación se recogen en la siguiente tabla.

Modelo P8.1: estimaciones MCO utilizando las 49 observaciones 1–49
Variable dependiente: $S$

| Variable | Coeficiente | Desv. Típica | Estadístico $t$ | Valor p |
|---|---|---|---|---|
| const | 434,821 | 258,871 | 1,6797 | 0,0999 |
| $ED$ | 133,551 | 31,5081 | 4,2386 | 0,0001 |
| $EX$ | 34,4543 | 12,1316 | 2,8401 | 0,0067 |
| $HB$ | 470,460 | 144,870 | 3,2475 | 0,0022 |

Suma cuadrados residuos 11089355 $R^2$ 0,450263

donde $HB$ es una variable que toma el valor 1 para los hombres y 0 para las mujeres.

### PRIMERA PARTE.

1. Escribe el modelo de regresión lineal y la recta de regresión muestral correspondientes al Modelo P8.1.
2. Interpreta los coeficientes estimados 34,4543 y 470,460.
3. Escribe la expresión del coeficiente de determinación e interpreta la bondad de ajuste del modelo.
4. Contrasta la significatividad conjunta de las variables explicativas del modelo.
5. ¿Existe discriminación salarial por razones de género?

## SEGUNDA PARTE.

La dirección de la empresa no está conforme con las conclusiones que se obtienen a partir de la estimación del Modelo P8.1. Cree que a la hora de determinar el salario no se ha tenido en cuenta una variable fundamental como es el tipo de ocupación de cada empleado. En particular en esta empresa, hay que distinguir entre empleados de taller, mantenimiento, administrativos y técnicos.

Los resultados de la estimación del modelo de regresión añadiendo la variable ocupación a las ya consideradas en el Modelo P8.1 son los siguientes:

Modelo P8.2: estimaciones MCO utilizando las 49 observaciones 1–49
Variable dependiente: $S$

| Variable | Coeficiente | Desv. Típica | Estadístico $t$ | Valor p |
|---|---|---|---|---|
| const | 502,034 | 239,845 | 2,0932 | 0,0424 |
| $ED$ | 63,9771 | 26,9073 | 2,3777 | 0,0221 |
| $EX$ | 29,2516 | 9,57726 | 3,0543 | 0,0039 |
| $HB$ | 528,154 | 151,238 | 3,4922 | 0,0011 |
| $ADM$ | 248,672 | 212,599 | 1,1697 | 0,2487 |
| $TAL$ | 382,987 | 164,408 | 2,3295 | 0,0247 |
| $TEC$ | 1110,05 | 203,871 | 5,4449 | 0,0000 |

Suma cuadrados residuos 5620313 $R^2$ 0,721382

donde:

$$ADM_i = \begin{cases} 1 & \text{si } i \in \text{administrativo} \\ 0 & \text{en otro caso} \end{cases} \qquad TAL_i = \begin{cases} 1 & \text{si } i \in \text{taller} \\ 0 & \text{en otro caso} \end{cases}$$

$$TEC_i = \begin{cases} 1 & \text{si } i \in \text{técnico} \\ 0 & \text{en otro caso} \end{cases} \qquad MAN_i = \begin{cases} 1 & \text{si } i \in \text{mantenimiento} \\ 0 & \text{en otro caso} \end{cases}$$

1. Escribe el modelo de regresión lineal general propuesto y explica detalladamente cómo se ha incluido la variable explicativa ocupación en el mismo.

2. ¿Tiene razón la empresa cuando afirma que la ocupación es una variable significativa en la determinación de los salarios?

3. Según la conclusión obtenida en el apartado anterior, ¿cuáles son las propiedades de los estimadores MCO en el Modelo P8.1? ¿Te puedes fiar de los contrastes realizados en la Primera Parte del ejercicio? ¿Existe, por tanto, discriminación salarial por razones de género?

4. Dados los resultados de la estimación del Modelo P8.2, ¿cuál es la diferencia salarial estimada entre dos administrativas con el mismo nivel de educación si la experiencia de una de ellas es de 3 años y la de la otra es de 4 años?

5. La mayor diferencia de salarios, a igualdad de condiciones, se produce entre técnicos y empleados de mantenimiento. La empresa se comprometió el año anterior a que esta diferencia no fuera superior a mil dólares. Dados los resultados de la estimación del Modelo P8.2, ¿ha cumplido la empresa su promesa?

### TERCERA PARTE.

Un experto de los sindicatos indica que no sólo existe discriminación salarial debida al género porque, a igualdad de condiciones, las mujeres ganan en media menos que los hombres, sino también porque por cada año más de experiencia se le paga más a un hombre que a una mujer. Para demostrar su teoría, los sindicatos presentan a la dirección de la empresa los resultados de la estimación del siguiente modelo:

Modelo P8.3: estimaciones MCO utilizando las 49 observaciones 1–49
Variable dependiente: $S$

| Variable | Coeficiente | Desv. Típica | Estadístico $t$ | Valor p |
|---|---|---|---|---|
| const | 739,825 | 240,696 | 3,0737 | 0,0038 |
| $ED$ | 67,1281 | 25,1391 | 2,6703 | 0,0108 |
| $EX$ | -8,1917 | 16,5544 | -0,4948 | 0,6234 |
| $HB$ | 52,1988 | 226,481 | 0,2305 | 0,8189 |
| $ADM$ | 237,551 | 198,455 | 1,1970 | 0,2382 |
| $TAL$ | 446,348 | 155,238 | 2,8752 | 0,0064 |
| $TEC$ | 1301,35 | 203,149 | 6,4059 | 0,0000 |
| $EX \times HB$ | 53,7363 | 19,9972 | 2,6872 | 0,0104 |

Suma cuadrados residuos 4778683 $R^2$ 0,763104

donde:

$$EX_i \times HB_i = \begin{cases} EX_i & \text{si } i \in \text{hombre} \\ 0 & \text{en otro caso} \end{cases}$$

1. Escribe el modelo de regresión lineal correspondiente al Modelo P8.3. ¿Cumple las hipótesis básicas del modelo de regresión lineal? En particular, ¿cumple la hipótesis de linealidad?

2. En igualdad de condiciones, ¿cuál es la diferencia salarial estimada entre un hombre y una mujer?

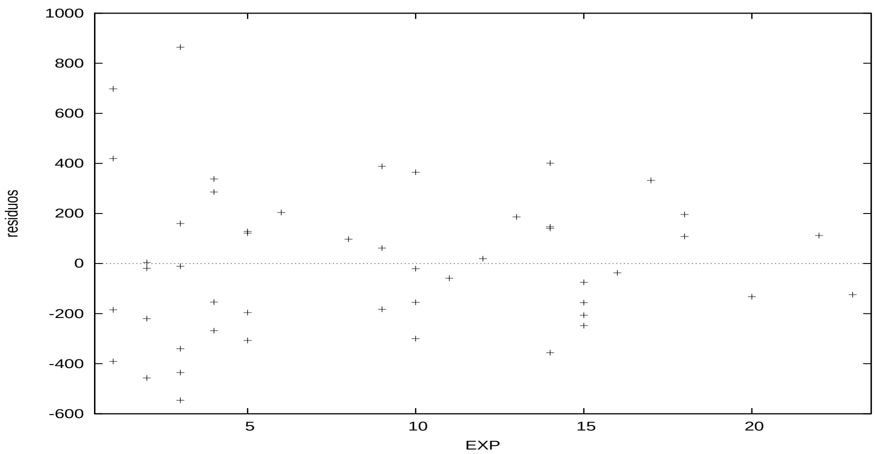
3. ¿Cuál es el incremento estimado en el salario por un año más de experiencia, manteniendo fijas el resto de las variables?

4. Interpreta el coeficiente estimado 53,7363.

5. ¿Existe evidencia en la muestra que apoye la hipótesis de los sindicatos: un año más de experiencia se valora de forma diferente a un hombre y a una mujer?

6. Utilizando los datos del fichero practicaP8.gdt, contrasta si existe o no discriminación salarial por razones de género en la empresa SXF.

7. En el siguiente gráfico se representan los residuos del Modelo P8.3 frente a la variable experiencia. Comentálo. ¿Crees que se podría incumplir alguna hipótesis básica del MRLG?




# PRÁCTICA P9.

Un agente inmobiliario quiere analizar los factores que influyen en el precio de las viviendas en Bilbao. Para ello dispone de una muestra correspondiente a 265 viviendas sobre las siguientes variables:

$P$: precio de venta de la vivienda en miles de euros.

$H$: número de habitaciones.

$B$: número de baños.

$S$: superficie de la vivienda en metros cuadrados.

## PRIMERA PARTE.

En la siguiente tabla se resumen los resultados de la estimación de un modelo de regresión lineal general para determinar el precio de la vivienda:

Modelo P9.1: estimaciones MCO utilizando las 265 observaciones 1–265
Variable dependiente: $P$

| Variable | Coeficiente | Desv. Típica | Estadístico $t$ | Valor p |
|---|---|---|---|---|
| const | -10,968 | 14,249 | -0,769 | 0,442 |
| $H$ | 0,655 | 2,660 | 0,246 | 0,805 |
| $B$ | 17,134 | 3,672 | 4,665 | 0,000 |
| $S$ | 0,304 | 0,041 | 7,354 | 0,000 |

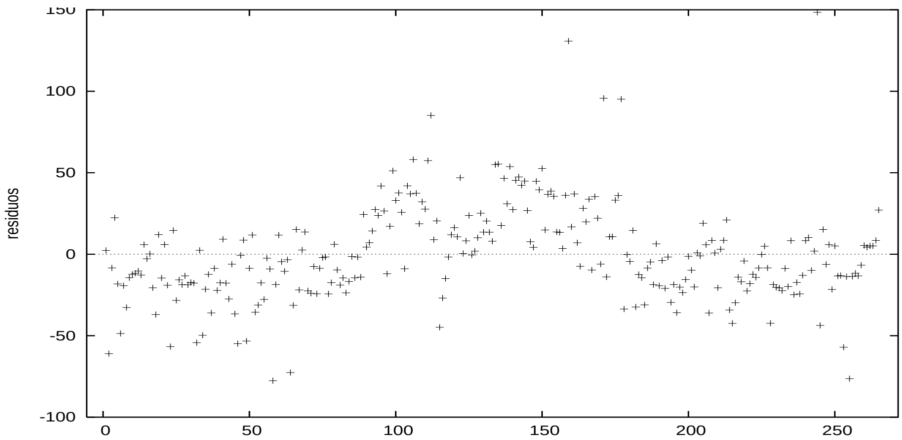
Suma cuadrados residuos 234507 $R^2$ 0,47943

**Matriz de covarianzas de los coeficientes**

| const | $H$ | $B$ | $S$ | |
|---|---|---|---|---|
| 203,059 | -33,3136 | 8,66392 | 0,00105675 | const |
| | 7,07734 | -3,84501 | -0,0226378 | $H$ |
| | | 13,4856 | -0,0777654 | $B$ |
| | | | 0,00170926 | $S$ |

1. Escribe el modelo de regresión lineal general propuesto por el agente inmobiliario y la recta de regresión muestral.
2. Manteniéndose el resto de las características de la vivienda constantes, ¿en cuántos euros se estima el valor de un metro cuadrado adicional?, ¿por qué?
3. Escribe la expresión del coeficiente de determinación. ¿Cuál es su valor? Interprétalo.
4. Contrasta la significatividad conjunta de las variables explicativas.
5. Contrasta la significatividad individual de la variable número de habitaciones.
6. Dadas las variables explicativas incluidas en el modelo, explica razonadamente a qué puede ser debido el resultado del apartado anterior.
7. ¿Existe evidencia en la muestra de que, ceteris paribus, se estaba dispuesto a pagar 20000 euros por un baño adicional?
8. En base a la recta de regresión muestral, ¿cuánto costaba un piso de 100 $m^2$ con cuatro habitaciones y dos baños?
9. Un amigo tuyo se compró un piso de las características señaladas en el apartado anterior y pagó por él 75000 euros. ¿Cómo le convencerías de que no le engañaron?
10. La muestra disponible cuenta con precios de viviendas correspondientes a tres zonas de Bilbao: las primeras 90 observaciones corresponden a viviendas ubicadas en la zona este, las siguientes 87 observaciones a casas situadas en la zona centro y las restantes 88 se encuentran en la zona oeste.

Teniendo en cuenta toda la información de que dispones, comenta el siguiente gráfico de los residuos del Modelo P9.1. ¿Crees que se podría incumplir alguna hipótesis básica del MRLG?




## SEGUNDA PARTE.

Dadas las caracterísiticas de la muestra y los resultados obtenidos en la Primera Parte, el agente inmobiliario mantiene la teoría de que existen diferencias en el precio de las viviendas según sea su localización. En la siguiente tabla se presentan los resultados de la estimación de un modelo de regresión que incluye la variable explicativa localización.

Modelo P9.2: estimaciones MCO utilizando las 265 observaciones 1–265
Variable dependiente: $P$

| Variable | Coeficiente | Desv. Típica | Estadístico $t$ | Valor p |
|---|---|---|---|---|
| const | 24,510 | 11,522 | 2,127 | 0,034 |
| $H$ | 1,919 | 2,086 | 0,919 | 0,358 |
| $B$ | 16,414 | 2,902 | 5,654 | 0,000 |
| $S$ | 0,227 | 0,032 | 6,894 | 0,000 |
| $O$ | -43,851 | 3,630 | -12,077 | 0,000 |
| $E$ | -37,114 | 3,617 | -10,259 | 0,000 |

Suma cuadrados residuos 141896 $R^2$ 0,685011

**Matriz de covarianzas de los coeficientes**

| const | $H$ | $B$ | $S$ | $O$ | $E$ | |
|---|---|---|---|---|---|---|
| 132,76 | -20,25 | 4,58 | -0,011 | -10,8524 | -5,063 | const |
| | 4,35 | -2,27 | -0,014 | -0,104 | -0,64 | $H$ |
| | | 8,42 | -0,048 | 0,78 | -0,83 | $B$ |
| | | | 0,00108 | 0,015 | 0,024 | $S$ |
| | | | | 13,18 | 6,64 | $O$ |
| | | | | | 13,08 | $E$ |

donde $O$ es una variable que toma valor 1 si la vivienda está en la zona oeste y 0 en otro caso y la variable $E$ toma valor 1 si la vivienda está en la zona este y 0 en otro caso.

1. Escribe el modelo de regresión lineal propuesto y estimado por el agente inmobiliario. Explica detalladamente cómo se ha introducido la nueva variable explicativa localización en este modelo.

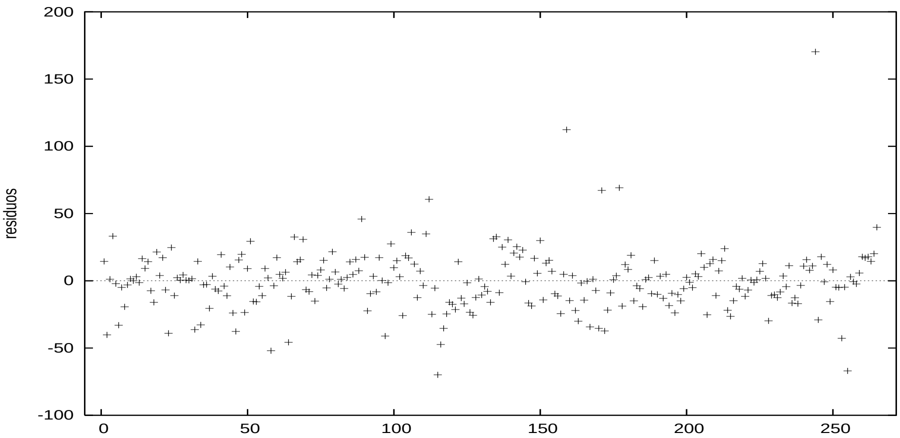
2. ¿Qué valores toman las variables $E$ y $O$ para la primera observación de la muestra?, ¿y para la centésima?, ¿y para la última?

3. Escribe la recta de regresión muestral.

4. Interpreta los coeficientes estimados -43,851 y -37,114.

5. Contrasta la hipótesis del agente inmobiliario de que la localización influye en el precio de venta.

6. Comenta el gráfico de los residuos del Modelo P9.2. Compáralo con el gráfico de los residuos del Modelo P9.1.




¿A qué crees que se debe la diferencia?

7. Dados los resultados de la estimación del Modelo P9.2 y tras observar con detenimiento el gráfico de los residuos del Modelo P9.1, el agente sospecha que existe diferencia de precio debido a la localización, pero que esta diferencia sólo se observa entre los pisos de la zona centro y el resto.

   Basándote en la información de la que dispones, ¿tiene razón este agente?

8. Dada la conclusión que has obtenido en el apartado anterior, ¿propondrías un modelo alternativo para la determinación del precio? Justifica tu respuesta.

## TERCERA PARTE.

Supongamos que el agente ha concluido que la variable localización sólo distingue entre la zona centro y la periferia (zonas este y oeste juntas). A pesar de todo, el agente inmobiliario se sigue cuestionando si el modelo recoge adecuadamente la influencia que ejerce la localización de la vivienda sobre su precio de venta.

La siguiente tabla muestra los resultados de la estimación del nuevo modelo que propone el agente.

Modelo P9.3: estimaciones MCO utilizando las 265 observaciones 1–265
Variable dependiente: $P$

| Variable | Coeficiente | Desv. Típica | Estadístico $t$ | Valor p |
|---|---|---|---|---|
| Variable | Coeficiente | Desv. típica | Estadístico $t$ | valor p |
| const | 3,826 | 13,866 | 0,276 | 0,782 |
| $H$ | 2,468 | 2,078 | 1,187 | 0,236 |
| $B$ | 17,275 | 2,859 | 6,041 | 0,000 |
| $S$ | 0,297 | 0,046 | 6,359 | 0,000 |
| periferia | -17,029 | 10,960 | -1,553 | 0,121 |
| periferia$\times S$ | -0,114 | 0,051 | -2,230 | 0,026 |

Suma cuadrados residuos 141100 $R^2$ 0,686779

1. Escribe el modelo de regresión lineal general propuesto. ¿En qué se diferencia del modelo que has propuesto en el último apartado de la Segunda Parte? ¿Qué se pretende recoger con la inclusión del nuevo término (periferia $\times S$)?

2. Si la superficie de la casa aumentara en un metro cuadrado, manteniéndose fijas el resto de las características, ¿en cuánto estimarías la variación del precio medio de venta de las viviendas?, ¿y si la casa está en la zona centro?, ¿y si la casa está en la zona oeste?

3. Contrasta la significativad de la variable localización en el Modelo P9.3.

4. Con la información de que dispones, ¿crees que un metro cuadrado adicional cuesta, ceteris paribus, lo mismo en la zona centro que en la periferia?

5. De acuerdo a tu respuesta en el apartado anterior, ¿qué modelo elegirías para determinar el precio de la vivienda?, ¿por qué? ¿Cuáles son los problemas que presentan los modelos que no has elegido? Razona tu respuesta y explica cuáles son las propiedades de los estimadores en los modelos que no has elegido.

# PRÁCTICA P10.

El dueño de un restaurante italiano que sirve a domicilio desea analizar el consumo anual de pizza de los residentes de su área ($C$, en euros) en función de las siguientes variables explicativas:

$R$: renta anual en miles de euros.

$E$: edad en años por encima de la mayoría de edad (18 años).

Género: hombre o mujer.

Nivel máximo de estudios: sin estudios, estudios primarios, estudios secundarios y estudios universitarios.

Se ha realizado una encuesta a 40 de los residentes en la zona, algunos de cuyos resultados se recogen en el siguiente cuadro.

| Obs. | $C_i$ | $R_i$ | $E_i$ | Género | Nivel de Estudios |
|---|---|---|---|---|---|
| 1 | 109 | 15 | 27 | Mujer | Sin Estudios |
| 2 | 0 | 30 | 2 | Mujer | Sin Estudios |
| 3 | 0 | 12 | 10 | Mujer | Sin Estudios |
| 4 | 108 | 20 | 7 | Mujer | Sin Estudios |
| 5 | 220 | 15 | 17 | Mujer | Estudios Primarios |
| 6 | 189 | 30 | 22 | Mujer | Estudios Primarios |
| 7 | 64 | 12 | 4 | Mujer | Estudios Primarios |
| 8 | 262 | 12 | 12 | Mujer | Estudios Primarios |
| 9 | 64 | 28 | 3 | Mujer | Estudios Primarios |
| 10 | 35 | 22 | 22 | Mujer | Estudios Secundarios |
| 11 | 94 | 44 | 3 | Mujer | Estudios Secundarios |
| 12 | 71 | 10 | 27 | Mujer | Estudios Secundarios |
| 13 | 403 | 22 | 18 | Mujer | Estudios Universitarios |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |

## PRIMERA PARTE.

El experto contratado por el dueño para realizar el estudio comienza estimando un modelo de regresión lineal simple del consumo en función de la renta:

$$C_i = \beta_1 + \beta_2 \ R_i + u_i \qquad i = 1, \ldots, 40. \tag{P10.1}$$

Empleando la siguiente información muestral,

$$\sum_{i=1}^{40} C_i = 7662 \quad \sum_{i=1}^{40} R_i = 17170 \quad \sum_{i=1}^{40} (C_i - \bar{C})^2 = 947651,9$$
$$\sum_{i=1}^{40} (R_i - \bar{R})^2 = 6041478 \quad \sum_{i=1}^{40} C_i R_i = 4169550 \quad \sum_{i=1}^{40} (C_i - \bar{C})(R_i - \bar{R}) = 880636,5$$

1. Estima los coeficientes del modelo (P10.1) y escribe la recta de regresión muestral.
2. Interpreta el coeficiente estimado que acompaña a la variable renta.
3. Calcula el coeficiente de correlación muestral entre el consumo y la renta. ¿Qué relación tiene con el coeficiente de determinación? Calcula este último e interprétalo.
4. ¿Qué hipótesis básicas del modelo de regresión lineal son necesarias para que el estimador MCO sea insesgado?
5. Considera el siguiente modelo de regresión lineal simple:

$$R_i = \alpha_1 + \alpha_2 \ C_i + v_i \qquad i = 1, \ldots, 40. \tag{P10.2}$$

   5.1. ¿Qué diferencia existe entre los Modelos P10.1 y P10.2?
   5.2. ¿Coincide la estimación de $\alpha_2$ con la de $\beta_2$? ¿Por qué?
   5.3. ¿Qué relación existe entre los coeficientes de determinación de los Modelos P10.1 y P10.2? ¿Por qué?

## SEGUNDA PARTE.

En una segunda etapa el experto decide utilizar toda la información para estimar un modelo para el consumo, incluyendo todas las variables explicativas de que dispone. La siguiente tabla presenta los resutados obtenidos en la estimación del modelo.

Modelo P10.3: estimaciones MCO utilizando las 40 observaciones 1–40
Variable dependiente: $C$

| Variable | Coeficiente | Desv. Típica | Estadístico $t$ | Valor p |
|---|---|---|---|---|
| const | 280,735 | 39,0137 | 7,196 | 0,0000 |
| $R$ | 0,283674 | 0,0428834 | 6,6150 | 0,0000 |
| $E$ | -8,8409 | 1,49434 | -5,9163 | 0,0000 |
| $M$ | -181,85 | 25,8175 | -7,0438 | 0,0000 |
| $EP$ | 73,1179 | 37,1903 | 1,9661 | 0,0578 |
| $ES$ | -6,4715 | 40,2046 | -0,1610 | 0,8731 |
| $EU$ | -48,752 | 59,1766 | -0,8238 | 0,4159 |

Suma cuadrados residuos 215476,3 $R^2$ 0,772621

donde:

$M$ toma el valor 1 cuando el cliente es una mujer y 0 si es un hombre;

$EP$ toma el valor 1 cuando el cliente tiene como máximo estudios primarios y 0 en otro caso;

$ES$ toma el valor 1 cuando el cliente tiene como máximo estudios secundarios y 0 en otro caso y

$EU$ toma el valor 1 cuando el cliente tiene como máximo estudios universitarios y 0 en otro caso.

1. Escribe el modelo de regresión lineal general propuesto (Modelo P10.3). Explica brevemente cómo ha sido especificado: ¿cuántas variables explicativas hay?, ¿cuáles son?
2. Escribe la función de regresión muestral e interpreta los coeficientes estimados.
3. Si se conociera la recta de regresión poblacional, ¿coindiría con la recta de regresión muestral?
4. Interpreta la suma de coeficientes estimados $(280,735 - 181,85)$.
5. ¿Cuál es el consumo estimado de una universitaria según el Modelo P10.3?
6. Contrasta la significatividad individual de la variable explicativa género.
7. Dada la muestra disponible, ¿es razonable concluir que las mujeres consumen 200 euros menos de pizza que los hombres, ceteris paribus?
8. Contrasta la significatividad individual de la variable explicativa nivel de estudios utilizando el fichero de datos practicaP10.gdt.

**TERCERA PARTE.**

El experto opina que, ante incrementos de la renta, la propensión marginal a consumir pizza es menor para las mujeres que para los hombres.

Hipótesis de trabajo:

> *El efecto de un incremento de la renta sobre el consumo de pizza depende del género.*

La siguiente tabla recoge los resultados de la estimación de un modelo de regresión lineal que contempla esta hipótesis de trabajo.

Modelo P10.4: estimaciones MCO utilizando las 40 observaciones 1–40
Variable dependiente: $C$

| Variable | Coeficiente | Desv. Típica | Estadístico $t$ | Valor p |
|---|---|---|---|---|
| const | 236,640 | 41,8363 | 5,656 | 0,0000 |
| $R$ | 0,395667 | 0,0645538 | 6,1293 | 0,0000 |
| $E$ | -8,9736 | 1,41315 | -6,3501 | 0,0000 |
| $M$ | -115,72 | 38,4135 | -3,0126 | 0,0050 |
| $EP$ | 71,7391 | 35,1439 | 2,0413 | 0,0495 |
| $ES$ | -11,153 | 38,0446 | -0,2932 | 0,7713 |
| $EU$ | -56,106 | 56,0092 | -1,0017 | 0,3240 |
| $R \times M$ | -0,152513 | 0,0684368 | -2,2285 | 0,0330 |

Suma cuadrados residuos 186527,9 $R^2$ 0,803168

1. Escribe el modelo de regresión propuesto. ¿Cumple las hipótesis básicas del MRLG?
2. Escribe las diez primeras filas de la matriz de datos $X$ para el modelo estimado.
3. Si en la muestra se dispusiera sólo de estas diez primeras observaciones, ¿qué resultados se obtendrían?
4. ¿Cuál es el efecto sobre el consumo medio de un incremento de 1000 euros en la renta, ceteris paribus? ¿Existe evidencia muestral de que sea constante?
5. ¿Cuál es el consumo estimado de una universitaria?
6. Contrasta si la variable explicativa renta es significativa utilizando el fichero de datos `practicaP10.gdt`.

### CUARTA PARTE.

Es razonable suponer que conforme un individuo tiene más edad su propensión marginal a consumir pizza es menor, es decir, va a dedicar menos parte de esa renta extra en aumentar su consumo de pizza.

Hipótesis de trabajo:

> *El efecto de un incremento de la renta sobre el consumo de pizza depende de la edad del individuo.*

1. Especifica un modelo que sea capaz de recoger este efecto interacción entre las variables renta y edad.
   - 1.1. ¿Cuál es el incremento esperado, ceteris paribus, en el consumo de pizza de un individuo si su renta anual aumenta en 500 euros? ¿Y si aumenta en 1000 euros?
   - 1.2. ¿Cuál es el incremento esperado, ceteris paribus, en el consumo de pizza de un individuo de 20 años si su renta anual aumenta en 1000 euros? ¿Y si tiene 40 años?
2. Estima el modelo que hayas propuesto por MCO utilizando el fichero de datos `practicaP10.gdt` y escribe la recta de regresión muestral.
   - 2.1. ¿Cuál es el incremento estimado, ceteris paribus, en el consumo de pizza de un individuo si su renta anual aumenta en 500 euros? ¿Y si aumenta en 1000 euros?
   - 2.2. ¿Cuál es el incremento estimado, ceteris paribus, en el consumo de pizza de un individuo de 20 años si su renta anual aumenta en 1000 euros? ¿Y si tiene 40 años?
3. Contrasta la hipótesis de trabajo planteada.
4. Contrasta la significatividad de la variable explicativa renta.

### QUINTA PARTE.

Es razonable suponer que conforme aumenta el nivel de renta del individuo su propensión marginal a consumir pizza disminuye.

Hipótesis de trabajo:

> *El efecto de un incremento de la renta sobre el consumo de pizza depende del nivel de renta del individuo.*

1. Especifica un modelo que sea capaz de recoger el efecto económico que plantea la hipótesis de trabajo.

2. Según el modelo especificado en el apartado anterior, ¿cuál es el incremento esperado, ceteris paribus, en el consumo de pizza de un individuo si su renta aumenta en 1000 euros? ¿Es constante?

3. Estima el modelo que hayas propuesto por MCO utilizando el fichero de datos `practicaP10.gdt` y escribe la recta de regresión muestral.

4. Contrasta la hipótesis de trabajo planteada.

5. Contrasta la significatividad de la variable explicativa renta.

# PRÁCTICA P11.

El Ministerio de Sanidad quiere analizar los determinantes del consumo de tabaco en Turquía con datos anuales desde 1960 hasta 1988 (ambos incluidos) sobre las siguientes variables:

$C$: consumo de cigarrillos por adulto en kilos.

$R$: renta real per capita en liras turcas.

$P$: precio real de los cigarrillos en liras turcas por kilo.

## PRIMERA PARTE.

Se formula el siguiente modelo de regresión lineal en el que el consumo de cigarrillos depende de su precio y de la renta per capita:

$$\log C_t = \beta_1 + \beta_2 \log P_t + \beta_3 \log R_t + u_t \quad u_t \sim NID(0, \sigma_u^2) \quad t = 1960, \ldots, 1988. \quad \text{(P11.1)}$$

1. ¿Cumple el Podelo P11.1 las hipótesis básicas del modelo de regresión lineal general?

2. Escribe la recta de regresión poblacional e interpreta los coeficientes que acompañan a las variables explicativas.

3. Escribe la matrices $X, X'X$ y $X'Y$ en función de la variable explicada y las variables explicativas del Modelo P11.1.

4. La estimación MCO del Modelo P11.1 con la muestra disponible proporciona los siguientes resultados ( $l\_X = \log X$ ):

Modelo P11.1: estimaciones MCO, usando las 29 observaciones 1960–1988
Variable dependiente: $l\_C$

| | Coeficiente | Desv. Típica | Estadístico $t$ | Valor p |
|---|---|---|---|---|
| const | −4,58987 | 0,724913 | −6,3316 | 0,0000 |
| $l\_P$ | −0,485683 | 0,101394 | −4,7900 | 0,0001 |
| $l\_R$ | 0,688498 | 0,0947276 | 7,2682 | 0,0000 |

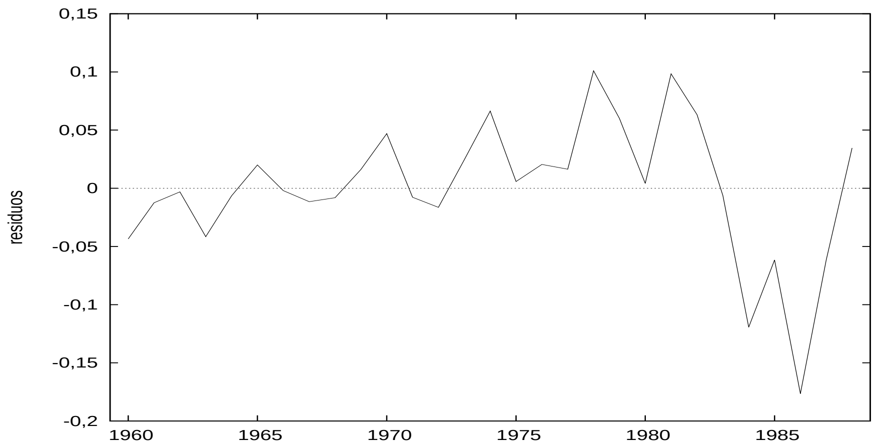
Suma cuadrados residuos 0,094911 $R^2$ 0,712058

Escribe la recta de regresión muestral.

5. ¿Existe evidencia en la muestra de que el precio influye en el consumo medio de tabaco?

6. ¿Qué hipótesis básicas del modelo de regresión lineal general son necesarias para que el contraste realizado en el apartado anterior sea válido?

7. El siguiente gráfico muestra los residuos de la estimación del Modelo P11.1. Comenta los resultados obtenidos. ¿A qué pueden ser debidos?




## SEGUNDA PARTE.

A partir de 1982, en Turquía se comenzaron a hacer públicos los anuncios del departamento de Sanidad avisando sobre los peligros de fumar para la salud. Además, en 1986 uno de los principales periódicos de tirada nacional lanzó una fuerte campaña antitabaco.

Se estima un modelo que quiere tener en cuenta la influencia de estas dos políticas antitabaco en el consumo de cigarrillos con los siguientes resultados:

Modelo P11.2: estimaciones MCO utilizando las 29 observaciones 1960–1988
Variable dependiente: $l\_C$

| Variable | Coeficiente | Desv. Típica | Estadístico $t$ | Valor p |
|---|---|---|---|---|
| const | −4,22729 | 0,515213 | −8,2049 | 0,0000 |
| $l\_P$ | −0,213012 | 0,0863821 | −2,4659 | 0,0212 |
| $l\_R$ | 0,626370 | 0,0678141 | 9,2366 | 0,0000 |
| $D82$ | −0,101307 | 0,0258239 | −3,9230 | 0,0006 |
| $D86$ | −0,100729 | 0,0360802 | −2,7918 | 0,0101 |

Suma cuadrados residuos 0,040520 $R^2$ 0,877069

donde $D82$ toma el valor 1 para las observaciones de 1982 en adelante y 0 para el resto, y $D86$ toma el valor 1 para las observaciones de 1986 en adelante y 0 para el resto.

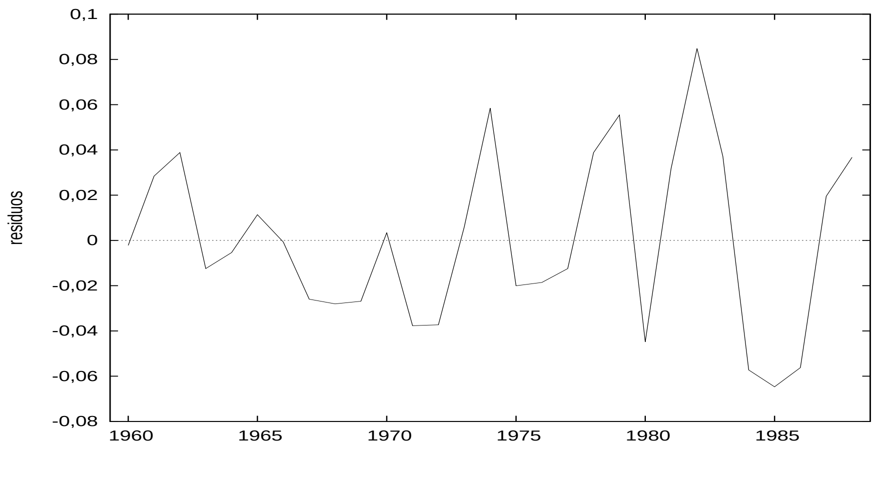
1. Escribe el modelo de regresión que corresponde al Modelo P11.2. ¿En qué se diferencia del Modelo P11.1? Explícalo detalladamente.

2. ¿Cuál es el consumo promedio de cigarrillos en el periodo 1960-1981? ¿Y en el periodo 1982-1985? ¿Y en el periodo 1986-1988?

3. Interpreta los coeficientes estimados de las variables $D82$ y $D86$.

4. ¿Ha sido efectiva cada una de las políticas antitabaco implantada en Turquía?

5. El siguiente gráfico muestra los residuos de la estimación del Modelo P11.2. Comenta los resultados y comparálos con los obtenidos en la estimación del Modelo P11.1.




6. Vistos los resultados obtenidos hasta ahora, ¿qué modelo utilizarías para la determinación del consumo de tabaco? ¿Por qué?

**TERCERA PARTE.**

Un investigador piensa que el efecto de las campañas publicitarias sobre el consumo de tabaco es más complejo que el reflejado en el Modelo P11.2. Concretamente, sospecha que existe cambio estructural, es decir, que las elasticidades renta y precio son diferentes en cada uno de los periodos de la muestra:

- Sin campaña: 1960-1981.
- Campaña de Sanidad: 1982-1985.
- Campaña de Sanidad y del periódico: 1986-1988.

La siguiente tabla muestra los resultados de la estimación de un modelo de regresión lineal que contempla la hipótesis de trabajo del investigador.

Modelo P11.3: estimaciones MCO utilizando las 29 observaciones 1960–1988
Variable dependiente: $l\_C$

| Variable | Coeficiente | Desv. Típica | Estadístico $t$ | Valor p |
|---|---|---|---|---|
| const | −5,02489 | 0,462283 | −10,8697 | 0,0000 |
| $l\_P$ | −0,381857 | 0,0882174 | −4,3286 | 0,0003 |
| $l\_R$ | 0,735837 | 0,0620130 | 11,8659 | 0,0000 |
| $D82$ | 23,3924 | 5,23205 | 4,4710 | 0,0002 |
| $D86$ | −36,2588 | 12,1341 | −2,9882 | 0,0073 |
| $l\_P \times D82$ | 0,416367 | 0,193047 | 2,1568 | 0,0434 |
| $l\_P \times D86$ | −0,236248 | 0,243390 | −0,9707 | 0,3433 |
| $l\_R \times D82$ | −2,80248 | 0,623550 | −4,4944 | 0,0002 |
| $l\_R \times D86$ | 4,25135 | 1,43051 | 2,9719 | 0,0075 |

Suma cuadrados residuos 0,016601 $R^2$ 0,949634

1. Escribe el modelo de regresión lineal general y la función de regresión muestral correspondientes al Modelo P11.3. Explica detalladamente la especificación de este modelo.
2. ¿Cuál es el consumo promedio de cigarrillos en el periodo 1960-1981? ¿Y en el periodo 1982-1985? ¿Y en el periodo 1986-1988?
3. Estima la elasticidad renta para los distintos periodos de la muestra: sin campaña, con una campaña, con dos campañas. ¿Es constante en los tres periodos?
4. Interpreta los coeficientes -2,80248 y 4,25135.
5. ¿Qué modelo consideras más adecuado para representar la determinación del consumo de cigarrillos, el Modelo P11.3 o el Modelo P11.2? Razónalo.

# PRÁCTICA P12.

El gerente de una empresa dedicada a la producción y venta de txakoli con Eusko Label quiere analizar la evolución de sus ventas en función de algunas variables que cree relevantes como son los precios y los gastos en publicidad. Para ello recoge datos trimestrales desde el primer trimestre de 1989 hasta el último trimestre de 2010 para las siguientes variables:

$V$: número de botellas de txakoli vendidas por la empresa.

$P$: precio medio en euros de la botella de txakoli vendida por la empresa.

$C$: precio medio en euros de la botella de txakoli vendida por las empresas competidoras.

$GP$: gastos en publicidad medidos en euros.

### PRIMERA PARTE.

La siguiente tabla muestra los resultados de la estimación de un modelo de regresión lineal que determina las ventas de txakoli en función de las tres variables explicativas mencionadas anteriormente.

Modelo P12.1: estimaciones MCO utilizando las 88 observaciones 1989:1–2010:4
Variable dependiente: $V$

| Variable | Coeficiente | Desv. Típica | Estadístico $t$ | Valor p |
|---|---|---|---|---|
| const | 304,430 | 173,557 | 1,7541 | 0,0831 |
| $P$ | −1,41396 | 0,983204 | −1,4381 | 0,1541 |
| $C$ | 2,56275 | 0,798534 | 3,2093 | 0,0019 |
| $GP$ | 0,743241 | 0,0757153 | 9,8163 | 0,0000 |

Suma cuadrados residuos 96083,63 $R^2$ 0,994851

1. Escribe el modelo de regresión lineal general propuesto por el gerente.
2. Escribe la función de regresión muestral.
3. Interpreta los coeficientes estimados de las variables relacionadas con los precios, $P$ y $C$. ¿Tienen el signo esperado?
4. Escribe la expresión del coeficiente de determinación junto con su valor. Interprétalo.
5. El siguiente cuadro muestra las primeras observaciones de la muestra disponible.

| Año | $V_t$ | $P_t$ | $C_t$ | $GP_t$ |
|---|---|---|---|---|
| 1989:1 | 1800,5 | 44,7 | 24,5 | 1990,6 |
| 1989:2 | 1807,5 | 45,0 | 23,9 | 2020,1 |
| 1989:3 | 1824,7 | 44,5 | 23,3 | 2045,3 |
| 1989:4 | 1821,2 | 47,5 | 23,1 | 2045,2 |
| 1990:1 | 1849,9 | 40,9 | 23,8 | 2073,9 |
| 1990:2 | 1863,5 | 39,4 | 23,7 | 2098,0 |
| 1990:3 | 1876,9 | 36,9 | 23,8 | 2106,6 |
| 1990:4 | 1904,6 | 35,9 | 23,7 | 2121,1 |
| $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ |

¿Cuántos botellas de txakoli se vendieron en el segundo trimestre de 1990? Basándote en el modelo estimado, ¿en cuánto se estima dicha venta?, ¿en cuántos botellas nos hemos equivocado a la hora de estimarlo?, ¿cómo se llama esta diferencia?

6. Si de un trimestre al siguiente los gastos en publicidad aumentaran en 200 euros ¿cuál sería el efecto estimado en las ventas?, ¿y si las demás variables se mantuvieran constantes?

7. Si los gastos en publicidad disminuyen en 100 euros, manteniendo el resto de variables fijas, ¿cuál es la variación mínima estimada en las ventas?, ¿y la máxima?, ¿por qué?

8. Contrasta la significatividad individual de las variables explicativas $P$ y $C$ del modelo. ¿Te parecen lógicos los resultados obtenidos?

9. Contrasta la significatividad conjunta de las variables explicativas, ¿crees que hay algún tipo de problema muestral?

10. Comenta el siguiente gráfico de los residuos del Modelo P12.1. ¿Podría incumplirse alguna hipótesis básica del modelo de regresión lineal?

## SEGUNDA PARTE.

El gerente recuerda que el nivel de lluvias en el periodo desde 1999 hasta 2003 no fue el más adecuado para la uva con la que se elabora el txakoli. Con el objeto de analizar la influencia de la climatología sobre las ventas, construye la variable Clima que toma valor uno en el periodo 1999-2003 y cero en otro caso.

Los resultados que obtiene al incluir la nueva variable se proporcionan a continuación:

Modelo P12.2: estimaciones MCO utilizando las 88 observaciones 1989:1–2010:4
Variable dependiente: $V$

| Variable | Coeficiente | Desv. Típica | Estadístico $t$ | Valor p |
|---|---|---|---|---|
| const | 467,625 | 129,134 | 3,6213 | 0,0005 |
| $P$ | −2,88326 | 0,743793 | −3,8764 | 0,0002 |
| $C$ | 3,02811 | 0,590079 | 5,1317 | 0,0000 |
| $GP$ | 0,687606 | 0,0560921 | 12,2585 | 0,0000 |
| Clima | −55,9927 | 6,59108 | −8,4952 | 0,0000 |

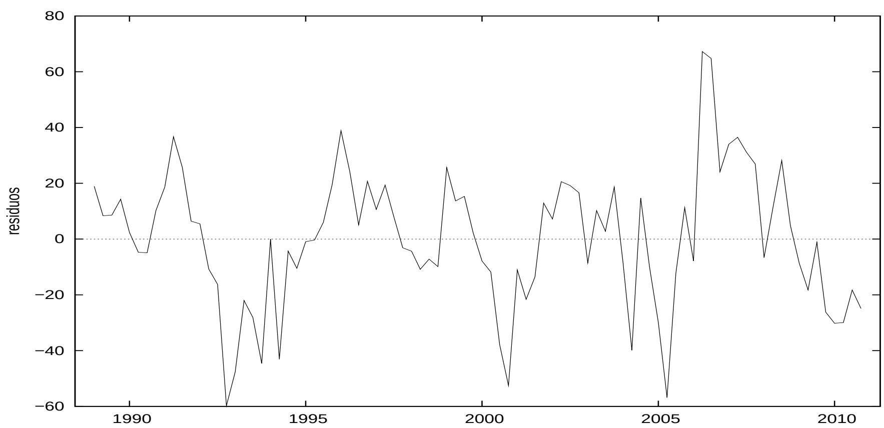
Suma cuadrados residuos 51395,28 $\quad R^2$ 0,997246

1. Escribe el modelo de regresión lineal general propuesto. ¿Cuál es su diferencia con respecto al Modelo P12.1?

2. Interpreta el coeficiente estimado que acompaña a la variable Clima. ¿Tiene el signo esperado?

3. ¿Existe evidencia muestral de que la mala climatología haya disminuido las ventas?

4. Comenta el gráfico de los residuos del Modelo P12.2 y compáralo con el gráfico de los residuos del Modelo P12.1.




5. ¿Qué puedes decir sobre la significatividad individual de las variables $P$ y $C$? Compara este resultado con el obtenido en el apartado 8 de la Primera Parte.

¿A qué puede ser debido este cambio?, ¿qué implicaciones tiene sobre las propiedades del estimador empleado en dicho apartado?

## TERCERA PARTE.

Debido a que los datos disponibles son trimestrales, el gerente quiere analizar si sus ventas presentan un comportamiento estacional. Los resultados de la estimación de un modelo que incluye el efecto estacional son los siguientes:

Modelo P12.3: estimaciones MCO utilizando las 88 observaciones 1989:1–2010:4
Variable dependiente: $V$

| Variable | Coeficiente | Desv. Típica | Estadístico $t$ | Valor p |
|---|---|---|---|---|
| const | 458,366 | 131,504 | 3,4856 | 0,0008 |
| $P$ | −2,84476 | 0,755292 | −3,7664 | 0,0003 |
| $C$ | 2,99899 | 0,599282 | 5,0043 | 0,0000 |
| $GP$ | 0,690514 | 0,0569771 | 12,1192 | 0,0000 |
| Clima | −55,9285 | 6,68217 | −8,3698 | 0,0000 |
| $dq1$ | 1,93326 | 7,61884 | 0,2537 | 0,8003 |
| $dq2$ | 1,40220 | 7,61273 | 0,1842 | 0,8543 |
| $dq3$ | 6,30854 | 7,62133 | 0,8277 | 0,4103 |

Suma cuadrados residuos 50907,59 $R^2$ 0,997272

donde $dqj$, $j = 1, 2, 3$, toma valor 1 si la observación pertenece al trimestre j-ésimo y cero en otro caso.

1. Escribe el modelo de regresión lineal que corresponde al Modelo P12.3. Explica cómo se ha incluido en este modelo el efecto estacional.

2. Escribe la recta de regresión muestral.

3. Interpreta los coeficientes estimados: 458,366 y 1,4022.

4. ¿En cuánto se estiman las ventas correspondientes al primer trimestre de 2003?, ¿y las del primer trimestre de 2004?

5. ¿Existe evidencia muestral de que las ventas de txakoli son estacionales?

6. Vistos los resultados obtenidos a lo largo del ejercicio, ¿qué modelo te parece más adecuado para explicar la evolución de las ventas de txakoli? ¿Por qué?

# PRÁCTICA P13.

Se quiere analizar el salario de los trabajadores de una empresa con sucursales en los tres territorios históricos para lo que se dispone de una muestra que recoge la información de 49 empleados sobre las siguientes variables:

$S$: salario del empleado en euros.

$E$: años de educación adicionales a los estudios básicos.

$X$: años de experiencia en el mercado laboral.

$H$: 1 si el empleado es hombre y 0 si es mujer.

$A$: 1 si el empleado trabaja en Alava y 0 en otro caso.

$B$: 1 si el empleado trabaja en Bizkaia y 0 en otro caso.

$G$: 1 si el empleado trabaja en Gipuzkoa y 0 en otro caso.

El siguiente cuadro recoge una parte de la muestra disponible.

| Obs. | 1 | 2 | 3 | 4 | 5 | ...... | 49 |
|---|---|---|---|---|---|---|---|
| $S_i$ | 1345 | 2435 | 1715 | 1461 | 1639 | ...... | 1288 |
| $H_i$ | 0 | 1 | 1 | 1 | 1 | ...... | 0 |
| $E_i$ | 6 | 4 | 6 | 6 | 9 | ...... | 6 |
| $X_i$ | 2 | 18 | 4 | 4 | 3 | ...... | 4 |
| $A_i$ | 0 | 0 | 1 | 0 | 0 | ...... | 0 |
| $B_i$ | 1 | 1 | 0 | 1 | 1 | ...... | 1 |

### PRIMERA PARTE.

Considera el siguiente modelo de regresión simple:

$$S_i = \beta_1 + \beta_2 \ E_i + u_i \qquad u_i \sim NID(0, \sigma_u^2) \qquad i = 1, 2, \ldots, 49. \tag{P13.1}$$

1. Empleando la siguiente información muestral, estima por MCO los coeficientes del modelo (P13.1).

$$\sum_{i=1}^{49} S_i = 89190 \quad \sum_{i=1}^{49}(S_i - \bar{S})^2 = 20172111,96 \quad \sum_{i=1}^{49}(E_i - \bar{E})^2 = 270,5306$$

$$\sum_{i=1}^{49} E_i = 305 \quad \sum_{i=1}^{49} S_iE_i = 585584 \quad \sum_{i=1}^{49}(S_i - \bar{S})(E_i - \bar{E}) = 30421,7551$$

2. Interpreta el coeficiente estimado que acompaña a la variable años de educación adicionales.

3. Calcula e interpreta el coeficiente de determinación.

4. ¿Es la variable años de educación adicionales significativa?

5. Si todos los individuos de la muestra tuvieran estudios básicos, ¿habría algún problema en la estimación del modelo?

### SEGUNDA PARTE.

Basándote en los siguientes resultados obtenidos en la estimación de un modelo de regresión para determinar el salario en base a los años de educación, los años de experiencia y el género:

Modelo P13.2: estimaciones MCO utilizando las 49 observaciones 1–49
Variable dependiente: $S$

| Variable | Coeficiente | Desv. Típica | Estadístico $t$ | Valor p |
|---|---|---|---|---|
| const | 434,821 | 258,871 | 1,6797 | 0,0999 |
| $E$ | 133,551 | 31,5081 | 4,2386 | 0,0001 |
| $X$ | 34,4543 | 12,1316 | 2,8401 | 0,0067 |
| $H$ | 470,460 | 144,870 | 3,2475 | 0,0022 |

Suma cuadrados residuos 11089355 $R^2$ 0,450263

1. Escribe el modelo de regresión que corresponde al Modelo P13.2. ¿En qué se diferencia del Modelo P13.1?

2. Interpreta los coeficientes estimados del Modelo P13.2.

3. Contrasta la significatividad individual y conjunta de las variables explicativas. ¿Crees que puede existir algún problema muestral?

4. ¿Hay evidencia en la muestra de que los hombres ganen en media más que las mujeres?

5. ¿Cuál es el salario estimado de una empleada con 5 años de educación adicional pero sin experiencia? ¿Y el de una empleada con estudios básicos y con 5 años de experiencia?

6. Si la experiencia de una empleada aumenta en un año, ¿en cuánto estimas que incrementa su salario medio si se mantienen constantes los años de educación? ¿Y si fuera un hombre?

## TERCERA PARTE.

Considera el modelo de regresión lineal cuya estimación MCO se recoge en la siguiente tabla.

Modelo P13.3: estimaciones MCO utilizando las 49 observaciones 1–49
Variable dependiente: $S$

| Variable | Coeficiente | Desv. Típica | Estadístico $t$ | Valor p |
|---|---|---|---|---|
| const | 1580,34 | 262,733 | 6,0150 | 0,0000 |
| $E$ | 60,8324 | 26,3964 | 2,3046 | 0,0261 |
| $X$ | 31,4564 | 9,01817 | 3,4881 | 0,0011 |
| $H$ | 594,659 | 118,738 | 5,0082 | 0,0000 |
| $A$ | -1151,2 | 194,498 | -5,9189 | 0,0000 |
| $G$ | -800,65 | 145,565 | -5,5003 | 0,0000 |

Suma cuadrados residuos 5689030 $R^2$ 0,717976

1. Escribe el modelo de regresión que corresponde al Modelo P13.3. Explica detalladamente en qué se diferencia del Modelo P13.2.
2. ¿Es la nueva variable explicativa relevante para determinar el salario medio?
3. ¿En cuánto estimas la diferencia salarial entre hombres y mujeres si el resto de características se mantienen iguales? ¿Es esta diferencia significativa?
4. ¿En cuánto estimas el salario medio de una empleada con estudios básicos y sin experiencia que trabaja en Bizkaia?, ¿y si trabaja en Araba?, ¿y si trabaja en Gipuzkoa?
5. Interpreta el valor (1580,34 + 594,659) que se obtiene de sumar las estimaciones correspondientes al primer y al quinto coeficientes.

## CUARTA PARTE.

La siguiente tabla muestra los resultados de la estimación de un modelo de regresión lineal general para los salarios que resulta de incluir una restricción lineal en el modelo Modelo P13.3.

Modelo P13.4: estimaciones MCO utilizando las 49 observaciones 1–49
Variable dependiente: $S$

| Variable | Coeficiente | Desv. Típica | Estadístico $t$ | Valor p |
|---|---|---|---|---|
| const | 1503,49 | 271,977 | 5,5280 | 0,0000 |
| $E$ | 81,9257 | 25,7337 | 3,1836 | 0,0027 |
| $X$ | 28,7986 | 9,33421 | 3,0853 | 0,0035 |
| $H$ | 475,994 | 110,845 | 4,2942 | 0,0001 |
| $A + G$ | -857,85 | 149,622 | -5,7334 | 0,0000 |

Suma cuadrados residuos 6347297 $R^2$ 0,685343

donde $A+G$ toma el valor 1 para los trabajadores de Alava y Gipuzkoa y el valor 0 para los trabajadores de Bizkaia.

1. ¿Cuál es la restricción en los coeficientes que relaciona el Modelo P13.3 y el Modelo P13.4? Interpreta su significado y contrasta si se cumple.

2. ¿En cuánto estimas el salario medio de una empleada con estudios básicos y sin experiencia que trabaja en Bizkaia?, ¿y si trabaja en Araba?, ¿y si trabaja en Gipuzkoa?

   Compara estos resultados con los obtenidos en el apartado 4. de la Tercera Parte. ¿A qué se debe la diferencia?

3. En base a todos los resultados obtenidos, ¿qué modelo elegirías para determinar el salario de los empleados de esta empresa? Razona tu respuesta en base a las propiedades del estimador utilizado.

# PRÁCTICA P14.

La función de producción es la relación que existe entre el producto obtenido y la combinación de factores que se utiliza en su obtención. Dado el estado de la tecnología en un momento dado de tiempo, la función de producción nos indica la cantidad de producto $Q$ que se puede obtener en función de las cantidades de los factores de producción que se utilizan $X_1, X_2, \ldots, X_n$, de modo que:

$$Q = f(X_1, X_2, \ldots, X_n)$$

Sea $f$ la función de producción Cobb-Douglas. Las propiedades técnicas de esta función de producción se establecen en torno al concepto de rendimientos a escala:

- Existen rendimientos a escala crecientes cuando al variar la cantidad utilizada de todos los factores en una determinada proporción, la cantidad obtenida del producto varía en una proporción mayor. Por ejemplo, si duplicamos la cantidad utilizada de todos los factores, se obtiene como resultado que la cantidad de producto se multiplica por un factor mayor que dos.

- Existen rendimientos a escala constantes cuando al variar la cantidad utilizada de todos los factores en una determinada proporción, la cantidad obtenida del producto varía en la misma proporción.

- Existen rendimientos a escala decrecientes cuando al variar la cantidad utilizada de todos los factores en una determinada proporción, la cantidad obtenida del producto varía en una proporción menor.

Sea un agricultor que se dedica al cultivo del maíz en una granja. Este empresario utilizará los siguientes factores de producción para producir $Q$ toneladas de maíz:

$K$: número de bienes de equipo duradero (máquinas) que utiliza.

$L$: número de trabajadores a tiempo completo.

$E$: miles de KW utilizados.

$H$: número de hectáreas cultivadas.

$F$: kilos de fertilizantes utilizados.

$S$: kilos de semillas empleados.

**PRIMERA PARTE.** Estimación de la función de producción.

Supongamos que en la función de producción de esta granja la relación entre el producto $Q$ y los factores de producción es doble logarítmica, es decir, tanto la variable endógena como las exógenas se miden en logaritmos.

1. Estima la función de producción con los datos desde el año 1948 al 1993. Justifica detalladamente la especificación de tu modelo, la selección de los regresores e interpreta los resultados de su estimación.

2. ¿Existe evidencia en la muestra de que la función de producción que has elegido tiene rendimientos a escala constantes?

**SEGUNDA PARTE.** Cambios tecnológicos.

La crisis del petróleo de principios de los años 70 trajo consigo cambios en la tecnología y, como es bien sabido, los cambios tecnológicos afectan a la función de producción. Supongamos que estos cambios en la tecnología se ponen en marcha a partir del año 1976.

1. ¿Existe evidencia en la muestra de que, para los mismos niveles de los factores de producción, el cambio tecnológico experimentado a partir de 1976 ha hecho que el nivel de producción sea mayor que antes de 1976? Justifica detalladamente la especificación del modelo utilizado y responde a la pregunta interpretando los resultados de su estimación.

2. ¿Existe evidencia de que la función de producción sea distinta antes y después de 1976, es decir, que el efecto sobre la cantidad producida de maíz de un incremento en los factores de producción sea diferente antes y después de 1976? Explica detalladamente la especificación del modelo utilizado y responde a la pregunta interpretando los resultados de su estimación.

**TERCERA PARTE.** Selección de modelos.

1. Dados los resultados obtenidos, ¿cuál es el modelo más adecuado para la función de producción? Justifica tu respuesta.
2. Con el modelo seleccionado en el apartado anterior, predice la cantidad de maíz producida por esta granja para el año 1994, sabiendo que cuenta con 59 máquinas y 79 trabajadores a tiempo completo, y que ha cultivado 87 hectáreas utilizando 92 kilos de fertilizante, 96 kilos de semillas y 92000 KW de energía.

# PRÁCTICA P15.

Se cuenta con 52 datos semanales sobre las ventas de la marca de latas de atún AL en un determinado hipermercado, así como su precio y el precio de las marcas de la competencia IS y CA.

**PRIMERA PARTE.** Modelo para las ventas de latas de atún AL.

1. Estima el modelo de regresión para la determinación de las ventas de la marca AL en función de su propio precio y el de los competidores. Interpreta los coeficientes estimados y la bondad de ajuste del modelo propuesto.
2. ¿Existe evidencia en la muestra de que el efecto de las políticas de precios de las dos marcas competidoras es el mismo sobre las ventas medias de la marca AL?
3. Dados los resultados obtenidos hasta ahora, especifica y estima un modelo adecuado para las ventas de atún. Justifica tu elección e interpreta los resultados obtenidos.
4. El gerente cree que si en la semana 53 mantiene el precio de su marca AL y los competidores lo suben en 0,15 euros con respecto de sus precios en la semana anterior, sus ventas medias van a aumentar en 50 unidades. ¿Hay evidencia empírica que apoye este supuesto?

**SEGUNDA PARTE.** Políticas de marketing.

El gerente de la empresa AL se ha planteado hacer algunas campañas publicitarias para fomentar las ventas. De hecho, ha probado la efectividad de dos tipos de campañas diferentes:

CAMPAÑA A. Montar exclusivamente un stand en el hipermercado que hace promoción de la marca (de la semana 30 a la 52).

CAMPAÑA B. Poner un anuncio en el periódico (semanas 1, 7, 10, 11, 12, 13, 17, 22 y 26).

1. Estima un modelo para las ventas que tenga en cuenta no sólo los precios sino también el posible efecto de las dos campañas publicitarias. Justifica la especificación del modelo e interpreta los resultados obtenidos.

2. Si el gerente sólo pudiera pagar un tipo de campaña publicitaria ¿qué política de marketing le aconsejarías?

**TERCERA PARTE.** Selección de modelos.

1. Dados los resultados obtenidos, ¿qué modelo escoges para determinar las ventas de la marca AL? Justifícalo razonadamente.

2. Basándote en el modelo escogido, ¿qué recomendaciones le darías al gerente de la empresa sobre política de precios y política de marketing?

# Parte IV

# Soluciones a las Prácticas

# Solución PRÁCTICA P1.

### PRIMERA PARTE.

1. Tamaño muestral: $N = 15$.

   Precio de la cuarta vivienda de la muestra: $P_4 = 306$ miles de dólares.

   Número de habitaciones de la cuarta vivienda de la muestra: $H_4 = 2$.

2. La variable explicada es el precio, $P$, y la variable explicativa es la superficie, $S$.

   Los elementos aletarios son la variable explicada, $P$, y la perturbación aleatoria, $u$.

3. Recta de regresión poblacional: $E(P_i) = \alpha + \beta\, S_i, \quad i = 1, \ldots, 15.$

4. $\beta =$ variación esperada en el precio de una vivienda, en miles de dólares, cuando la superficie aumenta en un pie cuadrado.

5. El criterio de estimación de Mínimos Cuadrados Ordinarios se basa en la minimización de la suma de los residuos al cuadrado, es decir, en minimizar

$$\sum_{i=1}^{15} \hat{u}_i^2 = \sum_{i=1}^{15} (P_i - \hat{\alpha} - \hat{\beta}\, S_i)^2.$$

   Las condiciones de primer orden resultan de obtener las derivadas parciales respecto de cada estimador e igualarlas a cero:

$$-2 \sum (P_i - \hat{\alpha} - \hat{\beta}\, S_i) = 0 \quad \Rightarrow \quad \sum (P_i - \hat{\alpha} - \hat{\beta}\, S_i) = 0$$

$$-2 \sum (P_i - \hat{\alpha} - \hat{\beta}\, S_i)\, S_i = 0 \quad \Rightarrow \quad \sum (P_i - \hat{\alpha} - \hat{\beta}\, S_i)\, S_i = 0$$

   Ecuaciones normales:

$$\begin{aligned} \sum P_i &= N\,\hat{\alpha} + \hat{\beta} \sum S_i \\ \sum P_i\, S_i &= \hat{\alpha} \sum S_i + \hat{\beta} \sum S_i^2 \end{aligned}$$

   Sustituyendo los valores muestrales:

$$\begin{aligned} 3600 &= 15 + \hat{\beta}\ 21909 \\ 5322245 &= \hat{\alpha}\ 21909 + \hat{\beta}\ 32780493 \end{aligned}$$

6. La estimación MCO de los coeficientes del Modelo P1.1 se puede llevar a cabo de varias formas:

   a) Resolviendo el sistema de ecuaciones normales del apartado anterior:

   $$\hat{\alpha} = 120,029 \qquad \hat{\beta} = 0,082.$$

   b) Mediante las fórmulas explícitas resultantes del sistema de ecuaciones normales:

   $$\hat{\beta} = \frac{\sum_{i=1}^{N}(S_i - \bar{S})(P_i - \bar{P})}{\sum_{i=1}^{N}(S_i - \bar{S})^2} = \frac{64085}{780207,6} = 0,082$$

   $$\hat{\alpha} = \bar{P} - \hat{\beta}\,\bar{S} = \frac{3600}{15} - 0,082 \times \frac{21909}{15} = 120,029$$

   c) Matricialmente: $\hat{\beta}_{MCO} = (X'X)^{-1}\,X'Y$

$$\begin{bmatrix} \hat{\alpha} \\ \hat{\beta} \end{bmatrix}_{MCO} = \begin{bmatrix} N & \sum_{i=1}^{15} S_i \\ \sum_{i=1}^{15} S_i & \sum_{i=1}^{15} S_i^2 \end{bmatrix}^{-1} \begin{bmatrix} \sum_{i=1}^{15} P_i \\ \sum_{i=1}^{15} P_i\,S_i \end{bmatrix} =$$

$$= \begin{bmatrix} 15 & 21909 \\ 21909 & 32780493 \end{bmatrix}^{-1} \begin{bmatrix} 3600 \\ 5322245 \end{bmatrix}$$

$$= \frac{1}{11703114} \begin{bmatrix} 32780493 & -21909 \\ -21909 & 15 \end{bmatrix} \begin{bmatrix} 3600 \\ 5322245 \end{bmatrix} = \begin{bmatrix} 120,029 \\ 0,082 \end{bmatrix}$$

   Recta de regresión muestral: $\widehat{P}_i = 120,029 + 0,082\,S_i \qquad i = 1,2,\ldots,15.$

7. Precio estimado: $\widehat{P}_1 = 120,029 + 0,082 \times 1185 = 217,199$ miles de dólares

   Residuo: $\hat{u}_1 = P_1 - \widehat{P}_1 = 219 - 217,199 = 1,801$ miles de dólares.

8. Intervalo de confianza.

$$IC(\beta)_{95\,\%} = \left[\hat{\beta} \pm t(N-k)_{\alpha/2}\ \hat{\sigma}_{\hat{\beta}}\right] =$$

$$= \left[0,082 \pm t(15-2)_{0,05/2}\ \sqrt{0,00566932}\right] = [-0,0806 \quad 0,2446]$$

dado que:

$$\hat{\sigma}^2_{\hat{\beta}} = \frac{\hat{\sigma}^2_u}{\sum_{i=1}^{15}(S_i - \bar{S})^2} = \frac{\sum \hat{u}_i^2/(15-2)}{\sum_{i=1}^{15}(S_i - \bar{S})^2} = \frac{57511,06/(15-2)}{780207,6} = 0,00567019$$

y

$$\sum \hat{u}_i^2 = \sum P_i^2 - \hat{\alpha} \sum P_i - \hat{\beta} \sum P_i S_i = 57511,06$$

Como $0 \in IC(\beta)_{95\,\%}$, se puede concluir con un 95 % de confianza que la variable explicativa superficie no es estadísticamente significativa.

## SEGUNDA PARTE.

1. El orden de la matriz de datos $X$ es $(15 \times 3)$ porque hay 15 observaciones en la muestra y dos variables explicativas más el término constante.

$$X = \begin{bmatrix} 1 & 1185 & 2 \\ 1 & 1421 & 2 \\ 1 & 1478 & 2 \\ 1 & 2205 & 2 \\ 1 & 1171 & 3 \\ \vdots & \vdots & \vdots \end{bmatrix}$$

2. Estimador MCO:

$$\hat{\beta}_{MCO} = \begin{bmatrix} \hat{\alpha} \\ \hat{\beta}_1 \\ \hat{\beta}_2 \end{bmatrix} = \begin{bmatrix} N & \sum S_i & \sum H_i \\ \sum S_i & \sum S_i^2 & \sum S_i H_i \\ \sum H_i & \sum S_i H_i & \sum H_i^2 \end{bmatrix}^{-1} \begin{bmatrix} \sum P_i \\ \sum P_i S_i \\ \sum P_i H_i \end{bmatrix}$$

3. $\hat{\alpha} =$ 238,808: precio estimado de una vivienda en miles de dólares cuando su superficie y el número de habitaciones son cero.

   $\hat{\beta}_1 =$ 0,06593: variación en el precio estimado de la vivienda, en miles de dólares, cuando la superficie aumenta en un pie cuadrado manteniendo la variable número de habitaciones fija.

   Este coeficiente estimado tiene el signo esperado ya que indica que, ceteris paribus, existe una relación directa entre la superficie de la vivienda y su precio: a mayor superficie mayor precio de la vivienda aunque el número de habitaciones no varie.

$\hat{\beta}_2 =$ -33,968: se estima que si el número de habitaciones aumenta en una unidad, manteniéndose la variable superficie fija, el precio disminuye en 33,968 miles de dólares.

El signo de esta estimación señala una relación inversa entre el precio de la vivienda y el número de habitaciones, lo que parece contraintuitivo. Sin embargo, hay que tener en cuenta que la interpretación de los coeficientes es ceteris paribus, es decir, se incrementa el número de habitaciones pero se manteniene la superficie constante.

4. $\widehat{P}_p = 238,808 + 0,06593 \times 1250 - 33,968 \times 3 = 219,3165$ miles de dólares

5. Si el piso cuenta con una habitación más, manteniendo el resto de los factores fijos, se estima que su precio disminuye en 33968 dólares.

   $\Delta\widehat{P} = -33,968 \times \Delta H = -33,968 \times 1 = -33,968$ miles de dólares $= \hat{\beta}_2$.

6. Si el piso cuenta con una habitación más y 120 pies cuadrados más, se estima que su precio disminuye en 26056 dólares.

$$\begin{aligned} \Delta\widehat{P} &= 0,06593 \times \Delta S - 33,968 \times \Delta H = \\ &= 0,06593 \times 120 - 33,968 \times 1 = -26,056 \text{ miles de dólares.} \end{aligned}$$

7. $R^2 = 1 - \dfrac{SCR}{SCT} = 1 - \dfrac{\sum \hat{u}_i^2}{\sum(P_i - \bar{P})^2} = 1 - \dfrac{52630,2}{62766,03} = 0,161486$

   Interpretación: el 16,1486 % de la variabilidad del precio de venta de las viviendas en la muestra viene explicada por la variabilidad de las variables explicativas: superficie y número de habitaciones.

8. $\hat{\sigma}_u^2 = \dfrac{\sum \hat{u}_i^2}{N-k} = \dfrac{52630,2}{15-3} = 4385,85.$

9. MRLG: $P_i = \alpha + \beta_1 S_i + \beta_2 H_i + u_i \qquad i = 1, \ldots, N.$

   Contraste de significatividad conjunta.

$$\begin{array}{l} H_0 : \beta_1 = \beta_2 = 0 \\ H_a : \beta_1 \neq 0 \text{ y/o } \beta_2 \neq 0 \end{array} \qquad F = \frac{R^2/q}{(1-R^2)/(N-k)} \overset{H_0}{\sim} \mathcal{F}(q, N-k)$$

$$F = \frac{0,161486/2}{(1-0,161486)/(15-3)} = 1,155513 < 3,88529 = \mathcal{F}(2,12)_{0,05}$$

No se rechaza la hipótesis nula con un nivel de significación del 5 %, es decir, se puede concluir que las variables explicativas superficie y número de habitaciones no son conjuntamente significativas.

10. Contrastes de significatividad individual.

$$\begin{array}{l} H_0 : \beta_i = 0 \\ H_a : \beta_i \neq 0 \end{array} \qquad t = \frac{\hat{\beta}_i - 0}{\hat{\sigma}_{\hat{\beta}_i}} \overset{H_0}{\sim} t(N-k) \qquad i = 1, 2.$$

$$\text{Superficie } (S) \qquad |t| = \left| \frac{0,06593 - 0}{0,07653} \right| = 0,861 \, < \, 2,17881 = t(15-3)_{0,05/2}$$

$$\text{N}^{\text{o}} \text{ habitaciones } (H) \qquad |t| = \left| \frac{-33,968 - 0}{32,228} \right| = 1,054 \, < \, 2,17881 = t(15-3)_{0,05/2}$$

Por lo tanto, se puede concluir que, para un nivel de significación del 5 %, las variables superficie y número de habitaciones no son individualmente significativas.

**TERCERA PARTE.**

1. $$\underset{(\hat{\sigma}_{\hat{\beta}})}{\widehat{P}_i} = \underset{(158068)}{238808} + \underset{(76,53)}{65,93}\, S_i - \underset{(32228)}{33968}\, H_i \qquad \sum_{i=1}^{N} \hat{u}_i^2 = 52630188426,9231.$$

2. $\hat{\alpha} =$ 238808 dólares: precio estimado de un piso cuando las variables superficie y número de habitaciones toman el valor cero.

   $\hat{\beta}_1 =$ 65,93 dólares: se estima que un pie cuadrado de superficie adicional, manteniéndose el número de habitaciones fijo, incrementa el precio de un piso en 65,93 dólares.

   $\hat{\beta}_2 =$ -33968 dólares: se estima que una habitación adicional, manteniéndose la superficie fija, disminuye el precio de un piso en 33968 dólares.

3. $\sum \hat{u}_i^2 = 52630188426,9231 \qquad \hat{\sigma}_u^2 = \dfrac{\sum \hat{u}_i^2}{N-k} = \dfrac{52630188426,9231}{15-3} = 4385850000.$

4. Matriz de covarianzas de los estimadores:

$$\widehat{V}(\hat{\beta}) = \hat{\sigma}_u^2\,(X'X)^{-1} = \begin{bmatrix} 24985500000 & -9943300 & -3632110000 \\ -9943300 & 5857,77 & 495514 \\ -3632110000 & 495514 & 1038700000 \end{bmatrix}$$

5. $R^2 = 1 - \dfrac{SCR}{SCT} = 1 - \dfrac{\sum \hat{u}_i^2}{\sum (P_i - \bar{P})^2} = 0,161486$

   Interpretación: el 16,1486 % de la variabilidad del precio de venta de las viviendas en la muestra viene explicada por la variabilidad de las variables explicativas: superficie y número de habitaciones.

6. Contraste de significatividad conjunta. Dado que el coeficiente de determinación es igual que en el caso de los precios medidos en miles de dólares, el resultado ha de ser el mismo: las variables explicativas no son conjuntamente significativas. Comprobémoslo.

$$\begin{array}{ll} H_0 : \beta_1 = \beta_2 = 0 & \\ H_a : \beta_1 \neq 0 \text{ y/o } \beta_2 \neq 0 & \end{array} \qquad F = \frac{R^2/q}{(1-R^2)/(N-k)} \overset{H_0}{\sim} \mathcal{F}(q, N-k)$$

$$F = \frac{0,161486/2}{(1-0,161486)/(15-3)} = 1,155513 \, < \, 3,88529 = \mathcal{F}(2,12)_{0,05}$$

No se rechaza la hipótesis nula con un nivel de significación del 5 %, es decir, se puede concluir que las variables explicativas superficie y número de habitaciones no son conjuntamente significativas.

Contraste de significatividad individual.

$$\begin{array}{l} H_0 : \beta_i = 0 \\ H_a : \beta_i \neq 0 \end{array} \qquad t = \frac{\hat{\beta}_i - 0}{\hat{\sigma}_{\hat{\beta}_i}} \overset{H_0}{\sim} t(N-k) \qquad i = 1, 2.$$

Superficie ($S$) $$|t| = \left| \frac{65,93 - 0}{76,53} \right| = 0,861 \, < \, 2,17881 = t(15-3)_{0,05/2}$$

N° habitaciones ($H$) $$|t| = \left| \frac{-33968 - 0}{32228} \right| = 1,054 \, < \, 2,17881 = t(15-3)_{0,05/2}$$

Por lo tanto, se puede concluir que, para un nivel de significación del 5 %, las variables superficie y número de habitaciones no son individualmente significativas.

7. Comparación de los resultados de la Segunda Parte y la Tercera Parte :

**Coeficientes.** Se puede observar que la estimación de los coeficientes es diferente. Cuando se miden los precios en dólares los coeficientes estimados se multiplican por 1000. Sin embargo, hay que señalar que la información que proporcionan es la misma. Por ejemplo,
$\hat{\beta}_1$ = 65,93 dólares: se estima que un pie cuadrado de superficie adicional, manteniéndose el número de habitaciones constante, incrementa el precio de un piso en 65,93 dólares = 0,06593 miles de dólares.

**Bondad de ajuste.** La suma de cuadrados de los residuos y la estimación de la varianza de las perturbaciones son distintas cuando la variable explicada se mide en dólares. Específicamente, están multiplicadas por $1000^2$.

Sin embargo, como es lógico, la bondad de ajuste, es decir, la proporción de la variabilidad del precio en la muestra explicada por la variabilidad de las variables explicativas, es la misma independientemente de las unidades en que se mida la variable explicada, en este caso el precio.

**Inferencia.** Cuando la variable explicada está medida en miles de dólares, se tiene que:

$$\begin{aligned}
\widehat{V}(\hat{\alpha}) &= (158,068)^2 = 24985,49262 \\
\widehat{V}(\hat{\beta}_1) &= (0,07653)^2 = 0,005856841 \\
\widehat{V}(\hat{\beta}_2) &= (32,228)^2 = 1038,643984
\end{aligned}$$

Por lo tanto, las varianzas (y también las covarianzas) se multiplican por $1000^2$ cuando la variable explicada pasa a medirse en dólares.

Ahora bien, la significatividad, tanto individual como conjunta, de las variables explicativas es independiente de las unidades en las que se mida la variable explicada.

## CUARTA PARTE.

1. $\underset{(\hat{\sigma}_{\hat{\beta}})}{\widehat{P}_i} = \underset{(158,068)}{238,808} + \underset{(0,8238)}{0,7097}\, S_i - \underset{(32,228)}{33,968}\, H_i \qquad \sum_{i=1}^{N} \hat{u}_i^2 = 52630,2.$

2. $\hat{\alpha} =$ 238,808 miles de dólares: precio estimado de un piso cuando la superficie y el número de habitaciones toman el valor cero.

   $\hat{\beta}_1 =$ 0,7097 miles de dólares: se estima que un metro cuadrado de superficie adicional, manteniéndose el número de habitaciones fijo, incrementa el precio de un piso en 709,7 dólares.

   $\hat{\beta}_2 =$ -33,968 miles de dólares: se estima que una habitación adicional, manteniéndose la superficie fija, disminuye el precio de un piso en 33968 dólares.

3. $\sum \hat{u}_i^2 = 52630,2 \qquad \hat{\sigma}_u^2 = \dfrac{\sum \hat{u}_i^2}{N-k} = \dfrac{52630,2}{15-3} = 4385,849.$

4. Matriz de covarianzas de los estimadores:

$$\widehat{V}(\hat{\beta}) = \hat{\sigma}_u^2\,(X'X)^{-1} = \begin{bmatrix} 24985,5 & -107,029 & -3632,11 \\ -107,029 & 0,678692 & 5,33367 \\ -3632,11 & 5,33367 & 1038,7 \end{bmatrix}$$

5. $R^2 = 1 - \dfrac{SCR}{SCT} = 1 - \dfrac{\sum \hat{u}_i^2}{\sum (P_i - \bar{P})^2} = 1 - \dfrac{52630,2}{62766,03} = 0,161486$

   Interpretación: el 16,1486 % de la variabilidad del precio de venta de las viviendas en la muestra viene explicada por la variabilidad de las variables explicativas: superficie y número de habitaciones.

6. Contraste de significatividad conjunta. Dado que el coeficiente de determinación es igual que en el caso de la superficie medida en pies cuadrados, el resultado ha de ser el mismo: las variables explicativas no son conjuntamente significativas.

   Contraste de significatividad individual.

$$\begin{array}{l} H_0 : \beta_i = 0 \\ H_a : \beta_i \neq 0 \end{array} \qquad t = \frac{\hat{\beta}_i - 0}{\hat{\sigma}_{\hat{\beta}_i}} \overset{H_0}{\sim} t(N-k) \qquad i = 1, 2.$$

$$\text{Superficie } (S) \qquad |t| = \left| \frac{0,7097 - 0}{0,8238} \right| = 0,861 < 2,17881 = t(15-3)_{0,05/2}$$

$$\text{N}^{\text{o}} \text{ habitaciones } (H) \qquad |t| = \left| \frac{-33,968 - 0}{32,228} \right| = 1,054 < 2,17881 = t(15-3)_{0,05/2}$$

   Por lo tanto, se puede concluir que, para un nivel de significación del 5 %, las variables superficie y número de habitaciones no son individualmente significativas.

7. Comparación de los resultados de la Segunda Parte y la Cuarta Parte:

   En la Parte Cuarta del ejericio se han cambiado las unidades de media en una de las variables explicativas, la superficie.

   **Coeficientes.** Se puede observar que sólo cambia la estimación del coeficente que acompaña a la variable superficie, los demás permanecen iguales. Notése, sin embargo, que la interpretación de este coeficiente es la misma, pero en términos de metros cuadrados.

   Como 1 *pie*$^2$ = 0,09290304 $m^2$, el coeficiente se ha dividido por 0,09290304:

$$\frac{0,06593}{0,09290304} = 0,7097$$

   **Bondad de ajuste.** La suma de cuadrados de residuos y la estimación de la varianza de las perturbaciones es la misma porque no cambian las unidades en las que está medida la variable explicada.

   En consecuencia y, como es lógico, la bondad de ajuste, es decir, la proporción de la variabilidad del precio en la muestra explicada por la variabilidad de las variables explicativas, es la misma independientemente de las unidades en que se mida las variables explicativas, en este caso la superficie.

**Inferencia.** En la matriz de covarianzas solo se modifican la varianza del estimador que acompaña a la variable superficie y las covarianzas de este estimador con los otros dos.

La significatividad, tanto individual como conjunta, de las variables explicativas es indenpendiente de las unidades en las que se midan las variables explicativas.

# Solución PRÁCTICA P2.

## PRIMERA PARTE.

1. $\hat{\beta}_{MCO} = (X'X)^{-1} X'Y$

$$\begin{bmatrix} \widehat{\beta}_1 \\ \widehat{\beta}_2 \end{bmatrix}_{MCO} = \begin{bmatrix} N & \sum_{i=1}^{718} ED_i \\ \sum_{i=1}^{718} ED_i & \sum_{i=1}^{718} ED_i^2 \end{bmatrix}^{-1} \begin{bmatrix} \sum_{i=1}^{718} S_i \\ \sum_{i=1}^{718} S_i \, ED_i \end{bmatrix} =$$

$$= \begin{bmatrix} 718 & 9814 \\ 9814 & 137732 \end{bmatrix}^{-1} \begin{bmatrix} 1204591 \\ 11821856074 \end{bmatrix} =$$

$$= \frac{1}{2576980} \begin{bmatrix} 137732 & -9814 \\ -9814 & 718 \end{bmatrix} \begin{bmatrix} 1204591 \\ 11821856074 \end{bmatrix} = \begin{bmatrix} 873,519 \\ 58,8348 \end{bmatrix}$$

Función de regresión muestral: $\widehat{S}_i = 873,519 + 58,8348 \, ED_i \quad i = 1, 2, \ldots, 718.$

2. $\hat{\beta}_2 = 58{,}8348$ euros: variación estimada en el salario del individuo por cada año de educación adicional.

3. Para que el estimador MCO sea insesgado y de varianza mínima, han de cumplirse las siguientes hipótesis básicas sobre la perturbación aleatoria:

   - Media cero: $E(u_i) = 0, \ \forall i.$
   - Homocedasticidad: $V(u_i) = \sigma_u^2, \ \forall i.$

- Ausencia de autocorrelación: $E(u_i\, u_j) = 0$, $\forall i \neq j$.

Para poder hacer inferencia, además de las tres hipótesis básicas anteriores, la perturbación aleatoria ha de seguir una distribución normal.

4. Con las seis últimas observaciones porque, según los datos que conocemos de la muestra, en las seis primeras observaciones la variable educación, $ED$, toma el mismo valor para todos los individuos. Por lo tanto, en esta submuestra no se cumple el supuesto de ausencia de colinealidad perfecta con lo que el rango de la matriz de datos $X$ no es completo y no se pueden estimar los coeficientes por MCO. Sin embargo, la submuestra formada por los últimos seis valores de la muestra no presenta este problema.

5. Función de regresión muestral:

$$\widehat{S}_i \;=\; 0,873519 \;+\; 0,0588348\, ED_i \qquad i = 1, 2, \ldots, 718.$$

$\hat{\beta}_2 = 0{,}0588348$ miles de euros. Se estima que si la educación aumenta en un año, el salario del individuo aumenta en 0,0588348 miles de euros = 58,8348 euros.

### SEGUNDA PARTE.

1. MRLG:

$$S_i = \beta_1 + \beta_2\, ED_i + \beta_3\, FED_i + \beta_4\, MED_i + u_i \qquad u_i \sim NID(0, \sigma_u^2) \quad i = 1, 2, \ldots, 718.$$

2. FRM:

$$\underset{(\hat{\sigma}_{\hat{\beta}})}{\widehat{S}_i} \;=\; \underset{(92{,}6237)}{803,84} \;+\; \underset{(7{,}2261)}{46,6073}\, ED_i + \underset{(5{,}5797)}{11,1165}\, FED_i + \underset{(6{,}4935)}{11,2765}\, MED_i$$

$$R^2 = 0,120364 \qquad\qquad SCR = 106000000$$

3. $\hat{\beta}_3$ = 11,1165 euros. Se estima que por cada año de educación adicional del padre, el salario del individuo aumenta en 11,1165 euros, manteniéndose constantes las variables educación del individuo y educación de la madre del individuo.

   $\hat{\beta}_3$ = 11,2765 euros. Se estima que por cada año de educación adicional de la madre, el salario del individuo aumenta en 11,2765 euros, manteniéndose constantes las variables educación del individuo y educación del padre del individuo.

4. $R^2 = 1 - \dfrac{SCR}{SCT} = 1 - \dfrac{\sum \hat{u}_i^2}{\sum (S_i - \bar{S})^2} = 0,120364$

Interpretación: el 12,0364 % de la variabilidad del salario del individuo en la muestra viene explicada por la variabilidad de las variables explicativas: educación del individuo, educación de la padre y educación de la madre.

5. Según el Modelo P2.2:

$$i = 1 \qquad ED = 12 \qquad FED = 8 \qquad MED = 8$$

$$\begin{aligned} \widehat{S}_1 &= 803,84 + 46,6073\, ED_1 + 11,1165\, FED_1 + 11,2765\, MED_1 = \\ &= 803,84 + 46,6073 \times 12 + 11,1165 \times 8 + 11,2765 \times 8 = 1542,3 \,\text{euros} \end{aligned}$$

$$i = 718 \qquad ED = 13 \qquad FED = 6 \qquad MED = 8$$

$$\begin{aligned} \widehat{S}_{718} &= 803,84 + 46,6073\, ED_{718} + 11,1165\, FED_{718} + 11,2765\, MED_{718} = \\ &= 803,84 + 46,6073 \times 13 + 11,1165 \times 6 + 11,2765 \times 8 = 1566,6 \,\text{euros} \end{aligned}$$

6. Contraste de hipótesis.

$$\begin{array}{l} H_0 : \beta_2 = 50 \\ H_a : \beta_2 \neq 50 \end{array} \qquad t = \frac{\hat{\beta}_2 - 50}{\hat{\sigma}_{\hat{\beta}_2}} \overset{H_0}{\sim} t(N-k)$$

$$|t| = \left| \frac{46,6073 - 50}{7,2261} \right| = 0,46951 < 1,96329 = t(718-4)_{0,05/2}$$

No se rechaza la hipótesis nula con un nivel de significación del 5 % por lo que se puede concluir que, manteniendo el resto de las variables constantes, por cada año de educación adicional del individuo su salario aumenta en 50 euros.

7. Intervalo de confianza.

$$\begin{aligned} IC(\beta_2)_{95\,\%} &= \left[\hat{\beta}_2 \pm t(N-k)_{\alpha/2}\, \hat{\sigma}_{\hat{\beta}_2}\right] = \\ &= \left[46,6073 \pm t(718-4)_{0,05/2} \times 7,2261\right] = [32,4204 \quad 60,7943] \end{aligned}$$

Dado que $50 \in IC(95\,\%)$, llegamos a la misma conclusión que en el apartado anterior: con un nivel de confianza del 95 % por cada año de educación adicional del individuo, su salario podría aumentar en 50 euros.

8. Contraste de significatividad individual.

$$\begin{array}{l} H_0 : \beta_i = 0 \\ H_a : \beta_i \neq 0 \end{array} \qquad t = \frac{\hat{\beta}_i - 0}{\hat{\sigma}_{\hat{\beta}_i}} \overset{H_0}{\sim} t(N-k) \qquad i = 3, 4.$$

Educación padre $(FED)$ $\quad |t| = \left| \frac{11,1165 - 0}{5,57971} \right| = 1,9923 > 1,9633 = t(714)_{0,05/2}$

Educación madre $(MED)$ $\quad |t| = \left| \frac{11,2765 - 0}{6,49354} \right| = 1,7366 < 1,9633 = t(714)_{0,05/2}$

Por lo tanto, se puede concluir que, para un nivel de significación del 5 %, la variable educación de la madre no es individualmente significativa, mientras que la variable nivel de eduación del padre sí es individualmente significativa.

Contraste de significatividad conjunta.

$$H_0 : \beta_3 = \beta_4 = 0 \qquad F = \frac{(SCR_R - SCR_{NR})/q}{SCR_{NR}/(N-k)} \overset{H_0}{\sim} \mathcal{F}(q, N-k)$$
$$H_a : \beta_3 \neq 0 \text{ y/o } \beta_4 \neq 0$$

donde $SCR_{NR}$ es la suma de cuadrados de residuos del modelo no restringido, Modelo P2.2, $SCR_R$ es la suma de cuadrados de residuos del modelo restringido, Modelo P2.1, y $q = 2$ es el número de restricciones.

$$F = \frac{(107598000 - 106000000)/2}{106000000/(718-4)} = 6,83883 \; > \; 3,00834 = \mathcal{F}(2,714)_{0,05}$$

Se rechaza la hipótesis nula con un nivel de significación del 5 %, es decir, se puede concluir que las variables educación de la padre y educación de la madre son conjuntamente significativas.

9. Intervalos de confianza.

$$\begin{aligned} IC(\beta_3)_{95\,\%} &= \left[\hat\beta_3 \pm t(N-k)_{\alpha/2}\ \hat\sigma_{\hat\beta_3}\right] = \\ &= \left[11,1165 \pm t(718-4)_{0,05/2} \times 5,5797\right] = [0,161883 \quad 22,0711] \end{aligned}$$

$$\begin{aligned} IC(\beta_4)_{95\,\%} &= \left[\hat\beta_4 \pm t(N-k)_{\alpha/2}\ \hat\sigma_{\hat\beta_4}\right] = \\ &= \left[11,2765 \pm t(718-4)_{0,05/2} \times 6,4935\right] = [-1,47218 \quad 24,0252] \end{aligned}$$

Dado que $0 \in IC(\beta_4)_{95\,\%}$ y $0 \notin IC(\beta_3)_{95\,\%}$, llegamos a la misma conclusión que en el apartado anterior: con un nivel de confianza del 95 %, la variable educación del padre es individualemte significativa, mientras que la variable nivel de eduación de la madre no es individualmente significativa.

## TERCERA PARTE.

1. Hipótesis del experto: $\beta_3 = \beta_4$. Sustituyéndola en el modelo Modelo P2.2:

$$\begin{aligned} S_i &= \beta_1 + \beta_2\, ED_i + \beta_3\, FED_i + \beta_4\, MED_i + u_i \quad \rightarrow \\ S_i &= \beta_1 + \beta_2\, ED_i + \beta_3\, FED_i + \beta_3\, MED_i + u_i \quad \rightarrow \\ S_i &= \beta_1 + \beta_2\, ED_i + \beta_3\,(FED_i + MED_i) + u_i \end{aligned}$$

2. Contraste de hipótesis.

$$\begin{array}{l} H_0 : \beta_3 = \beta_4 \\ H_a : \beta_4 \neq \beta_4 \end{array} \qquad F = \frac{(SCR_R - SCR_{NR})/q}{SCR_{NR}/(N-k)} \overset{H_0}{\sim} \mathcal{F}(q, N-k)$$

Este estadístico se puede escribir también en términos del coeficiente de determinación:

$$F = \frac{(R^2_{NR} - R^2_R)/q}{(1 - R^2_{NR})/(N-k)} \overset{H_0}{\sim} \mathcal{F}(q, N-k)$$

donde $R^2_{NR}$ es el coeficiente de determinación del modelo no restringido, Modelo P2.2, $R^2_R$ es el coeficiente de determinación del modelo restringido, Modelo P2.3, y $q = 1$ es el número de restricciones.

$$F = \frac{(0,120364 - 0,12)/1}{(1 - 0,120364)/(718 - 4)} = 0,000235316 \; < \; 3,85452 = \mathcal{F}(1, 714)_{0,05}$$

No se rechaza la hipótesis nula con un nivel de significación del 5 %, es decir, se puede concluir que los niveles de educación del padre y de la madre tienen el mismo efecto en la determinación del salario del individuo.

3. El Modelo P2.3 porque este modelo es el Modelo P2.2 en el que se ha incluido una restricción que es cierta, dado el resultado obtenido en el contraste del apartado anterior. Estimar el Modelo P2.3 por MCO es lo mismo que estimar el Modelo P2.2 por Mínimos Cuadrados Restringidos. Como la restricción es cierta, los estimadores de Mínimos Cuadrados Restringidos siguen siendo insesgados y tienen una varianza menor que los estimadores MCO del Modelo P2.2.

# Solución PRÁCTICA P3.

### PRIMERA PARTE.

1. MRLG: $\quad S_i = \beta_1 + \beta_2 \, A_i + \beta_3 \, X_i + u_i \quad u_i \sim NID(0, \sigma^2_u) \quad i = 1, 2, \ldots, 336.$

2. Problema de optimización MCO se basa en minimizar la suma de los residuos al cuadrado:

$$\sum \hat{u}^2_i \; = \; \sum (S_i - \hat{\beta}_1 - \hat{\beta}_2 \, A_i - \hat{\beta}_3 \, X_i)^2.$$

Las condiciones de primer orden, resultantes de igualar a cero las derivadas parciales de la función objetivo respecto de cada uno los estimadores, son:

$$-2\sum(S_i-\hat\beta_1-\hat\beta_2\,A_i-\hat\beta_3\,X_i)=0 \quad\Rightarrow\quad \sum(S_i-\hat\beta_1-\hat\beta_2\,A_i-\hat\beta_3\,X_i)=0$$

$$-2\sum(S_i-\hat\beta_1-\hat\beta_2\,A_i-\hat\beta_3\,X_i)\,A_i=0 \quad\Rightarrow\quad \sum(S_i-\hat\beta_1-\hat\beta_2\,A_i-\hat\beta_3\,X_i)\,A_i=0$$

$$-2\sum(S_i-\hat\beta_1-\hat\beta_2\,A_i-\hat\beta_3\,X_i)\,X_i=0 \quad\Rightarrow\quad \sum(S_i-\hat\beta_1-\hat\beta_2\,A_i-\hat\beta_3\,X_i)\,X_i=0$$

Ecuaciones normales:

$$\begin{aligned}\sum S_i &= N\,\hat\beta_1+\hat\beta_2\sum A_i+\hat\beta_3\sum X_i\\ \sum S_i\,A_i &= \hat\beta_1\sum A_i+\hat\beta_2\sum A_i^2+\hat\beta_3\sum X_i\,A_i\\ \sum S_i\,X_i &= \hat\beta_1\sum X_i+\hat\beta_2\sum A_i\,X_i+\hat\beta_3\sum X_i^2\end{aligned}$$

3. $\hat\beta_2=61{,}4109$. Si la experiencia previa aumenta en un año, ceteris paribus, el incremento estimado del salario es de 61,4109 dólares.

   La estimación tiene el signo esperado, ya que indica una relación positiva entre la experiencia y el salario, a más experiencia más salario.

4. $\widehat{S}_5=3943,29\,-\,26,2979\times 18\,+\,61,4109\times 3=3654,160$ dólares.

5. $\widehat{S}(A=3,X=0)=\hat\beta_1+\hat\beta_2\times 3=3943,29-26,2979\times 3=3864,3963$ dólares.

6. Como la única diferencia entre los individuos es su edad, la diferencia entre sus salarios proviene de este factor. Dada la interpretación del coeficiente $\hat\beta_2$:

   Diferencia salarial: $\hat\beta_2\,\times\,10=-26,2979\,\times\,10=-262,979$ dólares.

7. $R^2=1-\dfrac{SCR}{SCT}=1-\dfrac{\sum\hat u_i^2}{\sum(S_i-\bar S)^2}=0,0493792$

   Interpretación: el 4,93792 % de la variabilidad de salario actual en la muestra está explicada por la variabilidad de las variables explicativas: edad y experiencia previa.

8. Contraste de significatividad conjunta.

   $H_0:\beta_2=\beta_3=0$
   $H_a:\beta_2\neq 0$ y/o $\beta_3\neq 0$

   $$F=\frac{R^2/q}{(1-R^2)/(N-k)}\overset{H_0}{\sim}\mathcal{F}(q,N-k)$$

$$F = \frac{0,0493792/2}{(1-0,0493792)/(336-3)} = 8,6487 \; > \; 3,02284 = \mathcal{F}(2,333)_{0,05}$$

Se rechaza la hipótesis nula con un nivel de significación del 5 %, es decir, se puede concluir que las variables explicativas, experiencia previa y edad son conjuntamente significativas.

Contraste de significatividad individual.

$$\begin{array}{l} H_0 : \beta_i = 0 \\ H_a : \beta_i \neq 0 \end{array} \qquad t = \frac{\hat{\beta}_i - 0}{\hat{\sigma}_{\hat{\beta}_i}} \overset{H_0}{\sim} t(T-k) \qquad i = 2,3.$$

$$\text{Edad } (A) \qquad |t| = \left| \frac{-26,2979 - 0}{15,8930} \right| = 1,655 \; < \; 1,96711 = t(336-3)_{0,05/2}$$

$$\text{Experiencia } (X) \quad |t| = \left| \frac{61,4109 - 0}{14,9234} \right| = 4,115 \; > \; 1,96711 = t(336-3)_{0,05/2}$$

Por lo tanto, se puede concluir que, para un nivel de significación del 5 %, la variable edad no es individualmente significativa, mientras que la variable experiencia previa sí es individualmente significativa.

## SEGUNDA PARTE.

1. MRLG:

$$S_i = \beta_1 \, + \, \beta_2 \, A_i \, + \, \beta_3 \, X_i \, + \, \beta_4 \, M_i \, + \, \beta_5 \, L_i \, + \, u_i \quad u_i \sim NID(0, \sigma_u^2) \quad i = 1,..,336.$$

2. La variable nivel de estudios es una variable cualitativa dividida en tres categorías. Para incluirla en el modelo, se definen tres variables ficticias, una para cada categoría.

   Para evitar problemas de colinealidad perfecta en la matriz de datos $X$, se ha especificado el Modelo P3.2 incluyendo el término constante y sólo dos de las tres variables ficticias, las correspondientes a los estudios medios ($M$) y a los estudios superiores ($L$).

3. $\beta_1 = E(S_i | A = 0, X = 0, M = 0, L = 0)$ : salario actual medio de un empleado de 18 años con estudios básicos y sin experiencia previa.

$\beta_2$ : variación esperada en el salario actual si aumenta en un un año la edad del empleado (mayor de edad) manteniendo constante la experiencia previa e independientemente del nivel de estudios.

$\beta_3$ : variación esperada en el salario actual si aumenta en un año la experiencia previa manteniendo constante la edad e independientemente del nivel de estudios.

$\beta_4$ : diferencia esperada en el salario actual entre un trabajador con estudios medios y un trabajador con estudios básicos para los mismos niveles de edad y experiencia previa.

$\beta_5$ : diferencia esperada en el salario actual entre un trabajador con estudios superiores y un trabajador con estudios básicos para los mismos niveles de edad y experiencia previa.

4.

$$X = \begin{bmatrix} & \text{A} & \text{X} & \text{M} & \text{L} \\ 1 & 15 & 7 & 1 & 0 \\ 1 & 36 & 14 & 1 & 0 \\ 1 & 12 & 11 & 1 & 0 \\ 1 & 16 & 9 & 1 & 0 \\ 1 & 18 & 3 & 0 & 1 \\ 1 & 36 & 9 & 0 & 0 \\ 1 & 29 & 8 & 0 & 1 \\ 1 & 26 & 1 & 0 & 1 \\ 1 & 24 & 14 & 0 & 1 \\ 1 & 19 & 6 & 1 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \end{bmatrix}$$

5. El primer individuo de la muestra tiene estudios medios, 33 años y 7 años de experiencia laboral previa.

$$\begin{aligned} \widehat{S}_1 &= 3028,05 - 27,8645 \times 15 + 63,5817 \times 7 + 508,577 \times 1 + 2106,81 \times 0 = \\ &= 3563,733 \text{ dólares.} \end{aligned}$$

6. Contraste de significatividad conjunta.

$$\begin{array}{l} H_0 : \beta_2 = \beta_3 = \beta_4 = \beta_5 = 0 \\ H_a : \beta_2 \neq 0 \text{ y/o } \beta_3 \neq 0 \text{ y/o } \beta_4 \neq 0 \text{ y/o } \beta_5 \neq 0 \end{array}$$

$$F = \frac{R^2/q}{(1-R^2)/(N-k)} \overset{H_0}{\sim} \mathcal{F}(q, N-k)$$

$$F = \frac{0,235795/4}{(1-0,235795)/(336-5)} = 25,5325 \; > \; 2,39893 = \mathcal{F}(4,331)_{0,05}$$

Se rechaza la hipótesis nula con un nivel de significación del 5 %, es decir, se puede concluir que las variables explicativas, edad, experiencia previa y nivel de estudios son conjuntamente significativas.

7. Contraste de significatividad individual.

**A.** Edad y experiencia previa.

$$\begin{array}{l} H_0 : \beta_i = 0 \\ H_a : \beta_i \neq 0 \end{array} \qquad t = \frac{\hat{\beta}_i - 0}{\hat{\sigma}_{\hat{\beta}_i}} \overset{H_0}{\sim} t(N-k) \qquad i = 2,3.$$

$$\text{Edad } (A) \qquad |t| = \left| \frac{-27,8645 - 0}{14,3069} \right| = 1,948 \; < \; 1,96716 = t(331)_{0,05/2}$$

$$\text{Experiencia } (X) \quad |t| = \left| \frac{63,5817 - 0}{4,732} \right| = 3,4048 \; > \; 1,96716 = t(331)_{0,05/2}$$

Por lo tanto, se puede concluir que, para un nivel de significación del 5 %, la variable edad no es individualmente significativa mientras que la variable experiencia previa si es individualmente significativa.

**B.** La variable nivel de estudios no es significativa si, a igualdad de condiciones, no hay diferencias en el salario actual promedio de un trabajador por tener un nivel de estudios u otro.

$$\begin{array}{l} H_0 : \beta_4 = \beta_5 = 0 \\ H_a : \beta_4 \neq 0 \text{ y/o } \beta_5 \neq 0 \end{array} \qquad F = \frac{(SCR_R - SCR_{NR})/q}{SCR_{NR}/(N-k)} \overset{H_0}{\sim} \mathcal{F}(q, N-k)$$

donde $SCR_{NR}$ es la suma de cuadrados de residuos del modelo no restringido, Modelo P3.2, $SCR_R$ es la suma de cuadrados de residuos del modelo restringido, Modelo P3.1, y $q = 2$ es el número de restricciones.

$$F = \frac{(1146540000 - 921708000)/2}{921708000/(336-5)} = 40,3712 \; > \; 3,02301 = \mathcal{F}(2,331)_{0,05}$$

Se rechaza la hipótesis nula con un nivel de significación del 5 %, es decir, se puede concluir que el nivel de estudios es una variable significativa para explicar el salario actual, una vez incluidas el resto de las variables.

8. En el Modelo P3.2 la variable nivel de estudios es individualmente significativa para determinar los salarios. Este resultado implica que el Modelo P3.1 está mal especificado porque omite una variable relevante.

   Los estimadores MCO del Modelo P3.1 y el estimador de la varianza de las perturbaciones son, por lo tanto, sesgados. Además, los estadísticos $t$ y $F$ no siguen las distribuciones habituales por lo que si se utilizan para realizar inferencia, los resultados no serían válidos.

# Solución PRÁCTICA P4.

## PRIMERA PARTE.

1. MRLG: $P_i = \beta_1 + \beta_2 A_i + \beta_3 S_i + \beta_4 B_i + u_i \quad u_i \sim NID(0, \sigma_u^2) \quad i = 1, 2, \ldots, 321.$

2. FRM:
$$\underset{(\hat{\sigma}_{\hat{\beta}})}{\widehat{P}_i} = \underset{(6{,}2563)}{8,7629} - \underset{(0{,}0568)}{0,3075}\ A_i + \underset{(0{,}0357)}{0,3280}\ S_i + \underset{(3{,}2074)}{12,2529}\ B_i$$
$$R^2 = 0,529875 \qquad SCR = 281066.$$

3. Interpretación del coeficiente $\beta_4$:

   $\beta_4$ : variación esperada en el precio ante un cambio unitario en el número de baños de la vivienda manteniendo constantes la antigüedad y la superficie.

   $\hat{\beta}_4 = 12,25$ miles de dólares, es decir, se estima que, ceteris paribus, un cuarto de baño adicional se valora en 12250 dólares.

4. $R^2 = 1 - \dfrac{SCR}{SCT} = 1 - \dfrac{\sum \hat{u}_i^2}{\sum (P_i - \bar{P})^2} = 0,529875$

   Interpretación: el 52,99 % de la variabilidad del precio de las viviendas en la muestra viene explicada por la variabilidad de las variables explicativas: antigüedad, superficie y número de baños.

5. $\hat{\sigma}_u^2 = \dfrac{SCR}{N-k} = \dfrac{\sum \hat{u}_i^2}{N-k} = \dfrac{281066}{321-4} = 886,64.$

6. Contraste de significatividad conjunta.

$$\begin{array}{l} H_0 : \beta_2 = \beta_3 = \beta_4 = 0 \\ H_a : \beta_2 \neq 0 \text{ y/o } \beta_3 \neq 0 \text{ y/o } \beta_4 \neq 0 \end{array} \qquad F = \frac{R^2/q}{(1-R^2)/(N-k)} \overset{H_0}{\sim} \mathcal{F}(q, N-k)$$

$$F = \frac{0,5299/3}{(1-0,5299)/(321-4)} = 119,096 \; > \; 2,64 = \mathcal{F}(3,317)_{0,05}$$

Se rechaza la hipótesis nula con un nivel de significación del 5 %, es decir, se puede concluir que las variables explicativas antigüedad, superficie y número de baños son conjuntamente significativas.

7. Contraste de significatividad individual: variable antigüedad.

$$\begin{array}{l} H_0 : \beta_2 = 0 \\ H_a : \beta_2 \neq 0 \end{array} \qquad t = \frac{\hat{\beta}_2 - 0}{\hat{\sigma}_{\hat{\beta}_2}} \overset{H_0}{\sim} t(N-k)$$

$$|t| = \left| \frac{-0,3075 - 0}{0,05688} \right| = 5,4066 \; > \; 1,96748 = t(321-4)_{0,05/2}$$

Se rechaza la hipótesis nula con un nivel de significación del 5 %, es decir, se puede concluir que la antigüedad es una variable significativa para explicar el precio de venta.

8. Contraste de hipótesis (500 dólares = 0,5 miles de dólares).

$$\begin{array}{l} H_0 : \beta_3 = 0,5 \\ H_a : \beta_3 \neq 0,5 \end{array} \qquad t = \frac{\hat{\beta}_3 - 0,5}{\hat{\sigma}_{\hat{\beta}_3}} \overset{H_0}{\sim} t(N-k)$$

$$|t| = \left| \frac{0,3280 - 0,5}{0,0357} \right| = 4,8179 \; > \; 1,96748 = t(321-4)_{0,05/2}$$

Se rechaza la hipótesis nula con un nivel de significación del 5 % y concluimos que, manteniéndo el resto de las variables constantes, no se está dispuesto a pagar 500 dólares por un cuarto de baño más.

9. Contraste de restricción lineal,

$$\begin{array}{l} H_0 : \beta_2 = -\beta_3 \;\Rightarrow\; \beta_2 + \beta_3 = 0 \\ H_a : \beta_2 \neq -\beta_3 \;\Rightarrow\; \beta_2 + \beta_3 \neq 0 \end{array} \qquad t = \frac{\hat{\beta}_2 + \hat{\beta}_3 - 0}{\sqrt{\widehat{var}(\hat{\beta}_2 + \hat{\beta}_3)}} \overset{H_0}{\sim} t(N-k)$$

$$\begin{aligned} \widehat{var}(\hat{\beta}_2 + \hat{\beta}_3) &= \widehat{var}(\hat{\beta}_2) + \widehat{var}(\hat{\beta}_3) + 2\widehat{cov}(\hat{\beta}_2, \hat{\beta}_3) = \\ &= 0,0032357 + 0,00128121 + 2 \times (-0,000556436) = 0,003 \end{aligned}$$

$$|t| = \left| \frac{(-0,3075 + 0,3280) - 0}{\sqrt{0,003}} \right| = 0,3746 \; < \; 1,96748 = t(321-4)_{0,05/2}$$

No se rechaza la hipótesis nula con un nivel de significación del 5 %, es decir, se puede concluir que, manteniéndo el resto de las variables constantes, el efecto negativo sobre el precio de la antigüedad de una vivienda se ve compensado por el efecto positivo de un aumento de la superficie.

10. Dado el resultado del contraste del apartado anterior, propondríamos estimar un modelo alternativo que incluyera la restricción $\beta_2 + \beta_3 = 0$. El modelo restringido se construye incluyendo la restricción en el modelo inicial, Modelo P4.1:

$$\begin{aligned} P_i &= \beta_1 + \beta_2\, A_i + \beta_3\, S_i + \beta_4\, B_i + u_i \quad \rightarrow \\ P_i &= \beta_1 + \beta_2\, A_i - \beta_2\, S_i + \beta_4\, B_i + u_i \quad \rightarrow \\ P_i &= \beta_1 + \beta_2\,(A_i - S_i) + \beta_4\, B_i + u_i \end{aligned}$$

    Estimar este modelo por MCO es lo mismo que estimar el modelo inicial sujeto a la restricción $\beta_2 + \beta_3 = 0$ por Mínimos Cuadrados Restringidos. Como la restricción es cierta, sabemos que las propiedades de los estimadores por Mínimos Cuadrados Restringidos son las siguientes: lineales, insesgados y con menor varianza que los estimadores MCO del modelo inicial.

11. Los residuos de la estimación MCO de un modelo de regresión lineal que estuviera bien especificado, es decir, que cumpliera todos los supuestos, se deberían comportar de forma similar a las perturbaciones: oscilando en torno a cero con varianza constante y de forma aleatoria, es decir, sin ningún patrón de comportamiento reconocible.

    Analizando el gráfico de los residuos del Modelo P4.1, se puede ver claramente que el comportamiento de los residuos no es uniforme u homogéneo. De hecho, los residuos de la primera parte de la muestra son en su mayoría negativos, mientras que los de la segunda parte de la muestra son, en general, positivos. Por lo tanto, parece que alguna hipótesis básica del modelo no se cumple. Es posible que se haya cometido un error en la especificación de la parte sistemática del modelo, por ejemplo, que se haya omitido alguna variable explicativa relevante que clasifica los datos en dos grupos.

## SEGUNDA PARTE.

1. La implantación de la incineradora es una variable explicativa cualitativa con dos categorías:

    - la incineradora existe a partir de 1981,
    - la incineradora no existe antes de 1981.

Por lo tanto, para introducir este regresor en el modelo se han de definir dos variables ficticias, una por cada categoría:

$$D_i^{81} = \begin{cases} 1 & i \in \text{ vendidas en 1981} \\ 0 & \text{en otro caso} \end{cases} \qquad D_i^{80} = \begin{cases} 1 & i \in \text{ vendidas en 1980} \\ 0 & \text{en otro caso} \end{cases}$$

Al especificar el modelo hay que tener cuidado para no tener problemas de colinealidad perfecta. Para ello especificamos el modelo con término constante e incluyendo tantas variables ficticias como categorías menos una:

$$P_i = \beta_1 + \beta_2 \, A_i + \beta_3 \, S_i + \beta_4 \, B_i + \beta_5 \, D_i^{81} + u_i.$$

En este último caso, la variable ficticia $D^{81}$ toma el valor 0 para la primera observación de la muestra (casa vendida en 1980) y 1 para la última observación (casa vendida en 1981).

2. Dado que la variable ficticia $INC$ representa el mismo papel que la variable $D^{81}$ propuesta en el apartado anterior:

   $\hat{\beta}_5$: diferencia estimada en el precio de las viviendas vendidas en 1981 con respecto a las del año 1980 que tienen las mismas características de antigüedad, superficie y número de baños.

   En un principio, se podría esperar que la puesta en marcha de la incineradora con su impacto medioambiental tuviera un efecto negativo sobre el precio. El coeficiente estimado es, sin embargo, positivo. Esto puede ser debido a que la incineradora trajo consigo una mejora de las comunicaciones que se espera que tenga un efecto positivo sobre el precio y, al parecer, ha pesado más el efecto positivo (comunicaciones) que el negativo (incineradora).

3. Contraste de significatividad individual: puesta en marcha de la incineradora.

$$\begin{array}{l} H_0 : \beta_5 = 0 \\ H_a : \beta_5 \neq 0 \end{array} \qquad t = \frac{\hat{\beta}_5 - 0}{\hat{\sigma}_{\hat{\beta}_5}} \overset{H_0}{\sim} t(N-k)$$

$$|t| = \left| \frac{35,7839 - 0}{2,8174} \right| = 12,7007 \; > \; 1,9675 = t(321-5)_{0,05/2}$$

   Se rechaza la hipótesis nula con un nivel de significación del 5 %, es decir, se puede concluir que la puesta en marcha de la incineradora es una variable significativa para explicar el precio de venta, una vez incluidas el resto de las variables.

4. Comparando este gráfico con los residuos del Modelo P4.1, se observa que los residuos del Modelo P4.2 oscilan en torno a cero y no se aprecia ya que haya una

diferencia de comportamiento entre los residuos correspondientes a la primera parte de la muestra y los de la segunda parte. Este resultado parece indicar que la introducción de la variable explicativa cualitativa incineradora, que es significativa, ha recogido el comportamiento no aleatorio que presentaban los residuos del Modelo P4.1 y el problema se ha solucionado. Se concluye que el Modelo P4.1 estaba mal especificado por omisión de variable relevante.

## TERCERA PARTE.

1. Modelo P4.3: $P_i = \beta_1 + \beta_2 A_i + \beta_3 A_i^2 + \beta_4 S_i + \beta_5 B_i + \beta_6 INC_i + u_i$.

   Este modelo tiene las mismas variables explicativas que el Modelo P4.2. La diferencia entre ambos está en el término $A_i^2$, que supone incluir la variable antigüedad en el modelo con una forma funcional diferente: cuadrática. De esta forma, se trata de captar un posible efecto no lineal de la variable antigüedad sobre el precio de las viviendas, de manera que, aunque el precio disminuya con la antigüedad, no lo haga de forma lineal sino que el efecto vaya amortiguándose a medida que aumentan los años de la vivienda.

2. El Modelo P4.3 no incumple ninguna hipótesis básica del MRLG. Cuando se dice que el modelo tiene que ser "lineal", significa que debe ser lineal en los coeficientes y éste lo es. El hecho de que la variable $A$ aparezca elevada al cuadrado no supone ninguna dificultad para la estimación de los parámetros ni afecta a las propiedades de los estimadores. A lo que sí afecta es a la interpretación de los coeficientes, ya que el coeficiente que acompaña a $A_i^2$ no puede interpretarse ahora como un multiplicador de la manera habitual.

3. Contraste de significatividad de la variable antigüedad.

$$\begin{array}{l} H_0 : \beta_2 = \beta_3 = 0 \\ H_a : \beta_2 \neq 0 \text{ y/o } \beta_3 \neq 0 \end{array} \qquad F = \frac{(SCR_R - SCR_{NR})/q}{SCR_{NR}/(N-k)} \overset{H_0}{\sim} \mathcal{F}(q, N-k)$$

   donde $q = 2$ es el número de restricciones, $SCR_{NR}$ es la suma de cuadrados de residuos del modelo no restringido, Modelo P4.3, y $SCR_R$ es la suma de cuadrados de residuos del siguiente modelo restringido:

$$P_i = \beta_1 + \beta_4 S_i + \beta_5 B_i + \beta_6 INC_i + u_i.$$

$$F = \frac{(198008 - 178257)/2}{178257/(321-6)} = 17,4505 \; > \; 3,0244 = \mathcal{F}(2, 315)_{0,05}$$

   Se rechaza la hipótesis nula con un nivel de significación del 5 %, es decir, se puede concluir que la antigüedad es una variable significativa para explicar el precio de venta, una vez incluidas el resto de las variables.

4. El efecto de la antigüedad en el precio depende de los años que tenga la casa:

$$\frac{\partial \widehat{P}}{\partial A_i} = \hat{\beta}_1 + 2\hat{\beta}_2\, A_i = -0,7 + 2 \times 0,0033\, A_i = -0,7 + 0,0066\, A_i$$

$$5 \text{ años} \qquad \frac{\partial \widehat{P}}{\partial A_i} = -0,7 + 0,0066 \times 5 = -0,667 \,\text{miles de dólares}$$

$$50 \text{ años} \qquad \frac{\partial \widehat{P}}{\partial A_i} = -0,7 + 0,0066 \times 50 = -0,37 \,\text{miles de dólares}$$

5. La relación entre la variable explicativa antigüedad y el precio es no lineal si el término $A_i^2$ es estadísticamente significativo, es decir, si el parámetro $\beta_3$ es distinto de cero.

$$\begin{array}{l} H_0 : \beta_3 = 0 \\ H_a : \beta_3 \neq 0 \end{array} \qquad t = \frac{\hat{\beta}_3 - 0}{\hat{\sigma}_{\hat{\beta}_3}} \overset{H_0}{\sim} t(N-k)$$

$$|t| = \left| \frac{0,00328 - 0}{0,000882} \right| = 3,7176 \;>\; 1,96752 = t(321-6)_{0,05/2}$$

Se rechaza la hipótesis nula con un nivel de significación del 5 %, es decir, se puede concluir que la relación entre el precio y la antigüedad de las casas es no lineal.

6. En el Modelo P4.3 todas las variables son significativas individualmente: en el apartado 3 se comprobó la significatividad de la variable antigüedad y en la tabla correspondiente a este modelo se puede comprobar que los estadísticos t para las variables superficie, baños e $INC$ superan el valor critío por lo que también son significativas. Además, en el apartado anterior se ha comprobado que la forma funcional que relaciona la antigüedad y el precio no es lineal.

   Por lo tanto, si elegimos un modelo que no sea el Modelo P4.3, tendríamos problemas.

   En el Modelo P4.1 se incurre en omisión de variables relevantes por no incluir $INC$ y en mala especificación de la forma funcional por no considerar el término $A_i^2$, es decir, por considerar que la relación entre el precio y la antigüedad es lineal en vez de cuadrática. En el Modelo P4.2, tendríamos problemas sólo de mala especificación de la forma funcional.

   Pero, en ambos casos estaremos sesgando la estimación de los coeficientes (esto puede comprobarse observando que los valores de los coeficientes de todas las variables cambian de una estimación a otra, lo que indica que no hay ortogonalidad entre las variables de la matriz $X$). Además, será sesgada la estimación de la varianza de las perturbaciones, lo que llevará a calcular estadísticos que no siguen la distribución propuesta y por lo tanto invalidará los contrastes.

Todo ello nos lleva a elegir el Modelo P4.3 que, a priori, no muestra indicios de mala especificación.

7. Si tuviera que explicar qué es lo que determina el precio de los pisos, empezaría por lo más obvio: el tamaño del piso y sus condiciones de uso, representadas por el número de baños, son variables importantes para marcar el precio; evidentemente, a mayor tamaño y mejores condiciones, mayor precio. Por otro lado, también tiene importancia la antigüedad: cuanto más años tenga un piso, menos valor tiene en términos equivalentes de tamaño y condiciones. Sin embargo, este efecto no es siempre proporcional a los años de antigüedad; a partir de un número de años, la desvalorización del piso se empieza a frenar; podríamos pensar que ello sucede porque ese punto marca la diferencia entre un piso simplemente "viejo" y un piso que se considera "antiguo".

   Por último, el hecho de haber construido una incineradora en la zona (que a priori podría pensarse que iba a tener un efecto negativo en el precio de los pisos) ha hecho aumentar los precios. Esto pone de manifiesto que las comunicaciones tienen un efecto relevante en los precios y que se pagan más los pisos en zonas bien comunicadas, ya que una mejora de la comunicación por autopista ha sido capaz de contrarrestar los posibles efectos negativos medioambientales.

# Solución PRÁCTICA P5.

## PRIMERA PARTE.

1. Modelo P5.1:

$$V_t = \beta_1 + \beta_2\, dq1_t + \beta_3\, dq2_t + \beta_4\, dq3_t + u_t \qquad u_t \sim NID(0, \sigma_u^2) \quad t = 2001{:}1, \ldots, 2008{:}4.$$

   En el Modelo P5.1 hay una sola variable explicativa, la estacionalidad, que recoge el efecto de que las ventas promedio son distintas en los diferentes trimestres del año.

   La estacionalidad es una variable cualitativa, con cuatro categorías, cada uno de los cuatro trimestres. Se han definido cuatro variables ficticias estacionales, una por cada trimestre y, para evitar problemas de colinealidad perfecta en la matriz de datos $X$, se ha especificado el modelo incluyendo el término constante más tantas variables ficticias como número de categorías menos una. En este caso, se ha eliminado la variable ficticia estacional correpondiente al cuarto trimeste. En consecuencia, el cuarto trimestre se convierte en el de referencia en la interpretación de los coeficientes del Modelo P5.1.

$\beta_1 = E(V_t|dq1 = 0, dq2 = 0, dq3 = 0)$ : ventas promedio en el cuarto trimestre.

$\beta_2 = E(V_t|dq1 = 1) - E(V_t|dq1 = 0, dq2 = 0, dq3 = 0) = (\beta_1 + \beta_2) - \beta_1$
$=$ diferencia en las ventas medias entre el primer y el cuarto trimestre.

$\beta_3 = E(V_t|dq2 = 1) - E(V_t|dq1 = 0, dq2 = 0, dq3 = 0) = (\beta_1 + \beta_3) - \beta_1$
$=$ diferencia en las ventas medias entre el segundo y el cuarto trimestre.

$\beta_4 = E(V_t|dq3 = 1) - E(V_t|dq1 = 0, dq2 = 0, dq3 = 0) = (\beta_1 + \beta_4) - \beta_1$
$=$ diferencia en las ventas medias entre el tercer y el cuarto trimestre.

2. La variable estacionalidad no es significativa si no hay diferencias en las ventas promedio entre trimestres.

$$\begin{array}{l} H_0 : \beta_2 = \beta_3 = \beta_4 = 0 \\ H_a : \beta_2 \neq 0 \text{ y/o } \beta_3 \neq 0 \text{ y/o } \beta_4 \neq 0 \end{array}$$

Como la variable estacionalidad es la única variable explicativa del Modelo P5.1, el contraste de significatividad individual es a su vez un contraste de significatividad conjunta, por lo que podemos utilizar el siguiente estadístico:

$$F = \frac{R^2/q}{(1 - R^2)/(T - k)} \overset{H_0}{\sim} \mathcal{F}(q, T - k)$$

$$F = \frac{0,5318/3}{(1 - 0,5318)/(32 - 4)} = 12,1156 \ > \ 2,94669 = \mathcal{F}(3, 28)_{0,05}$$

Se rechaza la hipótesis nula con un nivel de significación del 5 %, es decir, se puede concluir que la estacionalidad es una variable significativa para explicar las ventas de refrigeradores.

3. $\widehat{V}_{\text{año}} = \widehat{V}_{1TR} + \widehat{V}_{2TR} + \widehat{V}_{3TR} + \widehat{V}_{4TR}$

Dada la interpretación de los coeficientes del Modelo P5.1 (véase apartado 1):

$$\widehat{V}_{\text{año}} = (\hat{\beta}_1 + \hat{\beta}_2) + (\hat{\beta}_1 + \hat{\beta}_3) + (\hat{\beta}_1 + \hat{\beta}_3) + \hat{\beta}_1 = 5419,375 \ \text{ miles de euros}$$

## SEGUNDA PARTE.

1. Modelo P5.3:

$$V_t = \beta_1 + \beta_2\, P_t \;+\; \beta_3\, dq1_t + \beta_4\, dq2_t + \beta_5\, dq3_t + u_t \qquad t = 2001{:}1, \ldots, 2008{:}4.$$

   El Modelo P5.3 tiene dos variables explicativas, gasto en publicidad y estacionalidad, mientras que el Modelo P5.1 solo tiene una variable explicativa, estacionalidad, y el Modelo P5.2 solo tiene una variable explicativa, gasto en publicidad.

2. Como contamos con 32 observaciones para 4 variables explicativas, el orden de la matriz de datos $X$ es $(32 \times 5)$.

$$X = \begin{bmatrix} 1 & 252,6 & 1 & 0 & 0 \\ 1 & 272,4 & 0 & 1 & 0 \\ 1 & 270,9 & 0 & 0 & 1 \\ 1 & 273,9 & 0 & 0 & 0 \\ 1 & 268,9 & 1 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 356,4 & 0 & 0 & 0 \end{bmatrix}$$

3. Contraste de significatividad conjunta.

$$\begin{array}{l} H_0 : \beta_2 = \beta_3 = \beta_4 = \beta_5 = 0 \\ H_a : \beta_2 \neq 0 \text{ y/o } \beta_3 \neq 0 \text{ y/o } \beta_4 \neq 0 \text{ y/o } \beta_5 \neq 0 \end{array}$$

$$F = \frac{R^2/q}{(1-R^2)/(T-k)} \overset{H_0}{\sim} \mathcal{F}(q, T-k)$$

$$F = \frac{0,7298/4}{(1-0,7298)/(32-5)} = 18,231495 \;>\; 2,72777 = \mathcal{F}(4,27)_{0,05}$$

   Se rechaza la hipótesis nula con un nivel de significación del 5 %, es decir, se puede concluir que las variables explicativas estacionalidad y gastos en publicidad son conjuntamente significativas.

   Contraste de significatividad individual: gasto en publicidad.

$$\begin{array}{l} H_0 : \beta_2 = 0 \\ H_a : \beta_2 \neq 0 \end{array} \qquad t = \frac{\hat\beta_2 - 0}{\hat\sigma_{\hat\beta_2}} \overset{H_0}{\sim} t(T-k)$$

$$|t| = \left| \frac{2,773 - 0}{0,623} \right| = 4,451043339 \;>\; 2,05183 = t(32-5)_{0,05/2}$$

Se rechaza la hipótesis nula con un nivel de significación del 5 %, es decir, se puede concluir que el gasto en publicidad es una variable significativa para explicar las ventas de refrigeradores.

Contraste de significatividad individual: estacionalidad.

$$H_0 : \beta_3 = \beta_4 = \beta_5 = 0 \qquad F = \frac{(SCR_R - SCR_{NR})/q}{SCR_{NR}/(T-k)} \overset{H_0}{\sim} \mathcal{F}(q, T-k)$$
$$H_a : \beta_3 \neq 0 \text{ y/o } \beta_4 \neq 0 \text{ y/o } \beta_5 \neq 0$$

donde $SCR_{NR}$ es la suma de cuadrados de residuos del modelo no restringido, Modelo P5.3, $SCR_R$ es la suma de cuadrados de residuos del modelo restringido, Modelo P5.2, y $q = 3$ es el número de restricciones.

$$F = \frac{(1377140 - 465085)/3}{465085/(32-5)} = 17,64945 \; > \; 2,96035 = \mathcal{F}(3,27)_{0,05}$$

Se rechaza la hipótesis nula con un nivel de significación del 5 %, es decir, se puede concluir que la estacionalidad es una variable significativa para explicar las ventas de refrigeradores, una vez incluido el gasto en publicidad en el modelo.

4. Incremento estimado en las ventas = 2,773 miles de euros = 2773 euros, porque:

   $\hat{\beta}_2 = 2,773$ mide el incremento estimado en las ventas en miles de euros cuando el gasto en publicidad aumenta en una unidad (100 euros) independientemente del trimestre del año.

5. $\widehat{V}_{1TR} = 370,164 + 2,733\, P_t + 86,080 = (456,224 + 2,733\, P_t)$ miles de euros.

   $$\widehat{V}_{1TR} = 370,164 + 2,733\, P_t + 86,080 = 456,224 + 2,733 \times 17 =$$
   $$= 502,685 \text{ miles de euros.}$$

6. Intervalo de confianza.

   $$\begin{aligned} IC(\beta_2)_{95\,\%} &= \left[\hat{\beta}_2 \pm t(T-k)_{\alpha/2}\, \hat{\sigma}_{\hat{\beta}_2}\right] = \left[2,773 \pm t(32-5)_{0,05/2} \times 0,623\right] = \\ &= [2,773 \pm 1,27829009] = [1,494701 \;\; 4,05129009] \text{ miles de euros} \end{aligned}$$

7. ¿$\Delta P = 1\,\text{euro} \;\; \rightarrow \;\; \Delta V = 50\,\text{euros}$?

   $\Delta P = 100\,\text{euros} \;\; \rightarrow \;\; \Delta V = 5000\,\text{euros} = 5\,\text{miles de euros} \;\; \Rightarrow \;\; \hat{\beta}_2 = 5$

   Como 5 no está incluido en el $IC(\beta_2)_{95\,\%}$ calculado en el apartado anterior, se puede concluir, con una confianza del 95 %, que no es posible que un aumento de un euro en la publicidad trimestral pueda aumentar las ventas medias trimestrales en 50 euros.

8. Incremento gastos en publicidad: 250 euros.

$$\begin{aligned}
\widehat{V}_t &= 370,164 + 2,773 \times P_t + 86,08\, dq1_t + 328,578\, dq2_t + 411,345\, dq3_t \\
\widehat{V}_{t+4} &= 370,164 + 2,773 \times P_{t+4} + 86,08\, dq1_{t+4} + 328,578\, dq2_{t+4} + \\
&+ 411,345\, dq3_{t+4} \\
\Delta_4 \widehat{V}_{t+4} &= \widehat{V}_{t+4} - \widehat{V}_t = \\
&= 2,773 \times \Delta_4\, P_{t+4} + 86,08\, \Delta_4\, dq1_{t+4} + 328,578\, \Delta\, dq2_{t+4} + \\
&+ 411,345\, \Delta\, dq3_{t+4} = 2,773 \times 2,50 = 6,9325 \text{ miles de euros}
\end{aligned}$$

Primer trimestre: $\Delta_4 \widehat{V}_{t+4} = 6,9325$ miles de euros.

Este resultado es debido a que al estar comparando un trimestre con el mismo trimestre del año anterior, el incremento en las ventas dado un incremento de los gastos en publicidad es independiente del trimestre en el que nos encontremos.

9. Predicción por intervalo para el primer trimestre de 2009.

$$IC(V_{2009:1})_{95\,\%} = \left[\widehat{V}_{2009:1} \;\pm\; t(32-5)_{0,025}\; \hat{\sigma}_p\right]$$

donde:

$$\widehat{V}_{2009:1} = 370,164 + 86,080 = 456,244 \text{ miles de euros}$$

$$\begin{aligned}
\hat{\sigma}_p^2 &= \hat{\sigma}_u^2\; \left[1 + X_p'(X'X)^{-1}\, X_p\right] = \\
&= \frac{465085}{32-5} \left[1 + \begin{bmatrix} 1 & 0 & 1 & 0 & 0 \end{bmatrix} (X'X)^{-1} \begin{bmatrix} 1 \\ 0 \\ 1 \\ 0 \\ 0 \end{bmatrix}\right] = 49003,79
\end{aligned}$$

$$IC(V_{2009:1})_{95\,\%} = \left[456,244 \;\pm\; 2,052\; \sqrt{49003,79}\right] = [2,03 \quad 910,25]$$

Como el valor 0 no está incluido en el intervalo de predicción para las ventas del primer trimestre de 2009, se puede concluir con una confianza del 95 % que el asesor del gerente no tiene razón.

### TERCERA PARTE.

1. En el Modelo P5.4 solo se introducen las variables ficticias estacionales referidas al segundo y tercer trimestres, dejando fuera el primer y cuarto trimestres. Por lo tanto, en este modelo el primer y el cuarto trimestre son tratados de la misma manera. En términos de los coeficientes del modelo Modelo P5.3, la restricción incluida es $\beta_3 = 0$, lo que significa que, a igualdad de condiciones, la diferencia entre las ventas del primer trimestre y el cuarto es nula, por lo tanto, basta distinguir entre el segundo trimestre, el tercero, y el resto de los trimestres (primero y cuarto).

2. Restricción en el Modelo P5.3: $\beta_3 = 0$

$$\begin{array}{l} H_0 : \beta_3 = 0 \\ H_a : \beta_3 \neq 0 \end{array} \qquad t = \frac{\hat{\beta}_3 - 0}{\hat{\sigma}_{\hat{\beta}_3}} \overset{H_0}{\sim} t(T-k)$$

$$|t| = \left| \frac{86,08 - 0}{65,843} \right| = 1,307 \, < \, 2,05183 = t(32-5)_{0,05/2}$$

No se rechaza la hipótesis nula con un nivel de significación del 5 %, es decir, se puede concluir que, a igualdad de gasto en publicidad, la diferencia entre las ventas del primer trimestre y el cuarto es nula.

3. $\widehat{V}_{1TR} = 431,890 + 2,706 \times 3 = 440,008$ miles de euros.
$\widehat{V}_{2TR} = 431,890 + 2,706 \times 3 + 285,319 = 725,327$ miles de euros.
$\widehat{V}_{3TR} = 431,890 + 2,706 \times 3 + 368,554 = 808,562$ miles de euros.
$\widehat{V}_{4TR} = \widehat{V}_{1TR} = 431,890 + 2,706 \times 3 = 440,008$ miles de euros.

### CUARTA PARTE.

Se utilizaría el Modelo P5.4 para especificar las ventas de refrigeradores porque en el Modelo P5.3 hemos comprobado que las dos variables explicativas consideradas a lo largo de este ejercicio, gasto en publicidad y estacionalidad, son individualmente significativas.

Por lo tanto, se han de rechazar los modelos Modelo P5.1 y Modelo P5.2 porque tanto en uno como en otro se está omitiendo una variable explicativa relevante, por lo que los estimadores MCO son sesgados y los estadísticos $t$ y $F$ no siguen las distribuciones habituales por lo que si se hace inferencia de la forma habitual podríamos llegar a conclusiones erróneas.

En lo que se refiere a la comparación entre el Modelo P5.3 y el Modelo P5.4, este último no es más que el Modelo P5.3 incluyendo una restricción que se ha contrastado en la Tercera Parte que es cierta. Por lo tanto, los estimadores MCR del Modelo P5.3 son insesgados y tienen una varianza menor que los estimadores MCO del Modelo P5.3.

# Solución PRÁCTICA P6.

### PRIMERA PARTE.

1. MRLG: $V_t = \beta_1 + \beta_2 M_t + \beta_3 C_t + \beta_4 S_t + u_t \quad u_t \sim NID(0, \sigma_u^2) \quad t = 1, 2, \ldots, 23.$

2. FRM:

$$\underset{(\hat\sigma_{\hat\beta})}{\widehat{V}_t} = \underset{(4{,}2865)}{33,7991} - \underset{(0{,}126)}{0,2223}\, M_t + \underset{(0{,}059)}{0,0249}\, C_t + \underset{(0{,}005)}{0,0129}\, S_t \qquad R^2 = 0,911653 \qquad SCR = 105,657.$$

$\hat\beta_1 = 33,7991$ : ventas promedio estimadas de marisco cuando las variables precio de la empresa, precio de los competidores y sueldo medio de los trabajadores toman el valor cero.

$\hat\beta_2 = -0,2223$ : se estima que las ventas disminuyen en 0,2223 miles de kilos cuando el precio del kilo de marisco de la empresa aumenta en un euro, manteniéndose constantes las variables precio de los competidores y sueldo medio de los trabajadores.

$\hat\beta_3 = 0,0249$ : se estima que las ventas aumentan en 0,0249 miles de kilos cuando el precio del kilo de marisco de las empresas competidoras aumenta en un euro, manteniéndose constantes las variables precio de la empresa y sueldo medio de los trabajadores.

$\hat\beta_4 = 0,0129$ : se estima que las ventas aumentan en 0,0129 miles de kilos cuando el sueldo medio de los trabajadores aumenta en cien euros, manteniéndose constantes las variables precio de la empresa y precio de los competidores.

3. $R^2 = 1 - \dfrac{SCR}{SCT} = 1 - \dfrac{\sum \hat u_t^2}{\sum (V_t - \bar V)^2} = 0,911653$

Interpretación: el 91,1653 % de la variabilidad de las ventas en la muestra viene explicada por la variabilidad de las variables explicativas: precio medio de venta del marisco, precio medio de los competidores y salario medio de la localidad.

4. Contraste de significatividad conjunta.

$$\begin{array}{l} H_0 : \beta_2 = \beta_3 = \beta_4 = 0 \\ H_a : \beta_2 \neq 0 \text{ y/o } \beta_3 \neq 0 \text{ y/o } \beta_4 \neq 0 \end{array} \qquad F = \frac{R^2/q}{(1-R^2)/(T-k)} \overset{H_0}{\sim} \mathcal{F}(q, T-k)$$

$$F = \frac{0,911653/3}{(1-0,911653)/(23-4)} = 65,3533 \, > \, 3,12735 = \mathcal{F}(3,19)_{0{,}05}$$

Se rechaza la hipótesis nula con un nivel de significación del 5 %, es decir, se puede concluir que las variables explicativas precio de la empresa, precio de los competidores y sueldo medio de los trabajadores son conjuntamente significativas.

5. Contraste de significatividad individual.

$$\begin{array}{ll} H_0 : \beta_i = 0 \\ H_a : \beta_i \neq 0 \end{array} \qquad t = \frac{\hat{\beta}_i - 0}{\hat{\sigma}_{\hat{\beta}_i}} \overset{H_0}{\sim} t(T-k) \qquad i = 2, 3, 4.$$

Precio marisco ($M$) $\quad |t| = \left| \frac{-0,2223 - 0}{0,12} \right| = 1,762 \; < \; 2,093 = t(23-4)_{0,05/2}$

Precio competidor ($C$) $\quad |t| = \left| \frac{0,0249 - 0}{0,059} \right| = 0,4229 \; < \; 2,093 = t(23-4)_{0,05/2}$

Sueldo ($S$) $\quad |t| = \left| \frac{0,0129 - 0}{0,00503} \right| = 2,577 \; > \; 2,093 = t(23-4)_{0,05/2}$

Por lo tanto, se puede concluir que, para un nivel de significación del 5 %, las variables precio del marisco y precio de los competidores no son individualmente significativas, mientras que la variable sueldo sí es individualmente significativa.

Estamos analizando la demanda de marisco, por lo que, en principio, sería de esperar que factores como renta y precios fueran determinantes en su evolución. Los resultados obtenidos son lógicos en cuanto a la variable explicativa sueldo se refiere, ya que sí es significativa, pero no son los esperados para las dos variables de precios, sobre todo para la variable de precio del propio producto, que según el razonamiento económico debería ser relevante a la hora de explicar las ventas.

6. Contraste de hipótesis (175 kilos = 0,175 miles de kilos).

$$\begin{array}{ll} H_0 : \beta_3 = 0,175 \\ H_a : \beta_3 < 0,175 \end{array} \qquad t = \frac{\hat{\beta}_3 - 0,175}{\hat{\sigma}_{\hat{\beta}_3}} \overset{H_0}{\sim} t(T-k)$$

$$t = \frac{0,0249 - 0,175}{0,05908} = -2,5406 \; < \; -1,72913 = t(23-4)_{0,05}$$

Se rechaza la hipótesis nula con un nivel de significación del 5 % y concluimos que, manteniendo el resto de las variables constantes, si el precio de las empresas competidoras aumenta en un euro no sería posible que la empresa incrementara sus ventas en más de 175 kilos.

7. Interpretación del coeficiente $\beta_4$:

$\beta_4 = \frac{\partial E(V_t)}{\partial S_t}$ : variación esperada en las ventas ante un incremento unitario (100 euros) en el sueldo, manteniendo constante el resto de las variables.

$\hat{\beta}_4 = 0,0129$ miles de kilogramos. Por lo tanto, se estima que si el sueldo disminuye en cien euros, manteniendose el resto de variables constantes, las ventas disminuyen en 0,0129 miles de kilogramos = 12,9 kilogramos.

8. MRLG:

$$V_t = \beta_1 + \beta_2 M_t + \beta_3 C_t + \beta_4 S_t + \beta_5 R_t + u_t \quad u_t \sim NID(0, \sigma_u^2) \quad t = 1, 2, \ldots, 23.$$

El hecho de que, en la muestra, la correlación entre dos variables explicativas, renta y sueldo, sea tan alta significa que existe un alto grado de multicolinealidad entre los regresores. La multicolinealidad no supone que se incumpla ninguna de las hìpótesis básicas del MRLG por lo que el modelo anterior se puede estimar por MCO, y estos estimadores seguirían teniendo buenas propiedades, es decir, serían lineales, insesgados y óptimos (en el sentido de varianza mínima dentro de la clase de estimadores lineales e insesgados). El único problema que genera un alto grado de multicolinealidad, es que al ser tan alta la correlación entre los datos de renta y sueldo, las varianzas estimadas de los estimadores MCO serán grandes y se tenderá a no rechazar la $H_0$ en los contrastes de significación individual correspondientes a las estas dos variables explicativas.

9. Analizando el gráfico de los residuos del Modelo P6.1, se puede ver claramente que no oscilan aleatoriamente en torno a cero. De hecho, existe un patrón claro de comportamiento: los residuos de la primera parte y la última parte de la muestra son negativos, mientras que los del centro de la muestra son positivos. Por lo tanto, parece que alguna hipótesis básica del modelo no se cumple. Podría ser que se haya cometido un error en la especificación de la parte sistemática del modelo, por ejemplo, que se haya omitido alguna variable explicativa relevante.

## SEGUNDA PARTE.

1. Modelo P6.2:

$$V_t = \beta_1 + \beta_2 PUB_t + \beta_3 M_t + \beta_4 C_t + \beta_5 S_t + u_t \quad u_t \sim NID(0, \sigma_u^2) \quad t = 1, 2, \ldots, 23.$$

La diferencia con respecto al Modelo P6.1 es la inclusión del factor explicativo campaña publicitaria. Este es un factor cualitativo con dos categorías (campaña o no campaña) que se introduce a través de variables ficticias. En este caso de las dos variables ficticias que se pueden definir, $PUB$ y $NOPUB$, se ha incluido en el modelo una sola ($PUB$) manteniendo el término constante para evitar problemas de multicolinealidad perfecta en la matriz de regresores $X$.

2. $\hat{\beta}_2 = 4,2235$: diferencia estimada en las ventas (en miles de kilos) entre hacer una campaña publicitaria o no hacerla, manteniéndose el resto de las variables, precio, precio del competidor y sueldo iguales.

Es decir, se estima que, en igualdad de condiciones, la campaña publicitaria aumenta las ventas medias en 4,2235 miles de kilos.

3. Contraste a una cola.

$$\begin{array}{l} H_0 : \beta_2 = 0 \\ H_a : \beta_2 > 0 \end{array} \qquad t = \frac{\hat{\beta}_2 - 0}{\hat{\sigma}_{\hat{\beta}_2}} \overset{H_0}{\sim} t(T-k)$$

$$t = \frac{4,2235 - 0}{0,34863} = 12,1147 \; > \; 1,73406 = t(23-5)_{0,05}$$

Se rechaza la hipótesis nula con un nivel de significación del 5 %, es decir, se puede concluir que existe evidencia de que la campaña publicitaria ha tenido un efecto positivo en las ventas.

4. Contraste de significatividad individual: precio del marisco.

$$\begin{array}{l} H_0 : \beta_3 = 0 \\ H_a : \beta_3 \neq 0 \end{array} \qquad t = \frac{\hat{\beta}_3 - 0}{\hat{\sigma}_{\hat{\beta}_3}} \overset{H_0}{\sim} t(T-k)$$

$$|t| = \left| \frac{-0,2754 - 0}{0,04307} \right| = 6,3954 \; > \; 2,10092 = t(23-5)_{0,05/2}$$

Se rechaza la hipótesis nula con un nivel de significación del 5 %, es decir, se puede concluir que la variable precio medio de venta de marisco es individualmente significativa.

Sin embargo, esta variable no era significativa individualmente en el Modelo P6.1. Las razones de esta diferencia de resultados es la siguiente. Como se ha comprobado en el apartado 2., la variable publicidad tiene efecto sobre las ventas por lo que el Modelo P6.1 estaba mal especificado por omisión de una variable relevante, la campaña publicitaria. Como se puede demostrar, si en un MRLG se omiten variables relevantes, los estimadores MCO están sesgados y la inferencia no es válida, porque los estadísticos $t$ y $F$ no siguen las distribuciones habituales.

5. Contraste de restricción lineal.

$$\begin{array}{ll} H_0 : \beta_3 = -\beta_4 & (\beta_3 + \beta_4 = 0) \\ H_a : \beta_3 \neq -\beta_4 & (\beta_3 + \beta_4 \neq 0) \end{array} \qquad t = \frac{\hat{\beta}_3 + \hat{\beta}_4 - 0}{\sqrt{\widehat{var}(\hat{\beta}_3 + \hat{\beta}_4)}} \overset{H_0}{\sim} t(T-k)$$

$$\begin{aligned} \widehat{var}(\hat{\beta}_3 + \hat{\beta}_4) &= \widehat{var}(\hat{\beta}_3) + \widehat{var}(\hat{\beta}_4) + 2\widehat{cov}(\hat{\beta}_3, \hat{\beta}_4) = \\ &= 0,0018554 + 0,000415674 + 2 \times (-0,000156414) = 0,001958 \end{aligned}$$

$$|t| = \left| \frac{(-0,2754 + 0,0687) - 0}{\sqrt{0,001958}} \right| = |-4,67| > 2,101 = t(23-5)_{0,05/2}$$

Se rechaza la hipótesis nula con un nivel de significación del 5 %, es decir, se puede concluir que, manteniéndo el resto de las variables constantes, el gerente no puede contrarrestar una bajada del precio de sus competidores con una disminución de sus precios en la misma cantidad.

6. Comparando este gráfico con los residuos del Modelo P6.1, se observa que los residuos del Modelo P6.2 oscilan aleatoriamente en torno a cero y no se aprecia ya que haya una diferencia de comportamiento entre los residuos correspondientes a la primera parte de la muestra, los de la parte central y los de la última parte. Este resultado parece indicar que la inclusión de la variable explicativa cualitativa publicidad (que tiene un efecto positivo sobre las ventas) ha recogido el comportamiento no aletorio que presentaban los residuos del Modelo P6.1 y el problema se ha solucionado. Se puede concluir que el Modelo P6.1 estaba mal especificado por omisión de variable relevante.

7. Sustituyendo en la función de regresión muestral:

$$\begin{aligned} \widehat{V}_{2004:4} &= 31,058 + 4,2235 \times 1 - 0,2754 \times 30 + 0,0687 \times 32 + 0,0103 \times 250 \\ \widehat{V}_{2004:4} &= 39,2287 \text{ miles de kilos} \end{aligned}$$

## TERCERA PARTE.

1. La estacionalidad es una variable explicativa cualitativa con cuatro categorías, por lo tanto, se introduce en el modelo a través de variables ficticias. Se pueden definir tantas variables ficticias como categorías:

$$T1_t = \begin{cases} 1 & t \in \ 1^{\text{o}} \text{ trimestre} \\ 0 & \text{en otro caso} \end{cases} \qquad T2_t = \begin{cases} 1 & t \in \ 2^{\text{o}} \text{ trimestre} \\ 0 & \text{en otro caso} \end{cases}$$

$$T3_t = \begin{cases} 1 & t \in \ 3^{\text{o}} \text{ trimestre} \\ 0 & \text{en otro caso} \end{cases} \qquad T4_t = \begin{cases} 1 & t \in \ 4^{\text{o}} \text{ trimestre} \\ 0 & \text{en otro caso} \end{cases}$$

Para evitar problemas de colinealidad perfecta entre las columnas de la matriz $X$, especificamos el modelo con término independiente y tantas variables ficticias como categorías menos una. Dejamos fuera la variable ficticia del trimestre 4, de forma que el modelo propuesto, Modelo P6.3, es:

$$V_t = \beta_1 + \beta_2 M_t + \beta_3 C_t + \beta_4 S_t + \beta_5 PUB_t + \beta_6 T1_t + \beta_7 T2_t + \beta_8 T3_t + u_t$$

2. La matriz de datos para el Modelo P6.3 es:

$$
\begin{array}{c}
\text{t} \\ 1999{:}1 \\ 1999{:}2 \\ 1999{:}3 \\ 1999{:}4 \\ 2000{:}1 \\ 2000{:}2 \\ 2000{:}3 \\ 2000{:}4 \\ \vdots
\end{array}
\qquad
\text{X} = \begin{bmatrix}
\text{C} & \text{PUB} & M_t & C_t & S_t & T1_t & T2_t & T3_t \\
1 & 0 & M_1 & C_1 & S_1 & 1 & 0 & 0 \\
1 & 0 & M_2 & C_2 & S_2 & 0 & 1 & 0 \\
1 & 0 & M_3 & C_3 & S_3 & 0 & 0 & 1 \\
1 & 0 & M_4 & C_4 & S_4 & 0 & 0 & 0 \\
1 & 0 & M_5 & C_5 & S_5 & 1 & 0 & 0 \\
1 & 0 & M_6 & C_6 & S_6 & 0 & 1 & 0 \\
1 & 1 & M_7 & C_7 & S_7 & 0 & 0 & 1 \\
1 & 1 & M_8 & C_8 & S_8 & 0 & 0 & 0 \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots
\end{bmatrix}
$$

3. Contrastes sobre el modelo Modelo P6.3.

3.1. Contraste de significación individual: estacionalidad.

$$
\begin{array}{l}
H_0 : \beta_6 = \beta_7 = \beta_8 = 0 \\
H_a : \beta_6 \neq 0 \text{ y/o } \beta_7 \neq 0 \text{ y/o } \beta_8 \neq 0
\end{array}
$$

$$F = \frac{(SCR_R - SCR_{NR})/q}{SCR_{NR}/(T-k)} \overset{H_0}{\sim} \mathcal{F}(q, T-k)$$

donde $SCR_{NR}$ es la suma de residuos al cuadrado del modelo no restringido, Modelo P6.3, $SCR_R$ es la suma de residuos al cuadrado del Modelo P6.2 y $q = 3$ es el número de restricciones.

$$F = \frac{(11,5427 - 8,13216)/3}{8,13216/(23-8)} = 2,09693 \; < \; 3,28738 = \mathcal{F}(3,15)_{0,05}$$

No se rechaza la hipótesis nula con un nivel de significación del 5 %, es decir, se puede concluir que la estacionalidad no es una variable significativa para explicar el precio de venta, una vez incluidas el resto de las variables.

3.2. Contraste de restricción lineal.

$$
\begin{array}{l}
H_0 : \beta_6 = \beta_7 = \beta_8 \\
H_a : \beta_6 \neq \beta_7 \,\text{y/o}\, \beta_6 \neq \beta_8
\end{array}
\qquad
F = \frac{(SCR_R - SCR_{NR})/q}{SCR_{NR}/(T-k)} \overset{H_0}{\sim} \mathcal{F}(q, T-k)
$$

donde $q = 2$ es el número de restricciones, $SCR_{NR}$ es la suma de residuos al cuadrado del modelo no restringido, Modelo P6.3, $SCR_R$ es la suma de residuos al cuadrado del siguiente modelo restringido, Modelo P6.4:

$$
\begin{aligned}
V_t &= \beta_1 + \beta_2 \, M_t + \beta_3 \, C_t + \beta_4 \, S_t + \beta_5 \, PUB_t + \beta_6 \, T1_t + \beta_6 \, T2_t + \beta_6 \, T3_t + u_t \\
V_t &= \beta_1 + \beta_2 \, M_t + \beta_3 \, C_t + \beta_4 \, S_t + \beta_5 \, PUB_t + \beta_6 \, (T1_t + T2_t + T3_t) + u_t
\end{aligned}
$$

$$F = \frac{(11,1822 - 8,13216)/2}{8,13216/(23-8)} = 2,81296 \; < \; 3,68232 = \mathcal{F}(2,15)_{0,05}$$

No se rechaza la hipótesis nula con un nivel de significación del 5 %, es decir, se puede concluir que no hay diferencia en las ventas medias de los tres primeros trimestres manteniendo el resto de las variables iguales.

4. El Modelo P6.2 porque:

   - Este modelo incluye la variable explicativa publicidad que es significativa. Por lo tanto, el Modelo P6.1 omite una variable relevante, lo que implica que los estimadores MCO de los coeficientes de regresión están sesgados, el estimador de la varianza de las perturbaciones está también sesgado y los estadísticos $t$ y $F$ no siguen las distribuciones habituales por lo que la inferencia no es válida.
   - Este modelo no incluye la estacionalidad que ha resultado ser una variable explicativa no relevante.

# Solución PRÁCTICA P7.

## PRIMERA PARTE.

1. $\hat{\beta}_{MCO} = (X'X)^{-1} X'Y$

$$\begin{bmatrix} \widehat{\beta}_1 \\ \widehat{\beta}_2 \end{bmatrix}_{MCO} = \begin{bmatrix} T & \sum P_t \\ \sum P_t & \sum P_t^2 \end{bmatrix}^{-1} \begin{bmatrix} \sum Q_t \\ \sum Q_t P_t \end{bmatrix} =$$

$$= \begin{bmatrix} 89 & 801,1691 \\ 801,1691 & 7407,915 \end{bmatrix}^{-1} \begin{bmatrix} 225,3035 \\ 1953,55 \end{bmatrix} =$$

$$= \begin{bmatrix} 0,4249483 & -0,0459583 \\ -0,0459583 & 0,0051054 \end{bmatrix} \begin{bmatrix} 225,3035 \\ 1953,55 \end{bmatrix} = \begin{bmatrix} 5,96205 \\ -0,381092 \end{bmatrix}$$

Recta de regresión muestral: $\widehat{Q}_t = 5,96205 - 0,381092 \; P_t \quad t = 1983{:}1, \ldots, 1990{:}5.$

2. $\hat{\beta}_2$: si el precio del galón de silicona aumenta en un dólar, se estima que las ventas disminuyen en 0,381 miles de galones (= 381 galones).

   En principio, el coeficiente estimado tiene el signo esperado ya que como se está estimando una función de demanda, se espera que la relación entre precio y ventas sea negativa, es decir, que un aumento del precio resulte en una disminución de las ventas.

3. Estimador insesgado de la varianza de las perturbaciones: $\hat{\sigma}_u^2 = \sum \hat{u}_t^2/(T-k)$.

   Estimación:

$$\begin{aligned}
\hat{\sigma}_u^2 &= \frac{\sum_{t=1}^{89} \hat{u}_t^2}{T-k} = \frac{Q'Q - \hat{\beta}'X'Q}{89-2} = \\
&= \frac{785,7899 - [5,96205 \quad -0,381092] \begin{bmatrix} 225,3035 \\ 1953,55 \end{bmatrix}}{87} = \\
&= \frac{186,99987}{87} = 2,1494
\end{aligned}$$

4. Contraste a una cola.

$$\begin{array}{l} H_0 : \beta_2 = 0 \\ H_a : \beta_2 < 0 \end{array} \qquad t = \frac{\hat{\beta}_2 - 0}{\hat{\sigma}_{\hat{\beta}_2}} \overset{H_0}{\sim} t(T-k)$$

$$\hat{\sigma}_{\hat{\beta}_2}^2 = \hat{\sigma}_u^2 \, a_{22} = \frac{\sum \hat{u}_t^2}{T-k} \, a_{22} = \frac{186,99987}{89-2} \times 0,005105404 = 0,0109759$$

$$t = \frac{-0,381092 - 0}{\sqrt{0,0109759}} = -3,6375 \, < \, -1,6625 = t(89-2)_{0,05}$$

   Se rechaza la hipótesis nula con un nivel de significación del 5 %, es decir, se puede concluir que existe evidencia en la muestra de que el precio influye negativamente en la cantidad media de silicona vendida.

5. Para que se cumpla el Teorema de Gauss-Markov y el estimador MCO sea en consecuencia el estimador lineal e insesgado de varianza mínima se han de cumplir todas las hipótesis básicas de modelo de regresión lineal salvo la de normalidad de las perturbaciones. Es decir, se han de cumplir las siguientes hipótesis:

   - Modelo bien especificado.
   - Modelo lineal en los coeficientes.

- Coeficientes constantes.
- Regresores fijos.
- Ausencia de colinealidad perfecta.
- Media de la perturbación cero.
- Varianza de la perturbación constante.
- No existencia de autocorrelación en las perturbaciones.

6. Se ha de construir en primer lugar un intervalo de predicción de 95 % de probabilidad de la cantidad vendida de silicona bajo el nuevo precio. Si la cantidad que espera vender el dueño pertenece a este intervalo de predicción, según el modelo y la información muestral disponible, el dueño está en lo cierto. En caso contrario, el dueño no está en lo cierto.

$$IC(Q_p)_{95\,\%} = \left[\widehat{Q}_p \;\pm\; t(89-2)_{0{,}025}\;\hat{\sigma}_p\right]$$

donde

$$\widehat{Q}_p \;=\; \hat{\beta}_1 + \hat{\beta}_2\,P_p = 5,96205\,-0,381092\,\times\,9,5 = 2,341676 \;\text{ miles de galones}$$

$$\begin{aligned}\hat{\sigma}_p^2 &= \hat{\sigma}_u^2\;\left[1 + X_p'(X'X)^{-1}\,X_p\right] = \\ &= 2,1494\;\left[1\;+\;[1 \qquad 9,5]\begin{bmatrix} 0,424948 & -0,045958 \\ -0,045958 & 0,0051054\end{bmatrix}\begin{bmatrix} 1 \\ 9,5\end{bmatrix}\right] = 2,1762\end{aligned}$$

$$\begin{aligned}IC(Q_p)_{95\,\%} &= \left[2,3416 - 1,9876\,\times\,\sqrt{2,1762} \quad 2,3416 + 1,9876\,\times\,\sqrt{2,1762}\right] = \\ &= [-0,590 \quad 5,274]\;\text{miles de galones.}\end{aligned}$$

El dueño prevee que las ventas pueden ser de 3500 galones = 3,5 miles de galones. Como $3,5\,\in\,IC(Q_p)_{95\,\%}$, existe evidencia de que el dueño está en lo cierto.

## SEGUNDA PARTE.

1. Modelo P7.2:

$$Q_t = \alpha_1 + \alpha_2\,P_t + \alpha_3\,CC_t + \alpha_4\,ICPP_t + u_t \quad u_t \sim NID(0,\sigma_u^2) \quad t = 1983{:}1,\ldots,1990{:}5.$$

2. Contrastes de significatividad.

A. Contraste de significatividad individual de las dos variables explicativas.

$$\begin{array}{l} H_0 : \alpha_i = 0 \\ H_a : \alpha_i \neq 0 \end{array} \qquad t = \frac{\hat{\alpha}_i - 0}{\hat{\sigma}_{\hat{\alpha}_i}} \overset{H_0}{\sim} t(T-k) \qquad i = 3, 4.$$

$$CC \qquad |t| = \left| \frac{0,0228 - 0}{0,0049} \right| = 4,6412 > 1,98827 = t(89-4)_{0,05/2}$$

$$ICCP \qquad |t| = \left| \frac{0,22376 - 0}{0,0081} \right| = 2,7514 > 1,98827 = t(89-4)_{0,05/2}$$

Por lo tanto, se puede concluir que, para un nivel de significación del 5 %, las variables casas que se van a comenzar a construir e índice combinado de construcción son individualmente significativas para explicar las ventas.

B. Contraste de significatividad conjunta de las dos variables explicativas.

$$\begin{array}{l} H_0 : \alpha_3 = \alpha_4 = 0 \\ H_a : \alpha_3 \neq 0 \text{ y/o } \alpha_4 \neq 0 \end{array} \qquad F = \frac{(SCR_R - SCR_{NR})/q}{SCR_{NR}/(T-k)} \overset{H_0}{\sim} \mathcal{F}(q, T-k)$$

donde $SCR_{NR}$ es la suma de cuadrados de residuos del modelo no restringido, Modelo P7.2, $SCR_R$ es la suma de cuadrados de residuos del modelo restringido, Modelo P7.1, y $q = 2$ es el número de restricciones.

$$F = \frac{(186,99987 - 140,69)/2}{140,69/(89-4)} = 13,9894 > 3,10384 = \mathcal{F}(2,85)_{0,05}$$

Se rechaza la hipótesis nula con un nivel de significación del 5 %, es decir, se puede concluir que las dos variables, $CC$ e $ICPP$ son conjuntamente significativas para explicar las ventas.

No hay ninguna contradicción en los resultados. Las dos variables son individualmente significativas y, por lo tanto, son también conjuntamente significativas.

3. En el apartado anterior se obtiene que las dos nuevas variables explicativas incluidas en el Modelo P7.2 son individualmente significativas. Por lo tanto, el Modelo P7.1 no cumple las hipótesis básicas del modelo de regresión lineal porque no está bien especificado: hay omisión de variables relevantes. Esto hace que el estimador MCO $\hat{\beta}_2$ en el Modelo P7.1 sea, en general, sesgado. Además, tampoco se cumple el Teorema de Gauss-Markov, por lo que el estimador MCO $\hat{\beta}_2$ no es óptimo, es decir, no es el de varianza mínima.

4.

$$
\begin{aligned}
\widehat{Q}_{t+1} &= -1,0331 - 0,326389\, P_{t+1} + 0,0228171\, CC_{t+1} + 0,22376\, ICCP_{t+1} \\
\widehat{Q}_{t} &= -1,0331 - 0,326389\, P_{t} + 0,0228171\, CC_{t} + 0,22376\, ICCP_{t} \\
\Delta\widehat{Q}_{t+1} &= -0,326389\, \Delta P_{t+1} + 0,0228171\, \Delta CC_{t+1} + 0,22376\, \Delta ICCP_{t+1} = \\
&= -0,326389 \times (-1) + 0,0228171 \times 0,5 + 0,22376 \times 20 = \\
&= 4,813 \text{ miles de galones}
\end{aligned}
$$

En este escenario, se estima que la venta de silicona disminuiría en 481,3 galones.

## TERCERA PARTE.

1. Modelo P7.3:

$$Q_t = \delta_1 + \delta_2\, P_t + \delta_3\, P_t^2 + \delta_4, CC_t + \delta_5\, ICPP_t + u_t \qquad t = 1983{:}1, \ldots, 1990{:}5.$$

El Modelo P7.3 tiene las mismas variables explicativas que el Modelo P7.2, precio ($P$), casas a construir ($CC$) e índice combinado de construcción ($ICPP$). Por lo tanto, la diferencia entre ambos modelos no se encuentra en el número de variables explicativas, sino en la forma funcional.

En el Modelo P7.2 la relación entre la variable dependiente $Q$ y cada una de las variables explicativas ($P, CC, ICPP$) es lineal de forma que, ceteris paribus, una variación unitaria en una variable explicativa tiene un efecto marginal constante en las ventas, $Q$.

En el caso del precio, en el Modelo P7.2:

$$\frac{\partial E(Q_t)}{\partial P_t} = \alpha_2 \qquad t = 1983{:}1, 1983{:}2, \ldots, 1990{:}5.$$

En el Modelo P7.3 la relación entre la variable dependiente $Q$ y cada una de las variables explicativas $CC$ y $ICPP$ sigue siendo lineal, mientras que la relación entre la variable endógena $Q$ y la variable explicativa precio ya no es lineal sino cuadrática, de forma que el efecto del precio sobre las ventas no va a ser constante sino que va a depender del nivel del precio. Así, en el Modelo P7.3:

$$\frac{\partial E(Q_t)}{\partial P_t} = \delta_2 \,+\, 2\;\; \delta_3\, P_t \qquad t = 1983{:}1, 1983{:}2, \ldots, 1990{:}5.$$

Por lo tanto, cuanto mayor sea el precio, el efecto marginal del precio sobre las ventas será mayor (o menor) dependiendo del signo de $\delta_2$ y $\delta_3$.

2. Sí. El Modelo P7.3 es lineal en los coeficientes (aunque no en las variables), no existe colinealidad perfecta (se han podido estimar los coeficientes del modelo) y, con la información disponible, el modelo está bien especificado. Además, la perturbacion sigue una distribución normal de media cero y varianza constante ($\sigma_u^2$). Por último, la covarianza entre pertubaciones distintas es cero, dado que son independientes.

3. El efecto marginal estimado de un incremento unitario en el precio sobre las ventas, manteniendo el resto de las variables constantes:

$$\frac{\partial \widehat{Q}_t}{\partial P_t} = \hat{\delta}_2 \, + \, 2 \, \times \, \hat{\delta}_3 \, P_t = -0,497857 + 2 \times 0,0082889 \, P_t \qquad t = 1983{:}1, \ldots, 1990{:}5.$$

   Precio = 5 dólares:

$$\frac{\partial \widehat{Q}_t}{\partial P_t} = \hat{\delta}_2 \, + \, 2 \, \hat{\delta}_3 \, P_t = -0,497857 + 2 \times 0,0082889 \times 5 = -0,415 \text{ miles de galones}$$

   Precio = 12 dólares:

$$\frac{\partial \widehat{Q}_t}{\partial P_t} = \hat{\delta}_2 + 2 \times \hat{\delta}_3 \, P_t = -0,497857 + 2 \times 0,0082889 \times 12 = -0,299 \text{ miles de galones}$$

   Se puede observar que cuanto mayor es el precio de partida, menos negativo es el efecto de un incremento unitario del precio sobre las ventas.

4. Contraste sobre la forma funcional apropiada.

$$\begin{array}{l} H_0 : \delta_3 = 0 \\ H_a : \delta_3 \neq 0 \end{array} \qquad t = \frac{\hat{\delta}_3 - 0}{\hat{\sigma}_{\hat{\delta}_3}} \overset{H_0}{\sim} t(T-k)$$

$$|t| = \left| \frac{0,00828892 - 0}{0,0380014} \right| = 0,2181 \, < \, 1,98861 = t(89-5)_{0,05/2}$$

   No se rechaza la hipótesis nula con un nivel de significación del 5 %, es decir, se puede concluir que la relación que existe entre la cantidad vendida y el precio es lineal. No hay evidencia en la muestra a favor de la hipótesis del agente sobre una relación funcional cuadrática entre la cantidad vendida y el precio.

**CUARTA PARTE.**

1. Modelo P7.4:

$$
\begin{aligned}
Q_t &= \gamma_1 + \gamma_2\, P_t + \gamma_3\, CC_t + \gamma_4\, ICPP_t + \gamma_5\, dm1_t + \gamma_6\, dm2_t + \gamma_7\, dm3_t + \gamma_8\, dm4_t + \\
&+ \gamma_9\, dm5_t + \gamma_{10}\, dm6_t + \gamma_{11}\, dm7_t + \gamma_{12}\, dm8_t + \gamma_{13}\, dm9_t + \gamma_{14}\, dm10_t + \\
&+ \gamma_{15}\, dm11_t + u_t \qquad t = 1983{:}1, 1983{:}2, \ldots, 1990{:}5.
\end{aligned}
$$

En este modelo se ha introducido una nueva variable explicativa de carácter cualitativo para recoger el hecho de que la cantidad vendida de silicona puede presentar comportamiento estacional, es decir, para un mismo nivel de las variables explicativas $(P, CC, IPCC)$, la venta media de silicona puede ser distinta cada mes.

Las variables explicativas cualitativas se han de introducir en la especificación del modelo mediante variables ficticias. Como los datos son mensuales, la variable cualitativa estacionalidad tiene doce categorías, una por cada mes, por lo que se definen doce variables ficticias:

$$dm1_t = \begin{cases} 1 & t \in \text{Enero} \\ 0 & \text{en otro caso} \end{cases} \qquad dm2_t = \begin{cases} 1 & t \in \text{Febrero} \\ 0 & \text{en otro caso} \end{cases}$$

$$dm3_t = \begin{cases} 1 & t \in \text{Marzo} \\ 0 & \text{en otro caso} \end{cases} \qquad dm4_t = \begin{cases} 1 & t \in \text{Abril} \\ 0 & \text{en otro caso} \end{cases}$$

$$dm5_t = \begin{cases} 1 & t \in \text{Mayo} \\ 0 & \text{en otro caso} \end{cases} \qquad dm6_t = \begin{cases} 1 & t \in \text{Junio} \\ 0 & \text{en otro caso} \end{cases}$$

$$dm7_t = \begin{cases} 1 & t \in \text{Julio} \\ 0 & \text{en otro caso} \end{cases} \qquad dm8_t = \begin{cases} 1 & t \in \text{Agosto} \\ 0 & \text{en otro caso} \end{cases}$$

$$dm9_t = \begin{cases} 1 & t \in \text{Septiembre} \\ 0 & \text{en otro caso} \end{cases} \qquad dm10_t = \begin{cases} 1 & t \in \text{Octubre} \\ 0 & \text{en otro caso} \end{cases}$$

$$dm11_t = \begin{cases} 1 & t \in \text{Noviembre} \\ 0 & \text{en otro caso} \end{cases} \qquad dm12_t = \begin{cases} 1 & t \in \text{Diciembre} \\ 0 & \text{en otro caso} \end{cases}$$

Para evitar problemas de multicolinealidad perfecta en la matriz $X$, a la hora de especificar el modelo sólo se introducen tantas variables ficticias como número de categorías menos una. En este caso, se introducen once variables ficticias, las correspondientes a los meses de enero a noviembre, eliminando la variable ficticia correspondiente al mes de diciembre. Por lo tanto, el mes de diciembre queda como referencia.

2. Interpretación coeficientes estimados:

$\hat{\gamma}_1 = -0,897081$ : se estima que se venden -0,897081 miles de galones de silicona cuando las variables $P, CC$ e $ICPP$ toman el valor cero y estamos en el mes de diciembre.

$\hat{\gamma}_{14} = 0,251659$ : se estima que, para los mismos niveles de las variables $P, CC$ e $ICPP$, se venden 0,251659 miles de galones más de silicona en noviembre que en diciembre.

3. Constraste de significación individual: estacionalidad.

$$H_0 : \gamma_5 = \gamma_6 = \gamma_7 = \ldots = \gamma_{15} = 0$$
$$H_a : \gamma_5 \neq 0 \text{ y/o } \gamma_6 \neq 0 \text{ y/o } \gamma_7 \neq 0 \ \ldots \text{ y/o } \gamma_{15} \neq 0$$

$$F = \frac{(SCR_R - SCR_{NR})/q}{SCR_{NR}/(T-k)} \overset{H_0}{\sim} \mathcal{F}(q, T-k)$$

donde $SCR_{NR}$ es la suma de cuadrados de residuos del modelo no restringido, Modelo P7.4, $SCR_R$ es la suma de cuadrados de residuos del modelo restringido, Modelo P7.2, y $q = 11$ es el número de restricciones.

$$F = \frac{(140,69 - 117,217)/11}{117,217/(89-15)} = 1,34715 \ < 1,92056 \ = \ \mathcal{F}(11,74)_{0,05}$$

No se rechaza la hipótesis nula con un nivel de significación del 5 %, es decir, se puede concluir que no existe evidencia en la muestra de que las ventas presenten, ceteris paribus, comportamiento estacional.

4. El Modelo P7.2 porque:

- Las variables, $CC$ y $ICCP$ sí son significativas por lo que hay que incluirlas en el modelo. En consecuencia, el Modelo P7.1 omite variables relevantes y los estimadores MCO están sesgados y la inferencia no es válida.
- La evidencia muestral apoya la hipótesis de que la relación funcional entre la cantidad y el precio es lineal por lo que el Modelo P7.3 no está bien especificado.
- La estacionalidad no es una variable relevante, luego no hay que incluirla en el modelo, por lo que descartamos también el Modelo P7.4.

# Solución PRÁCTICA P8.

## PRIMERA PARTE.

1. MRLG:

$$S_i = \beta_1 + \beta_2\, ED_i + \beta_3\, EX_i + \beta_4\, HB_i + u_i \qquad u_i \sim NID(0, \sigma_u^2) \quad i = 1, 2, \ldots, 49.$$

   FRM:

$$\underset{(\hat{\sigma}_{\hat{\beta}})}{\widehat{S}_i} = \underset{(258{,}871)}{434,821} + \underset{(31{,}5081)}{133,551}\, ED_i + \underset{(12{,}1316)}{34,4543}\, EX_i + \underset{(144{,}870)}{470,460}\, HB_i$$

$$R^2 = 0,450263 \qquad SCR = 11089355$$

2. $\hat{\beta}_3 = 34,4543$: incremento estimado en el salario en dólares si la experiencia aumenta en un año manteniendo constantes las variables educación y género.

   $\hat{\beta}_4 = 470,460$: diferencia estimada en el salario entre un hombre y una mujer para los mismos niveles de educación y experiencia.

3. $R^2 = 1 - \dfrac{SCR}{SCT} = 1 - \dfrac{\sum \hat{u}_i^2}{\sum (S_i - \bar{S})^2} = 0,450263$

   Interpretación: el 45,0263 % de la variabilidad de los salarios en la muestra está explicada por la variabilidad de las variables explicativas: educación, experiencia y género.

4. Contraste de significatividad conjunta.

$$\begin{array}{l} H_0 : \beta_2 = \beta_3 = \beta_4 = 0 \\ H_a : \beta_2 \neq 0 \text{ y/o } \beta_3 \neq 0 \text{ y/o } \beta_4 \neq 0 \end{array} \qquad F = \frac{R^2/q}{(1-R^2)/(N-k)} \overset{H_0}{\sim} \mathcal{F}(q, N-k)$$

$$F = \frac{0,450263/3}{(1-0,450263)/(49-4)} = 12,28577483 > 2,81154 = \mathcal{F}(3,45)_{0{,}05}$$

   Se rechaza la hipótesis nula con un nivel de significación del 5 %, es decir, se puede concluir que las variables explicativas educación, experiencia y género son conjuntamente significativas.

5. Existirá discriminación por razones de género si, para el mismo nivel de eduación y experiencia, el salario esperado es diferente para un hombre y para una mujer.

$$\begin{array}{l} H_0 : \beta_4 = 0 \\ H_a : \beta_4 \neq 0 \end{array} \qquad t = \frac{\hat{\beta}_4 - 0}{\hat{\sigma}_{\hat{\beta}_4}} \overset{H_0}{\sim} t(N-k)$$

$$|t| = \left| \frac{470,46 - 0}{144,87} \right| = 3,2475 \; > \; 2,0141 = t(49-4)_{0,05/2}$$

Se rechaza la hipótesis nula con un nivel de significación del 5 %, es decir, se puede concluir que existe discriminación salarial por razones de género.

## SEGUNDA PARTE.

1. Modelo P8.2:

$$S_i = \beta_1 + \beta_2\, ED_i + \beta_3\, EX_i + \beta_4\, HB_i + \beta_5\, ADM_i + \beta_6\, TAL_i + \beta_7\, TEC_i + u_i.$$

La variable ocupación es una variable cualitativa con cuatro categorías: administrativo, empleado de taller, técnico y empleado de mantenimiento. Para introducirla en el modelo se han definido cuatro variables ficticias, una por categoría. Si incluyéramos la variable ocupación en el modelo introduciendo las cuatro variables ficticias más el término constante, la matriz de datos $X$ presentaría colinealidad perfecta y no se podrían estimar los parámetros del modelo. Por lo tanto, se ha elegido especificar el modelo con término constante e incluyendo tantas variables ficticias como número de categorías menos una. En particular, se ha dejado fuera la categoría de empleado de mantenimiento, que servirá de referencia en la interpretación de los coeficientes relativos a la ocupación en el Modelo P8.2.

2. La variable ocupación no es significativa si, a igualdad de condiciones, no hay diferencias en el salario por ocupar un puesto u otro en la empresa.

$$\begin{aligned} H_0 &: \beta_5 = \beta_6 = \beta_7 = 0 \\ H_a &: \beta_5 \neq 0 \text{ y/o } \beta_6 \neq 0 \text{ y/o } \beta_6 \neq 0 \end{aligned} \qquad F = \frac{(SCR_R - SCR_{NR})/q}{SCR_{NR}/(N-k)} \overset{H_0}{\sim} \mathcal{F}(q, N-k)$$

donde $SCR_{NR}$ es la suma de cuadrados de residuos del modelo no restringido, Modelo P8.2, $SCR_R$ es la suma de cuadrados de residuos del modelo restringido, Modelo P8.1, y $q = 3$ es el número de restricciones.

$$F = \frac{(11089355 - 5620313)/3}{5620313/(49-7)} = 13,6232 \; > \; 2,82705 = \mathcal{F}(3,42)_{0,05}$$

Se rechaza la hipótesis nula con un nivel de significación del 5 %, es decir, se puede concluir que la ocupación es una variable significativa para explicar el salario, una vez incluidas el resto de las variables explicativas.

3. Según el resultado del apartado anterior, el Modelo P8.1 omite una variable relevante, ocupación. Los estimadores MCO del Modelo P8.1 estarán sesgados y no son de varianza mínima porque se incumple una de las hipótesis básicas necesarias para que se satisfaga el Teorema de Gauss-Markov.

   El estimador de la varianza de las perturbación es sesgado y los estadísticos $t$ y $F$ ya no siguen las distribuciones $t$ de Student y $\mathcal{F}$ de Snedecor, respectivamente, por lo que los resultados obtenidos en los contrastes de la Primera Parte no son fiabels. Esto significa, entre otras cosas, que no es fiable el resultado que se obtuvo sobre la existencia de discriminación salarial utilizando el Modelo P8.1. Para obtener una conclusión fiable, se vuelve a realizar el contraste de existencia de discriminación salarial en el Modelo P8.2.

$$\begin{array}{l} H_0 : \beta_4 = 0 \\ H_a : \beta_4 \neq 0 \end{array} \qquad t = \frac{\hat{\beta}_4 - 0}{\hat{\sigma}_{\hat{\beta}_4}} \overset{H_0}{\sim} t(N-k)$$

$$|t| = \left| \frac{528,154 - 0}{151,238} \right| = 3,4922 \; > \; 2,01808 = t(49-7)_{0,05/2}$$

   Se rechaza la hipótesis nula con un nivel de significación del 5 %, es decir, se puede concluir que existe discriminación salarial por razones de género.

4. La diferencia estimada es:

$$\begin{aligned} & (\widehat{S}|EX = 3, HB = 0, ADM = 1) - (\widehat{S}|EX = 4, HB = 0, ADM = 1) = \\ = \; & (\hat{\beta}_1 + \hat{\beta}_2\, ED_i + \hat{\beta}_3 \times 4 + \beta_5) - (\hat{\beta}_1 + \hat{\beta}_2\, ED_i + \hat{\beta}_3 \times 3 + \beta_5) = \\ = \; & \hat{\beta}_3 = 29,2526 \text{ dólares.} \end{aligned}$$

5. Contraste a una cola.

$$\begin{array}{l} H_0 : \beta_7 = 1000 \\ H_a : \beta_7 > 1000 \end{array} \qquad t = \frac{\hat{\beta}_7 - 1000}{\hat{\sigma}_{\hat{\beta}_7}} \overset{H_0}{\sim} t(N-k)$$

$$t = \frac{1110,05 - 1000}{203,871} = 0,53980213 \; < \; 1,68195 = t(49-7)_{0,05}$$

   No se rechaza la hipótesis nula con un nivel de significación del 5 %, es decir, se puede concluir que existe evidencia de que la empresa ha cumplido su compromiso y la diferencia de salario entre el personal técnico y el de mantenimiento no ha sido superior a 1000 dólares.

**TERCERA PARTE.**

1. MRLG:

   $$S_i = \beta_1+\beta_2\, ED_i+\beta_3\, EX_i+\beta_4\, HB_i+\beta_5\, ADM_i+\beta_6\, TAL_i+\beta_7\, TEC_i+\beta_8\,(EX_i\times HB_i)+u_i.$$

   La hipótesis de linealidad del MRLG requiere únicamente linealidad en los coeficientes y no en las variables, por lo que el Modelo P8.3, que es lineal en los coeficientes, la cumple.

   Sin embargo, el Modelo P8.3 no es lineal en las variables porque se ha introducido un término $(EX \times HB)$ que recoge la interacción entre la experiencia y el género, es decir, que la variación del salario por un año adicional de experiencia es diferente para hombres y para mujeres.

2. $\widehat{S}_i^H = \hat\beta_1 + \hat\beta_2\, ED_i^H + \hat\beta_3\, EX_i^H + \hat\beta_4 + \hat\beta_5\, ADM_i^H + \hat\beta_6\, TAL_i^H + \hat\beta_7\, TEC_i^H + \hat\beta_8\, EX_i^H$
   $\widehat{S}_i^M = \hat\beta_1 + \hat\beta_2\, ED_i^M + \hat\beta_3\, EX_i^M + \hat\beta_5\, ADM_i^M + \hat\beta_6\, TAL_i^M + \hat\beta_7\, TEC_i^M$

   En igualdad de condiciones:

   $$\widehat{S}_i^H - \widehat{S}_i^M = \hat\beta_4 + \hat\beta_8\, EX_i^H = 52,1988 + 53,7363 \times EX_i^H \text{ dólares}$$

   Es decir, a igualdad de condiciones, la diferencia salarial estimada entre un hombre y una mujer depende de sus años de experiencia.

3. $\dfrac{\partial \hat S_i}{\partial EX_i} = \hat\beta_3 + \hat\beta_8\, HB_i = -8,1917 + 53,7363\, HB_i$ dólares.

   El incremento estimado en el salario por un año más de experiencia depende del género.

4. $\hat\beta_8 = 53,7363$: diferencia en el incremento estimado en el salario por un año más de experiencia entre un hombre y una mujer, a igualdad de ocupación y manteniendo la variable educación constante.

5. Contraste del efecto interacción.

   $$\begin{array}{l} H_0 : \beta_8 = 0 \\ H_a : \beta_8 \neq 0 \end{array} \qquad t = \frac{\hat\beta_8 - 0}{\hat\sigma_{\hat\beta_8}} \overset{H_0}{\sim} t(N-k)$$

   $$|t| = \left|\frac{53,6373 - 0}{19,9972}\right| = 2,6872 \,>\, 2,01954 = t(49-8)_{0,05/2}$$

   Se rechaza la hipótesis nula con un nivel de significación del 5 %, es decir, se puede concluir que un año más de experiencia se valora de forma diferente a un hombre y a una mujer.

6. Existirá discriminación por razones de género si, a igualdad de educación, experiencia y ocupación, el salario esperado es diferente para un hombre y para una mujer.

$$H_0 : \beta_4 = \beta_8 = 0 \qquad F = \frac{(SCR_R - SCR_{NR})/q}{SCR_{NR}/(N-k)} \overset{H_0}{\sim} \mathcal{F}(q, N-k)$$
$$H_a : \beta_4 \neq 0 \text{ y/o } \beta_8 \neq 0$$

donde $q = 2$ es el número de restricciones, $SCR_{NR}$ es la suma de cuadrados de residuos del modelo no restringido, Modelo P8.3, $SCR_R$ es la suma de cuadrados de residuos del siguiente modelo restringido:

$$S_i = \beta_1 + \beta_2\, ED_i + \beta_3\, EX_i + \beta_5\, ADM_i + \beta_6\, TAL_i + \beta_7\, TEC_i + u_i$$

$$F = \frac{(7252280 - 4778683)/2}{4778683/(49-8)} = 10,6114 \; > \; 3,22568 = \mathcal{F}(2,41)_{0,05}$$

Se rechaza la hipótesis nula con un nivel de significación del 5 %, es decir, se puede concluir que existe discriminación salarial por razones de género en la empresa SXF.

7. Se puede observar que los residuos oscilan en torno a cero. Pero se ve que para los individuos con menos experiencia la variabilidad de los residuos es superior que para los individuos con mayor experiencia. Este comportamiento puede indicar que las perturbaciones no tengan la misma varianza para todos los individuos de la muestra, es decir, que es posible que no se cumpla la hipótesis básica de homocedasticidad de las perturbaciones.

# Solución PRÁCTICA P9.

## PRIMERA PARTE.

1. MRLG:

$$P_i = \beta_1 + \beta_2 H_i + \beta_3 B_i + \beta_4 S_i + u_i \qquad u_i \sim NID(0, \sigma_u^2) \quad i = 1, 2, \ldots, 265.$$

FRM:

$$\underset{(\hat{\sigma}_{\hat{\beta}})}{\widehat{P}_i} = \underset{(14{,}249)}{-\,10,968} + \underset{(2{,}66)}{0,655}\, H_i + \underset{(3{,}672)}{17,134}\, B_i + \underset{(0{,}041)}{0,304}\, S_i \qquad R^2 = 0,47943 \quad SCR = 234507.$$

2. Interpretación del coeficiente $\beta_4$:

   $\beta_4 = \dfrac{\partial E(P_i)}{\partial S_i}$ : variación esperada en el precio ante un cambio unitario en la superficie de la vivienda, manteniendo constante el resto de las variables.

   $\hat{\beta}_4 = 0,304$ miles de euros: se estima que un metro cuadrado de superficie adicional, manteniéndose el resto de las variables constantes, se valora en 3040 euros.

3. $R^2 = 1 - \dfrac{SCR}{SCT} = 1 - \dfrac{\sum \hat{u}_i^2}{\sum(P_i - \bar{P})^2} = 0,47943$

   Interpretación: el 47,943 % de la variabilidad del precio de venta de las viviendas en la muestra viene explicada por la variabilidad de las variables explicativas: número de habitaciones, número de baños y superficie.

4. Contraste de significatividad conjunta.

   $$H_0 : \beta_2 = \beta_3 = \beta_4 = 0 \qquad F = \frac{R^2/q}{(1-R^2)/(N-k)} \overset{H_0}{\sim} \mathcal{F}(q, N-k)$$
   $$H_a : \beta_2 \neq 0 \text{ y/o } \beta_3 \neq 0 \text{ y/o } \beta_4 \neq 0$$

   $$F = \frac{0,47943/3}{(1-0,47943)/(265-4)} = 80,1240 \; > \; 2,63919 = \mathcal{F}(3,261)_{0,05}$$

   Se rechaza la hipótesis nula con un nivel de significación del 5 %, es decir, se puede concluir que las variables explicativas habitaciones, baños y superficie son conjuntamente significativas.

5. Contraste de significatividad individual: número de habitaciones.

   $$H_0 : \beta_2 = 0 \qquad t = \frac{\hat{\beta}_2 - 0}{\hat{\sigma}_{\hat{\beta}_2}} \overset{H_0}{\sim} t(N-k)$$
   $$H_a : \beta_2 \neq 0$$

   $$|t| = \left| \frac{0,655 - 0}{2,660} \right| = 0,246 \; < \; 1,96909 = t(265-4)_{0,05/2}$$

   No se rechaza la hipótesis nula con un nivel de significación del 5 %, es decir, se puede concluir que el número de habitaciones no es una variable significativa para explicar el precio de venta, una vez incluidas en el modelo las variables superficie y número de baños.

6. Además de la variable explicativa "número de habitaciones", están incluidas las variables explicativas "número de baños" y "superficie de la vivienda", siendo ambas significativas ya que los estadísticos $t$ son superiores a $1,96909 = t_{(261)0,025}$.

Por otra parte, el coeficiente $\beta_2$ recoge la variación producida en el precio medio cuando la vivienda dispone de una habitación adicional manteniendo el resto de las características fijas, es decir, la superficie de la vivienda no varía ni tampoco el número de baños. Por tanto, el resultado obtenido en el contraste hace sospechar la presencia de un alto grado de multicolinealidad entre las tres variables explicativas, de forma que una vez incluidas las variables $B$ y $S$, la variable $H$ no aporta ninguna información adicional nueva relevante para determinar el precio de la vivienda. Para comprobar esta hipótesis habría que estudiar la matriz de correlaciones entre las variables explicativas.

7. Contraste de hipótesis (20000 euros = 20 miles de euros).

$$\begin{array}{l} H_0 : \beta_3 = 20 \\ H_a : \beta_3 \neq 20 \end{array} \qquad t = \frac{\hat{\beta}_3 - 20}{\hat{\sigma}_{\hat{\beta}_3}} \overset{H_0}{\sim} t(N-k)$$

$$|t| = \left| \frac{17,134 - 20}{3,672} \right| = 0,7805 \; < \; 1,96909 = t(265-4)_{0,05/2}$$

No se rechaza la hipótesis nula con un nivel de significación del 5 % y concluimos que, manteniendo el resto de las variables constantes, sí se está dispuesto a pagar 20000 euros por un cuarto de baño adicional.

8. En base a la recta de regresión muestral y dadas las características de la casa j, se tiene que:

$$\widehat{P}_j = -10,968 + 0,655 \times 4 + 17,134 \times 2 + 0,304 \times 100 = 56,32 \;\; \text{miles de euros.}$$

9. Se calcula el intervalo de confianza del precio de venta de una vivienda con las caracteríaticas citadas y se verifica si el precio pagado está o no dentro de este intervalo.

$$IC(P_j)_{1-\alpha} = \left[ \widehat{P}_j \pm t(N-k)_{\alpha/2} \; \hat{\sigma}_u \sqrt{1 + X_j (X'X)^{-1} X_j} \right]$$

$$\text{donde} \quad \widehat{P}_j = 56,32 \qquad \hat{\sigma}_u = \frac{234507}{265-4} = 29,9749 \qquad X_j = (1 \;\; 4 \;\; 2 \;\; 100)'$$

Utilizando el fichero de datos `practicaP9.gdt`, se obtiene que este intervalo es:

$$IC(P_j)_{95\,\%} = [-4,152 \quad 116,812]$$

Como $75 \in IC(P_j)_{95\,\%}$, el precio pagado por el amigo es un valor posible con una confianza del 95 % y se puede concluir que no hubo engaño.

10. Se observa que el comportamiento de los residuos no es uniforme u homogéneo. De hecho, los residuos de la primera parte y los de la última parte de la muestra son en su mayoría negativos, mientras que los del centro de la muestra son, en general, positivos. Por lo tanto, parece que alguna hipótesis básica del modelo no se cumple. Podría ser que se haya cometido un error en la especificación de la parte sistemática del modelo, por ejemplo, que se haya omitido alguna variable explicativa relevante.

## SEGUNDA PARTE.

1. Modelo P9.2: $P_i = \beta_1 + \beta_2 H_i + \beta_3 B_i + \beta_4 S_i + \beta_5 O_i + \beta_6 E_i + u_i \qquad i = 1, 2, \ldots, 265.$

   La localización es una variable explicativa cualitativa con tres categorías: oeste, este y centro. Por lo tanto, se definen tres variables ficticias, una por cada categoría:

$$O_i = \begin{cases} 1 & i \in \text{ zona oeste} \\ 0 & \text{en otro caso} \end{cases} \quad E_i = \begin{cases} 1 & i \in \text{ zona este} \\ 0 & \text{en otro caso} \end{cases} \quad C_i = \begin{cases} 1 & i \in \text{ zona centro} \\ 0 & \text{en otro caso} \end{cases}$$

   Para no tener problemas de colinealidad perfecta, el agente inmobiliario ha especificado el modelo introduciendo el término constante y tantas variables ficticias como categorías menos una: incluye las variables $O$ y $E$, dejando las viviendas del centro como referencia.

2.

| | 1ª obs. | 100ª obs. | 265ª obs. |
|---|---|---|---|
| $O$ | 0 | 0 | 1 |
| $E$ | 1 | 0 | 0 |

3. FRM:
$$\underset{(\hat\sigma_{\hat\beta})}{\widehat{P}_i} = \underset{(11{,}522)}{24,51} + \underset{(2{,}086)}{1,919}\, H_i + \underset{(2{,}902)}{16,414}\, B_i + \underset{(0{,}032)}{0,227}\, S_i - \underset{(3{,}63)}{43,851}\, O_i - \underset{(3{,}617)}{37,114}\, O_i$$

$$R^2 = 0,685011 \qquad SCR = 14189600$$

4. $\hat\beta_5$ = -43,851 miles de euros: la diferencia estimada en el precio de una vivienda en la zona oeste respecto a una vivienda en el centro, para las mismas características de superficie, número de baños y número de habitaciones.

   $\hat\beta_6$ = -37,114 miles de euros: la diferencia estimada en el precio de una vivienda en la zona este respecto a una vivienda en el centro, para las mismas características de superficie, número de baños y número de habitaciones.

5. La variable localización no es significativa si, a igualdad de condiciones, no hay diferencias en el precio promedio de la vivienda por estar en una zona o en otra.

$$H_0 : \beta_5 = \beta_6 = 0 \qquad H_a : \beta_5 \neq 0 \text{ y/o } \beta_6 \neq 0 \qquad F = \frac{(SCR_R - SCR_{NR})/q}{SCR_{NR}/(N-k)} \overset{H_0}{\sim} \mathcal{F}(q, N-k)$$

donde $SCR_{NR}$ es la suma de cuadrados de residuos del modelo no restringido, Modelo P9.2, $SCR_R$ es la suma de cuadrados de residuos del modelo restringido, Modelo P9.1, y $q = 2$ es el número de restricciones.

$$F = \frac{(234507 - 141896)/2}{141896/(265-6)} = 84,5202 \; > \; 3,03065 = \mathcal{F}(2,259)_{0,05}$$

Se rechaza la hipótesis nula con un nivel de significación del 5 %, es decir, se puede concluir que la localización es una variable significativa para explicar el precio de venta, una vez incluidas el resto de las variables.

6. Comparando este gráfico con los residuos del Modelo P9.1, se observa que los residuos del Modelo P9.2 oscilan en torno a cero y no se aprecia ya que haya una diferencia de comportamiento entre los residuos correspondientes a la primera parte de la muestra, los de la parte central y los de la última parte. Este resultado parece indicar que la introducción de la variable explicativa cualitativa localización (que es significativa) ha recogido el comportamiento no aletorio que presentaban los residuos del Modelo P9.1 y el problema se ha solucionado. Se puede concluir que el Modelo P9.1 estaba mal especificado por omisión de variable relevante.

7. Contraste de restricción lineal.

$$\begin{array}{ll} H_0 : \beta_5 = \beta_6 & (\beta_5 - \beta_6 = 0) \\ H_a : \beta_5 \neq \beta_6 & (\beta_5 - \beta_6 \neq 0) \end{array} \qquad t = \frac{\hat\beta_5 - \hat\beta_6 - 0}{\sqrt{\widehat{var}(\hat\beta_5 - \hat\beta_6)}} \overset{H_0}{\sim} t(N-k)$$

$$\widehat{var}(\hat\beta_5 - \hat\beta_5) = \widehat{var}(\hat\beta_5) + \widehat{var}(\hat\beta_6) - 2\widehat{cov}(\hat\beta_5, \hat\beta_6) = 13,18 + 13,08 - 2 \times 6,64 = 12,98$$

$$|t| = \left| \frac{-43,851 - (-37,114) - 0}{\sqrt{12,98}} \right| = |-1,86| \; < \; 1,96917 \approx t(265-6)_{0,05/2}$$

No se rechaza la hipótesis nula con un nivel de significación del 5 %, es decir, se puede concluir que, manteniendo el resto de las variables constantes, el efecto de la variable localización sobre el precio de la vivienda, diferencia sólo entre los pisos del centro y los restantes.

8. Sí, dado que el resultado del contraste del apartado anterior lleva a no rechazar la hipótesis nula $\beta_5 = \beta_6$, se propone estimar un modelo alternativo que incluya esta hipótesis. El modelo restringido sería:

$$\begin{aligned} P_i &= \beta_1 + \beta_2 H_i + \beta_3 B_i + \beta_4 S_i + \beta_5 O_i + \beta_6 E_i + u_i \quad \rightarrow \\ P_i &= \beta_1 + \beta_2 H_i + \beta_3 B_i + \beta_4 S_i + \beta_5 O_i + \beta_5 E_i + u_i \quad \rightarrow \\ P_i &= \beta_1 + \beta_2 H_i + \beta_3 B_i + \beta_4 S_i + \beta_5 (O_i + E_i) + u_i \end{aligned}$$

Estimar este Modelo Restringido por MCO es lo mismo que estimar el Modelo P9.2 sujeto a la restricción $\beta_5 = \beta_6$ por Mínimos Cuadrados Restringidos. Como la restricción es cierta, las propiedades de los estimadores por Mínimos Cuadrados Restringidos son las siguientes: lineales, insesgados y con menor varianza que los estimadores MCO del Modelo P9.2.

**TERCERA PARTE.**

1. Modelo P9.3: $P_i = \beta_1 + \beta_2\, H_i + \beta_3\, B_i + \beta_4\, S_i + \beta_5 \,\text{periferia}_i + \beta_6\, (\text{periferia}_i \times S_i) + u_i$.

   La localización ha pasado a ser una variable cualitativa con dos categorías: periferia (zona oeste y zona este) y centro. Además, se ha incluido en el modelo un nuevo término:
   $$\text{periferia}_i \times S_i = \begin{cases} S_i & i \in \text{ periferia} \\ 0 & \text{en otro caso} \end{cases}$$
   La variable localización aparece en el Modelo P9.3 de dos maneras diferentes: a través de la variable periferia y a través del término (periferia$\times S$). El nuevo término incorporado recoge una interacción entre superficie y localización que permite que el precio cambie debido a la acción conjunta de ambas variables, es decir, que el efecto del incremento de superficie de una vivienda sobre el precio puede no ser independiente de su localización.

   Por lo que se refiere al modelo planteado en el apartado 8 de la Segunda Parte, en él se incluye la variable localización pero sin el efecto de interacción con la variable superficie. Por lo tanto, en aquel modelo la influencia posible de la localización se recogería únicamente a través del coeficiente $\beta_5$.

2. La variación del precio de la vivienda ante un cambio en la superficie de un metro cuadrado, ceteris paribus, es:
   $$\frac{\partial E(P_i)}{\partial S_i} = \beta_4 + \beta_6 \ \text{periferia}_i = \begin{cases} \beta_4 & \text{si} \quad \text{periferia} = 0 \\ \beta_4 + \beta_6 & \text{si} \quad \text{periferia} = 1 \end{cases}$$
   Nótese que esta variación no es igual para todas las zonas.

   Por lo tanto, la variación estimada en el precio ante un incremento de un metro cuadrado en la superficie es:

   Piso en el centro $\Rightarrow$ $\hat{\beta}_4 = 0,297$ miles de euros
   Piso en la periferia $\Rightarrow$ $(\hat{\beta}_4 + \hat{\beta}_6) = 0,297 - 0,114 = 0,183$ miles de euros

3. El efecto de la localización está asociado a los coeficientes $\beta_5$ y $\beta_6$:
   $$E(P_i \mid \text{centro}) = \beta_1 + \beta_2 H_i + \beta_3 B_i + \beta_4 S_i$$
   $$E(P_i \mid \text{periferia}) = \beta_1 + \beta_2 H_i + \beta_3 B_i + \beta_4 S_i + \beta_5 + \beta_6 S_i.$$

$$H_0 : \beta_5 = \beta_6 = 0 \qquad F = \frac{(SCR_R - SCR_{NR})/q}{SCR_{NR}/(N-k)} \overset{H_0}{\sim} \mathcal{F}(q, N-k)$$
$$H_a : \beta_5 \neq 0 \quad y/o \quad \beta_6 \neq 0$$

El modelo restringido es el Modelo P9.1 y el modelo no restringido es el Modelo P9.3.

$$F = \frac{(234507 - 141000)/2}{141000/(265-6)} = 85,88 \; > \; 3,03065 = \mathcal{F}(2,259)_{0,05}$$

Se rechaza la hipótesis nula a un nivel de signifcación $\alpha$ =5 %, es decir, se puede concluir que el efecto de la localización es significativo para explicar el precio de las viviendas.

4. De acuerdo con el Modelo P9.3, un metro cuadrado adicional de superficie cuesta

$$\frac{\partial E(P_i)}{\partial S_i} = \beta_4 + \beta_6 \,\text{periferia}_i$$

Si $\beta_6 = 0$, un metro cuadrado adicional cuesta lo mismo con independencia del lugar en el que se encuentre la vivienda.

$$H_0 : \beta_6 = 0 \qquad t = \frac{\hat{\beta}_2 - 0}{\hat{\sigma}_{\hat{\beta}_2}} \overset{H_0}{\sim} t(N-k)$$
$$H_a : \beta_6 \neq 0$$

$$|t| = \left| \frac{-0,114 - 0}{0,051} \right| = |-2,230| > 1,96917 = t(265-6)_{0,05/2}$$

Se rechaza la hipótesis nula a un nivel de signifcación $\alpha$ =5 %, es decir, se puede concluir que el precio del metro cuadrado adicional no es el mismo en el centro que en la periferia.

5. El modelo que debemos elegir es el que mejor represente las variaciones de la variable precio, de acuerdo con la información disponible. Comparando los dos primeros modelos, comprobamos que el Modelo P9.1 no incorpora la variable localización que es significativa. El modelo Modelo P9.1 se rechaza porque omite una variable relevante, la localización, lo que supone que las estimaciones de los coeficientes y de la varianza de las perturbaciones están sesgadas y los contrastes no son válidos porque los estadísticos $t$ y $F$ no siguen las distribuciones adecuadas.
En el Modelo P9.2 se observa que la variable localización es significativa, aunque la manera de incluirla en el modelo no es la mejor, debido a que se ignora una restricción cierta (que las viviendas de la periferia tienen el mismo efecto en el precio por su localización). A priori, parece que si se incluye la restricción, para conseguir mayor eficiencia en la estimación, es un modelo correcto, ya que no hay evidencias claras de mala especificación.
No obstante, cuando en el Modelo P9.3 se incluye un efecto de interacción entre la localización y la superficie de la vivienda, este efecto resulta ser significativo, por lo

que también en el modelo Modelo P9.2 se ha cometido un error de especificación en la forma funcional, con las consecuencias de estimaciones sesgadas para los coeficientes, estimación sesgada de la varianza de las perturbaciones y contrastes que no son válidos.
Por lo tanto, de los tres modelos, el elegido debería ser el Modelo P9.3.

# Solución PRÁCTICA P10.

## PRIMERA PARTE.

1. Estimación MCO.

$$\hat{\beta}_2 = \frac{\sum_{i=1}^{N}(C_i - \bar{C})(R_i - \bar{R})}{\sum_{i=1}^{N}(R_i - \bar{R})^2} = \frac{880636,5}{6041478} = 0,145765$$

$$\hat{\beta}_1 = \bar{C} - \hat{\beta}_2\,\bar{R} = \frac{7662}{40} - 0,145765 \times \frac{17170}{40} = 128,98$$

Recta de regresión muestral:

$$\widehat{C}_i = 128,98 + 0,145765\,R_i \qquad i = 1, 2, \ldots, 40.$$

2. $\hat{\beta}_2 := 0{,}145765$: incremento estimado en el consumo de pizza, en euros, cuando la renta aumenta en una unidad, es decir, en mil euros.

3. Coeficiente de correlación muestral:

$$r_{CR} = \frac{\sum_{i=1}^{40}(C_i - \bar{C})(R_i - \bar{R})}{\sqrt{\sum_{i=1}^{40}(C_i - \bar{C})^2\ \sum_{i=1}^{40}(R_i - \bar{R})^2}} = \frac{880636,5}{\sqrt{947651,9 \times 6041478}} = 0,368045$$

Relación entre $r_{CR}$ y el coeficiente de determinación $R^2$:

$$R^2 = \frac{SCE}{SCT} = \frac{\sum_{i=1}^{40}(\hat{C}_i - \bar{C})^2}{\sum_{i=1}^{40}(C_i - \bar{C})^2} = \frac{\sum_{i=1}^{40}(\hat{\beta}_1 + \hat{\beta}_2\,R_i - \bar{C})^2}{\sum_{i=1}^{40}(C_i - \bar{C})^2} =$$

$$= \frac{\sum_{i=1}^{40}[(\bar{C} - \hat{\beta}_2\,\bar{R}) + \hat{\beta}_2\,R_i - \bar{C}]^2}{\sum_{i=1}^{40}(C_i - \bar{C})^2} = \frac{\hat{\beta}_2^2\ \sum_{i=1}^{40}(R_i - \bar{R})^2}{\sum_{i=1}^{40}(C_i - \bar{C})^2} =$$

$$= \frac{\left(\frac{\sum_{i=1}^{40}(C_i-\bar{C})(R_i-\bar{R})}{\sum_{i=1}^{40}(R_i-\bar{R})^2}\right)^2 \sum_{i=1}^{40}\,(R_i - \bar{R})^2}{\sum_{i=1}^{40}(C_i - \bar{C})^2} = \frac{\left(\sum_{i=1}^{40}(C_i - \bar{C})(R_i - \bar{R})\right)^2}{\sum_{i=1}^{40}(C_i - \bar{C})^2\ \sum_{i=1}^{40}(R_i - \bar{R})^2} =$$

$$= \left(\frac{\sum_{i=1}^{40}(C_i - \bar{C})(R_i - \bar{R})}{\sqrt{\sum_{i=1}^{40}(C_i - \bar{C})^2\ \sum_{i=1}^{40}(R_i - \bar{R})^2}}\right)^2 = r_{CR}^2 = (0,368045)^2 = 0,135457$$

Interpretación: el 13,5457 % de la variabilidad del consumo de pizza en la muestra viene explicada por la variabilidad de variable explicativa renta.

4. Para que el estimador MCO sea insesgado se han de cumplir las siguientes hipótesis básicas:

   - El modelo está bien especificado en el sentido de que incluye todas las variables explicativas relevantes, y todas las variables explicativas incluidas son relevantes.
   - Modelo lineal en los coeficientes.
   - Coeficientes constantes.
   - Regresores fijos.
   - Ausencia de colinealidad perfecta.
   - Media de la perturbación cero.

5. Modelo P10.2:

   5.1. Ambos modelos relacionan las variables consumo de pizza y renta. La diferencia entre ambos estriba en la dirección de la causalidad: en el Modelo P10.1 se desea explicar la variabilidad del consumo en función de la variable explicativa renta, y en el Modelo P10.2, la causalidad va en la otra dirección, explicando la variabilidad de la renta en función del consumo.

   5.2. La estimación MCO de $\alpha_2$:

$$\hat{\alpha}_2 = \frac{\sum_{i=1}^{N}(C_i - \bar{C})(R_i - \bar{R})}{\sum_{i=1}^{N}(C_i - \bar{C})^2} = \frac{880636,5}{947651,9} = 0,929282683$$

   La estimación de $\alpha_2$ no coincide con la de $\beta_2$, porque no están midiendo lo mismo:

$\hat{\beta}_2$ : incremento estimado en el consumo de pizza, en euros, cuando la renta aumenta en una unidad (mil euros).

$\hat{\alpha}_2$ : incremento estimado en la renta, en miles de euros, cuando el consumo de pizza aumenta en una unidad (1 euros).

5.3. Los coeficientes de determinación de ambos modelos son iguales. En el apartado 3 se ha demostrado que en el modelo de regresión lineal simple el coeficiente de determinación es igual al cuadrado del coeficiente de correlación entre el regresor y la variable dependiente, por lo tanto:

$$R^2_{P10.1} = r^2_{CR} \quad \text{y} \quad R^2_{P10.2} = r^2_{CR} \quad \Rightarrow \quad R^2_{P10.1} = R^2_{P10.2}$$

## SEGUNDA PARTE.

1. MRLG:

$$C_i = \beta_1 + \beta_2\, R_i + \beta_3\, E_i + \beta_4\, M_i + \beta_5\, EP_i + \beta_6\, ES_i + \beta_7\, EU_i + u_i$$

$$u_i \sim NID(0, \sigma_u^2) \quad i = 1, \ldots, 40$$

En este modelo el consumo de pizza se explica en función de cuatro variables explicativas: renta, edad, género y nivel de estudios. Dos de las variables explicativas, género y nivel de estudios, son cualitativas por lo que han de introducirse en el modelo a través de variables ficticias.

La variable género tiene dos categorías por lo que se pueden definir dos variables ficticias, una para los hombres y otra para las mujeres. La variable nivel de estudios tiene cuatro categorías, luego se construyen cuatro variables ficticias, una por cada categoría. Por último para evitar problemas de colinealidad exacta en la matriz de datos $X$ se ha optado por especificar el modelo con un término constante incluyendo tantas variables ficticas como número de categorías menos una para cada variable cualitativa. Para la variable género ha quedado fuera la categoría hombre y para la variable nivel de estudios la categoría sin estudios. Por lo tanto, los hombres sin estudios se convierten en la clase de referencia para interpretar los coeficientes.

2. FRM:

$$\begin{aligned}
\underset{(\hat{\sigma}_{\hat{\beta}})}{\widehat{C}_i} &= \underset{(39{,}0137)}{280,735} + \underset{(0{,}0428834)}{0,283674}\; R_i - \underset{(1{,}49434)}{8,8409}\; E_i - \underset{(25{,}8175)}{181,85}\; M_i + \underset{(37{,}1903)}{73,1179}\; EP_i - \\
&- \underset{(40{,}2046)}{6,4715}\; ES_i - \underset{(59{,}1766)}{48,752}\; EU_i \qquad R^2 = 0,772621 \qquad SCR = 215476,3
\end{aligned}$$

$\hat{\beta}_1$ = 280,735: consumo medio estimado de pizza para un hombre de 18 años sin estudios y sin renta.

$\hat{\beta}_2$ = 0,283674: incremento estimado en el consumo de pizza en euros, cuando la renta aumenta en mil euros, manteniendo la edad, el género y el nivel de estudios constantes.

$\hat{\beta}_3$ = -8,8409: si la edad del individuo aumenta en un año, manteniendo el resto de las variables (renta, género y nivel de estudios) constantes, se estima que el consumo de pizza disminuye en 8,8409 euros.

$\hat{\beta}_4$ = -181,85: diferencia estimada en el consumo de pizza en euros entre una mujer y un hombre a igualdad de condiciones de renta, edad y nivel de estudios.

$\hat{\beta}_5$ = 73,1179: diferencia estimada en el consumo de pizza en euros entre una persona con estudios primarios y una persona sin estudios, a igualdad de condiciones de renta, edad y género.

$\hat{\beta}_6$ = -6,4715: se estima que una persona con estudios secundarios consume en pizza 6,4719 euros menos que una sin estudios, a igualdad de condiciones de renta, edad y género.

$\hat{\beta}_7$ = -48,752: el consumo de pizza medio estimado para una persona con estudios universitarios es menor en 48,752 euros que el de una persona sin estudios, a igualdad de condiciones de renta, edad y género.

3. No, porque la recta de regresión poblacional se basa en los verdaderos valores de los coeficientes de regresión para la población, y la recta de regresión muestral se basa en los valores estimados de los coeficientes en base a una muestra de la población.

4. 280,735 - 181,85 = $\hat{\beta}_1 + \hat{\beta}_4$: consumo medio estimado de pizza para una mujer de 18 años sin estudios y sin renta.

5. $\widehat{C}_i = 280,735 + 0,283674\, R_i - 8,8409\, E_i - 181,85 - 48,752$ euros

6. Contraste de significatividad individual: variable género.

$$\begin{array}{l} H_0 : \beta_4 = 0 \\ H_a : \beta_4 \neq 0 \end{array} \qquad t = \frac{\hat{\beta}_4 - 0}{\hat{\sigma}_{\hat{\beta}_4}} \overset{H_0}{\sim} t(N-k)$$

$$|t| = \left| \frac{-181,85 - 0}{25,8175} \right| = |-7,0438| \; > \; 2,03452 = t(40-7)_{0,05/2}$$

Se rechaza la hipótesis nula con un nivel de significación del 5 %, es decir, se puede concluir que el género es una variable significativa para explicar el consumo de pizza.

7. Contraste de hipótesis.

$$\begin{array}{l} H_0 : \beta_4 = -200 \\ H_a : \beta_4 \neq -200 \end{array} \qquad t = \frac{\hat{\beta}_4 - 0}{\hat{\sigma}_{\hat{\beta}_4}} \overset{H_0}{\sim} t(N-k)$$

$$|t| = \left| \frac{-181,85 + 200}{25,8175} \right| = \ 0,703 < \ 2,03452 = t(40-7)_{0,05/2}$$

No se rechaza la hipótesis nula con un nivel de significación del 5 %, es decir, se puede concluir que las mujeres consumen 200 euros menos de pizza que los hombres a igualdad de condiciones.

8. La variable nivel de estudios no es significativa si, a igualdad de condiciones, no hay diferencias en el consumo medio de pizza por tener un nivel de estudios u otro.

$H_0 : \beta_5 = \beta_6 = \beta_7 = 0$
$H_a : \beta_5 \neq 0$ y/o $\beta_6 \neq 0$ y/o $\beta_7 \neq 0$

$$F = \frac{(SCR_R - SCR_{NR})/q}{SCR_{NR}/(N-k)} \overset{H_0}{\sim} \mathcal{F}(q, N-k)$$

donde $q = 3$ es el número de restricciones, $SCR_{NR}$ es la suma de cuadrados de residuos del modelo no restringido, Modelo P10.2, $SCR_R$ es la suma de cuadrados de residuos del siguiente modelo restringido:

$$C_i = \beta_1 + \beta_2\, R_i + \beta_3\, E_i + \beta_4\, M_i + u_i$$

$$F = \frac{(275888,8 - 215476,3)/3}{215476,3/(40-7)} = 3,08404 \ > \ 2,89156 = \mathcal{F}(3, 40-7)_{0,05}$$

Se rechaza la hipótesis nula con un nivel de significación del 5 %, es decir, se puede concluir que el nivel de estudios es una variable significativa para explicar el consumo de pizza, una vez incluidas el resto de las variables.

## TERCERA PARTE.

1. Modelo P10.3:

$$C_i = \beta_1 + \beta_2\, R_i + \beta_3\, E_i + \beta_4\, M_i + \beta_5\, EP_i + \beta_6\, ES_i + \beta_7\, EU_i + \beta_8\, (R_i \times M_i) + u_i \quad i = 1, \ldots, 40.$$

Sí, porque es un modelo lineal en los coeficientes, que es la hipótesis básica del MRLG. Ahora bien el modelo no es lineal en las variables porque presenta un efecto interacción entre las variables explicativas renta y género.

2.

$$X = \begin{array}{c} \begin{array}{cccccccc} \phantom{1} & R & E & M & EP & ES & EU & R \times M \end{array} \\ \begin{bmatrix} 1 & 15 & 27 & 1 & 0 & 0 & 0 & 15 \\ 1 & 30 & 2 & 1 & 0 & 0 & 0 & 30 \\ 1 & 12 & 10 & 1 & 0 & 0 & 0 & 12 \\ 1 & 20 & 7 & 1 & 0 & 0 & 0 & 20 \\ 1 & 15 & 17 & 1 & 1 & 0 & 0 & 15 \\ 1 & 30 & 22 & 1 & 1 & 0 & 0 & 30 \\ 1 & 12 & 4 & 1 & 1 & 0 & 0 & 12 \\ 1 & 12 & 12 & 1 & 1 & 0 & 0 & 12 \\ 1 & 28 & 3 & 1 & 1 & 0 & 0 & 28 \\ 1 & 22 & 22 & 1 & 0 & 1 & 0 & 22 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \end{bmatrix} \end{array}$$

3. Todas las personas de la submuestra formada por las diez primeras observaciones son mujeres. Por lo tanto, la matriz de datos correspondiente presenta colinealidad perfecta, ya que la columna de la variable $M$ es una columna de unos, igual a la del término constante. Además, en este caso, se daría también que $R_i \times M_i = R_i$, lo que genera una segunda combinación lineal entre las columnas de la matrix $X$. En consecuencia, el rango de la matriz $X$ no es completo y no se puede invertir la matriz $(X'X)^{-1}$ con lo que no se pueden estimar los coeficientes del modelo.

4. El efecto de la renta sobre el consumo medio es:

$$\frac{\partial\, E(C_i)}{\partial R_i} = \beta_2 + \beta_8\, M_i = \begin{cases} \beta_2 + \beta_8 & i \in \text{Mujer} \\ \beta_2 & i \in \text{Hombre} \end{cases}$$

Nótese que este efecto no es constante para toda la muestra ya que es distinto para los hombres y las mujeres debido a la introducción en el Modelo P10.3 del término de interacción entre la renta y el género. Específicamente, el coeficiente $\beta_8$ mide la diferencia en el incremento del consumo ante aumentos unitarios (mil euros) en la renta entre una mujer y un hombre manteniendo constante la edad y el nivel de educación.

$$\begin{array}{l} H_0 : \beta_8 = 0 \\ H_a : \beta_8 \neq 0 \end{array} \qquad t = \frac{\hat{\beta}_8 - 0}{\hat{\sigma}_{\hat{\beta}_8}} \overset{H_0}{\sim} t(N-k)$$

$$|t| = \left| \frac{-0,152513 - 0}{0,0684368} \right| = |-2,2285| \;>\; 2,03693 = t(40-8)_{0,05/2}$$

Se rechaza la hipótesis nula con un nivel de significación del 5 %, es decir, se puede concluir que el efecto de la renta sobre el consumo de pizza no es constante para toda la muestra, en concreto, es distinto para mujeres y para hombres.

5. Consumo estimado de una universitaria ($M = 1$, $EU = 1$):

$$\begin{aligned}\widehat{C}_i &= 236,640\, R_i - 8,9736\, E_i - 115,72 - 56,106 - 0,00152513\, R_i = \\ &= 236,640 + (0,395667 - 0,152513)\, R_i - 8,9736\, E_i - 115,72 - 56,106 = \\ &= 64,814 + 0,243154\, R_i - 8,9736\, E_i \text{ euros}\end{aligned}$$

6. La variable renta no es significativa si no afecta al consumo de pizza:

$$\begin{array}{l} H_0 : \beta_2 = \beta_8 = 0 \\ H_a : \beta_2 \neq 0 \text{ y/o } \beta_8 \neq 0 \end{array} \qquad F = \frac{(SCR_R - SCR_{NR})/q}{SCR_{NR}/(N-k)} \overset{H_0}{\sim} \mathcal{F}(q, N-k)$$

donde $q = 2$ es el número de restricciones, $SCR_{NR}$ es la suma de cuadrados de residuos del modelo no restringido, Modelo P10.3, $SCR_R$ es la suma de cuadrados de residuos del siguiente modelo restringido:

$$C_i = \beta_1 + \beta_3\, E_i + \beta_4\, M_i + \beta_5\, EP_i + \beta_6\, ES_i + \beta_7\, EU_i + u_i$$

$$F = \frac{(501199,2 - 186527,9)/2}{186527,9/(40-8)} = 26,9919 > 3,29454 = \mathcal{F}(2, 40-8)_{0,05}$$

Se rechaza la hipótesis nula con un nivel de significación del 5 %, es decir, se puede concluir que la renta es una variable significativa para explicar el consumo de pizza, una vez incluidas el resto de las variables.

## CUARTA PARTE.

1. Modelo P10.4:

$$C_i = \beta_1 + \beta_2\, R_i + \beta_3\, E_i + \beta_4\, M_i + \beta_5\, EP_i + \beta_6\, ES_i + \beta_7\, EU_i + \beta_8\, (R_i \times E_i) + u_i$$

La interacción entre las variables renta, $R$, y edad, $E$, se introduce en el modelo a través del término $(R_i \times E_i)$, que es capaz de recoger la hipótesis de trabajo planteada:

$$\frac{\partial E(C_i)}{\partial R_i} = \beta_2 + \beta_8\, E_i \qquad i = 1, 2, \ldots, 40.$$

Nótese que el efecto marginal de la renta sobre el consumo de pizza depende de la edad.

1.1. Incremento esperado en el consumo de pizza, ceteris paribus:

$\Delta R = 500$ euros : $\quad \Delta E(C) = 0,5 \times (\beta_2 + \beta_8\, E_i)$

$\Delta R = 1000$ euros : $\quad \Delta E(C) = \beta_2 + \beta_8\, E_i$

1.2. Incremento esperado en el consumo de pizza, ceteris paribus:

$\Delta R = 1000$ euros, $E_i = (20 - 18)$ años : $\quad \Delta E(C) = \beta_2 + \beta_8 \times 2$ euros

$\Delta R = 1000$ euros, $E_i = (40 - 18)$ años : $\quad \Delta E(C) = \beta_2 + \beta_8 \times 22$ euros

2. Recta de regresión muestral:

$$\begin{aligned} \widehat{C}_i &= \underset{(50{,}006)}{236,534} + \underset{(0{,}1536)}{0,4880}\, R_i - \underset{(2{,}2145)}{6,5534}\, E_i - \underset{(26{,}282)}{172,861}\, M_i + \underset{(37{,}831)}{60,327}\, EP_i - \\ &- \underset{(42{,}039)}{25,778}\, ES_i - \underset{(60{,}306)}{69,724}\, EU_i - \underset{(0{,}005961)}{0,008251}\, (R \times E)_i \quad i = 1, 2, \ldots, 40 \end{aligned}$$

$$R^2 = 0,785465 \quad SCR = 203304,1$$

2.1. Incremento estimado en el consumo de pizza, ceteris paribus:

$$\Delta R = 500 \text{ euros}: \quad \Delta \widehat{E(C)} = 0,5 \times (\hat\beta_2 + \hat\beta_8\, E_i)$$

$$\Delta \widehat{E(C)} = 0,5 \times (0,4880 - 0,008251\, E_i)$$

$$\Delta \widehat{E(C)} = 0,2440 - 0,004125\, E_i$$

$$\Delta R = 1000 \text{ euros}: \quad \widehat{\Delta} E(C) = \hat\beta_2 + \hat\beta_8\, E_i$$

$$\Delta \widehat{E(C)} = 0,4880 - 0,008251\, E_i$$

2.2. Incremento estimado en el consumo de pizza, ceteris paribus:

$\Delta R = 1000$ euros, $E_i = (20 - 18)$ años :

$$\Delta \widehat{E(C)} = \hat\beta_2 + \hat\beta_8 \times 2 = 0,4880 - 0,008251 \times 2 \quad \text{euros}$$

$\Delta R = 1000$ euros, $E_i = (40 - 18)$ años :

$$\Delta \widehat{E(C)} = \hat\beta_2 + \hat\beta_8 \times 22 = 0,4880 - 0,008251 \times 22 \quad \text{euros}$$

3. Contraste de la hipótesis de trabajo.

$$\begin{array}{l} H_0 : \beta_8 = 0 \\ H_a : \beta_8 \neq 0 \end{array} \qquad t = \frac{\hat\beta_8 - 0}{\hat\sigma_{\hat\beta_8}} \overset{H_0}{\sim} t(N-k)$$

$$|t| = \left| \frac{-0,008251 - 0}{0,005961} \right| = 1,384 < 2,03693 = t(265-4)_{0{,}05/2}$$

No se rechaza la hipótesis nula con un nivel de significación del 5 %, es decir, se puede concluir que no hay evidencia en la muestra a favor de la hipótesis de que el efecto de un incremento de la renta sobre el consumo de pizza depende de la edad del individuo.

4. La variable renta es significativa si no afecta al consumo de pizza:

$$H_0 : \beta_2 = \beta_8 = 0 \qquad F = \frac{(SCR_R - SCR_{NR})/q}{SCR_{NR}/(N-k)} \overset{H_0}{\sim} \mathcal{F}(q, N-k)$$
$$H_a : \beta_2 \neq 0 \text{ y/o } \beta_8 \neq 0$$

donde $q = 2$ es el número de restricciones, $SCR_{NR}$ es la suma de cuadrados de residuos del modelo no restringido, Modelo P10.4, $SCR_R$ es la suma de cuadrados de residuos del siguiente modelo restringido:

$$C_i = \beta_1 + \beta_3\, E_i + \beta_4\, M_i + \beta_5\, EP_i + \beta_6\, ES_i + \beta_7\, EU_i + u_i$$

$$F = \frac{(501199,2 - 203304,1)/2}{203304,1/(40-8)} = 23,4443 \;>\; 3,29454 = \mathcal{F}(2, 40-8)_{0,05}$$

Se rechaza la hipótesis nula con un nivel de significación del 5 %, es decir, se puede concluir que la renta es una variable significativa para explicar el consumo de pizza, una vez incluidas el resto de las variables.

## QUINTA PARTE.

1. Modelo P10.5:

$$C_i = \beta_1 + \beta_2\, R_i + \beta_3\, E_i + \beta_4\, M_i + \beta_5\, EP_i + \beta_6\, ES_i + \beta_7\, EU_i + \beta_8\, R_i^2 + u_i \quad i = 1, \ldots, 40.$$

Que el efecto de un incremento de la renta sobre el consumo de pizza dependa del nivel de renta del individuo significa que la relación entre la renta y el consumo no es lineal. En el Modelo P10.5 se propone una relación cuadrática entre la renta y el consumo de pizza que es capaz de recoger la hipótesis de trabajo propuesta.

2. El incremento esperado en el consumo de pizza, ceteris paribus, no es constante, sino que depende del nivel de renta del individuo:

$$\frac{\partial E(C_i)}{\partial R_i} = \beta_2 + 2\,\beta_8\, R_i$$

3. Recta de regresión muestral:

$$\begin{aligned} \hat{C}_i \;=\;& \underset{(41{,}996)}{250,494} + \underset{(0{,}1341)}{0,4991}\, R_i - \underset{(1{,}6429)}{10,1336}\, E_i - \underset{(26{,}2457)}{169,006}\, M_i + \underset{(36{,}4712)}{65,4304}\, EP_i - \\ -\;& \underset{(42{,}2678)}{33,5246}\, ES_i - \underset{(66{,}4836)}{104,923}\, EU_i - \underset{(0{,}000056)}{0,000095}\, R_i^2 \quad i = 1, 2, \ldots, 40 \end{aligned}$$

$$R^2 = 0,791253 \qquad SCR = 197819,8$$

4. Contraste de la hipótesis de trabajo.

$$H_0 : \beta_8 = 0 \qquad t = \frac{\hat{\beta}_8 - 0}{\hat{\sigma}_{\hat{\beta}_8}} \overset{H_0}{\sim} t(N-k)$$
$$H_a : \beta_8 \neq 0$$

$$|t| = \left| \frac{-0,000095 - 0}{0,000056} \right| = 1,69 \; < \; 2,03693 = t(40-8)_{0,05/2}$$

No se rechaza la hipótesis nula con un nivel de significación del 5 %, es decir, se puede concluir que no hay evidencia en la muestra a favor de la hipótesis de que el efecto de un incremento de la renta sobre el consumo de pizza depende del nivel de renta del individuo.

5. La variable renta es significativa si no afecta al consumo de pizza:

$$H_0 : \beta_2 = \beta_8 = 0 \qquad F = \frac{(SCR_R - SCR_{NR})/q}{SCR_{NR}/(N-k)} \overset{H_0}{\sim} \mathcal{F}(q, N-k)$$
$$H_a : \beta_2 \neq 0 \text{ y/o } \beta_8 \neq 0$$

donde $q = 2$ es el número de restricciones, $SCR_{NR}$ es la suma de cuadrados de residuos del modelo no restringido, Modelo P10.5, $SCR_R$ es la suma de cuadrados de residuos del siguiente modelo restringido:

$$C_i = \beta_1 + \beta_3\, E_i + \beta_4\, M_i + \beta_5\, EP_i + \beta_6\, ES_i + \beta_7\, EU_i + u_i$$

$$F = \frac{(501199,2 - 197819,8)/2}{197819,8/(40-8)} = 24,5378 \; > \; 3,29454 = \mathcal{F}(2, 40-8)_{0,05}$$

Se rechaza la hipótesis nula con un nivel de significación del 5 %, es decir, se puede concluir que la renta es una variable significativa para explicar el consumo de pizza, una vez incluidas el resto de las variables.

# Solución PRÁCTICA P11.

## PRIMERA PARTE.

1. Sí porque el modelo es lineal en los coeficientes (aunque no en las variables), no existe colinealidad perfecta ($\sum_{t=1}^{29} (\log P_t - \overline{\log P})^2 \neq 0$) y, con la información disponible, el modelo esta bien especificado. Además, la perturbación sigue una distribución normal de media cero y varianza constante, $\sigma_u^2$. Por último, la covarianza entre pertubaciones distintas es cero, dado que son independientes.

2. Recta de regresión poblacional:

$$E(\log C_t) \;=\; \beta_1 \;+\; \beta_2 \log P_t + \beta_3 \log R_t \qquad t = 1960, 1961, \ldots, 1988.$$

$\beta_2$ : Incremento porcentual en el consumo de cigarrillos cuando el precio aumenta un 1 %, manteniéndose la renta constante.

$\beta_3$ : Incremento porcentual en el consumo de cigarrillos cuando la renta aumenta un 1 %, manteniéndose el precio constante.

Como el modelo es doble-logarítmico, estos dos coeficientes, $\hat\beta_1$ y $\hat\beta_2$, son las elasticidades estimadas precio y renta, respectivamente.

3.

$$X = \begin{bmatrix} 1 & \log P_{1960} & \log R_{1960} \\ 1 & \log P_{1961} & \log R_{1961} \\ 1 & \log P_{1962} & \log R_{1962} \\ 1 & \log P_{1963} & \log R_{1963} \\ \vdots & \vdots & \vdots \\ 1 & \log P_{1988} & \log R_{1988} \end{bmatrix} \qquad X'Y = \begin{bmatrix} \sum \log C_t \\ \sum \log C_t \, \log P_t \\ \sum \log C_t \, \log R_t \end{bmatrix}$$

$$X'X = \begin{bmatrix} T & \sum \log P_t & \sum \log R_t \\ \sum \log P_t & \sum \log P_t^2 & \sum \log P_t \, \log R_t \\ \sum \log R_t & \sum \log P_t \, \log R_t & \sum \log R_t^2 \end{bmatrix}$$

4. Recta de regresión muestral:

$$\widehat{\log C_t} \;=\; \underset{(0{,}724913)}{-4,58987} - \underset{(0{,}101394)}{0,485683} \;\log P_t + \underset{(0{,}0947276)}{0,688498} \;\log R_t$$

$$R^2 = 0,712058 \qquad\qquad \sum \hat u_t^2 = 0,094911$$

5. Contraste de significatividad individual: precio del tabaco.

$$\begin{array}{l} H_0 : \beta_2 = 0 \\ H_a : \beta_2 \neq 0 \end{array} \qquad t = \frac{\hat\beta_2 - 0}{\hat\sigma_{\hat\beta_2}} \overset{H_0}{\sim} t(T-k)$$

$$|t| = \left| \frac{-0,485683 - 0}{\sqrt{0,101394}} \right| = 4,79 \;>\; 2,05553 = t(29-3)_{0{,}05/2}$$

Se rechaza la hipótesis nula con un nivel de significación del 5 %, es decir, existe evidencia en la muestra de que el precio influye en el consumo medio de tabaco.

6. Para que la inferencia en el Modelo P11.1 sea válida, tal y como se ha realizado en el apartado anterior, se han de cumplir todos los supuestos del modelo de regresión lineal:
   - Modelo bien especificado.
   - Modelo lineal en los coeficientes.
   - Coeficientes constantes.
   - Regresores fijos.
   - Ausencia de colinealidad perfecta.
   - Media de la perturbación cero: $E(u_t) = 0 \; \forall t$.
   - Varianza de la perturbación constante: $V(u_t) = \sigma_u^2 \; \forall t$.
   - No existencia de autocorrelación en las perturbaciones : $E(u_t \, u_s) = 0 \; \forall t \neq s$.
   - La perturbación sigue una distribución normal.

7. Los residuos de la estimación MCO de un modelo de regresión lineal que estuviera bien especificado, es decir, que cumpliera todos los supuestos, se deberían comportar de forma similar a las perturbaciones: oscilando en torno a cero con varianza constante y de forma aleatoria, es decir, sin ningún patrón de comportamiento reconocible.

   Los residuos del gráfico que corresponden a la estimación MCO del Modelo P11.1, oscilan en torno a cero pero no de forma aleatoria ya que se observa que los residuos siguen una tendencia, es decir, un comportamiento promedio creciente hasta el año 80, y a partir de los 80 decrecen bruscamente siendo casi todos negativos. Este resultado indica que algo está fallando en la especificación del modelo. Podría ser debido a que no se ha incluido en el modelo algún factor relevante; factor que fuera capaz de explicar el brusco cambio de comportamiento de los residuos a partir de los años 80.

## SEGUNDA PARTE.

1. Modelo P11.2:

$$\log C_t = \alpha_1 + \alpha_2 \log P_t + \alpha_3 \log R_t + \alpha_4 D82_t + \alpha_5 D86_t + u_t \quad t = 1960, 1961, \ldots, 1988.$$

   En el Modelo P11.2 se han introducido dos variables explicativas nuevas de carácter cualitativo para recoger el efecto sobre el consumo de cada una de las dos campañas publicitarias que vamos a denominar, por ejemplo, Campaña del 82 y Campaña del 86. Como estas variables explicativas son cualitativas se han de introducir en el modelo mediante variables ficticias.

La variable cualitativa Campaña del 82 tiene dos categorías: no hay campaña (antes del 82), sí hay campaña (después del 82) por lo que se definen dos variables ficticias:

$$ND82_t = \begin{cases} 1 & t \in [1960 - 1981] \\ 0 & t \in [1982 - 1988] \end{cases} \qquad D82_t = \begin{cases} 1 & t \in [1982 - 1988] \\ 0 & t \in [1960 - 1981] \end{cases}$$

Para evitar problemas de multicolinealidad perfecta en la matriz $X$, a la hora de especificar el modelo se introduce el término constante y tantas variables ficticias como número de categorías menos una. Es decir, en este caso, sólo se introduce una variable ficticia, $D82$.
De forma similar, se introduce la variable explicativa cualitativa Campaña del 86, a través de la variable ficticia $D86$.

2. Consumo promedio de cigarrillos:

$$\begin{aligned} 1960 - 1981: \quad & E(\log C_t) = \alpha_1 + \alpha_2 \log P_t + \alpha_3 \log R_t \\ 1982 - 1985: \quad & E(\log C_t) = \alpha_1 + \alpha_2 \log P_t + \alpha_3 \log R_t + \alpha_4 \\ 1986 - 1988: \quad & E(\log C_t) = \alpha_1 + \alpha_2 \log P_t + \alpha_3 \log R_t + \alpha_4 + \alpha_5 \end{aligned}$$

3. $\hat{\alpha}_4$ = -0,101307: diferencia en el consumo de cigarrillos (en logaritmos) entre el periodo 1960-1981 y el periodo 1982-1988, para el mismo precio y la misma renta.

   $\hat{\alpha}_5$ = -0,100729: diferencia en el consumo de cigarrillos (en logaritmos) entre el periodo 1982-1985 y el periodo 1986-1988, para el mismo precio y la renta.

4. Cada una de las campañas habría sido efectiva si hubiera logrado reducir el consumo de cigarrillos.

$$\begin{array}{l} H_0 : \alpha_i = 0 \\ H_a : \alpha_i < 0 \end{array} \qquad t = \frac{\hat{\alpha}_i - 0}{\hat{\sigma}_{\hat{\alpha}_i}} \overset{H_0}{\sim} t(T-k) \qquad i = 4, 5.$$

$$\text{Campaña 82 } (D82) \quad t = \frac{-0,101307 - 0}{0,0258239} = -3,923 < -1,71088 = t(29-5)_{0,05}$$

$$\text{Campaña 86 } (D86) \quad t = \frac{-0,100729 - 0}{0,0360802} = -2,7918 < -1,71088 = t(29-5)_{0,05}$$

En ambos casos se rechaza la hipótesis nula a un nivel de significación del 5 %, por lo que se puede concluir que tanto la campaña del departamento de Sanidad ($D82$) como la campaña del periódico ($D86$) han influido negativamente en el consumo de tabaco, es decir, han sido efectivas.

5. Los residuos de la estimación MCO del Modelo P11.2 oscilan en torno a cero sin observarse un patrón de comportamiento claro. El descenso brusco de los residuos al final de la muestra que se observaba en los residuos del Modelo P11.1 ha desaparecido. Ese patrón parece haber sido recogido mediante la inclusión de las dos nuevas variables explicativas cualitativas en el Modelo P11.2: Campaña del 82 y Campaña del 86. Por lo tanto, se puede afirmar que en el Modelo P11.1 había un problema de omisión de variables relevantes.

   Quizás se puede observar en el gráfico de los residuos que su variabilidad es mayor al final de la muestra que al principio de la misma, es decir, la varianza de los residuos no parece constante a lo largo de la muestra, lo que podría ser un indicio de heterocedasticidad en las perturbaciones.

6. El Modelo P11.2 porque en los contrastes realizados en el apartado 4 de esta Segunda Parte, se obtiene que las variables Campaña del 82 y Campaña del 86 son individualmente significativas. Por lo tanto, el Modelo P11.1 omite variables relevantes con lo que los estimadores MCO son sesgados y la inferencia no es válida.

## TERCERA PARTE.

1. Modelo P11.3:

$$\begin{aligned}\log C_t &= \delta_1 + \delta_2 \log P_t + \delta_3 \log R_t + \delta_4 D82_t + \delta_5 D86_t + \\ &+ \delta_6 \log P_t\, D82_t + \delta_7 \log P_t\, D86_t + \delta_8 \log R_t\, D82_t + \delta_9 \log R_t\, D86_t + u_t\end{aligned}$$

   El Modelo P11.3 tiene las mismas variables explicativas que el Modelo P11.2, pero la forma funcional es diferente: la relación de las variables $\log P$ y $\log R$ con las variables Campaña del 82 y Campaña del 86 ya no es lineal porque se han introducido interacciones entre ellas, de forma que el efecto del precio y de la renta sobre el consumo (las elasticidades precio y renta) puedan ser diferente antes y después de la puesta en marcha de cada una de las campañas publicitarias.

   En este modelo estamos suponiendo que hay un cambio estructural. No sólo es que las campañas hayan tenido efecto en el consumo de cigarrillos, sino que los coeficientes de la regresión son diferentes en cada uno de los tres periodos considerados.

2. Consumo promedio de cigarrillos:

$$1960 - 1981:$$

$$E(\log C_t) = \quad \delta_1 + \delta_2 \log P_t + \delta_3 \log R_t$$

$$1982 - 1985:$$

$$E(\log C_t) = \delta_1 + \delta_2 \log P_t + \delta_3 \log R_t + \delta_4 + \delta_6 \log P_t + \delta_8 \log R_t$$

$$= (\delta_1 + \delta_4) + (\delta_2 + \delta_6) \log P_t + (\delta_3 + \delta_8) \log R_t$$

$$1986 - 1988:$$

$$E(\log C_t) = \delta_1 + \delta_2 \log P_t + \delta_3 \log R_t + \delta_4 + \delta_5 + \delta_6 \log P_t + \delta_7 \log P_t + \delta_8 \log R_t$$

$$= (\delta_1 + \delta_4 + \delta_5) + (\delta_2 + \delta_6 + \delta_7) \log P_t + (\delta_3 + \delta_8 + \delta_9) \log R_t$$

3. Elasticidad renta $= \dfrac{E(\log C_t)}{\partial \log R_t}$.

   Estimación de la elasticidad renta:

   $$1960 - 1981: \quad \hat{\delta}_3 = 0,735837$$
   $$1982 - 1985: \quad \hat{\delta}_3 + \hat{\delta}_8 = -1,346643$$
   $$1986 - 1988: \quad \hat{\delta}_3 + \hat{\delta}_8 + \hat{\delta}_9 = 2,904707$$

   Se puede observar que no es constante en el tiempo, sino que toma tres valores diferentes para los tres periodos diferentes.

4. $\hat{\delta}_8$ = -2,80248: diferencia en el incremento porcentual estimado en el consumo ante incrementos de la renta de un 1 % entre el periodo 1982-1985 y el periodo 1960-1981, manteniendo el precio constante.
   = diferencia en la elasticidad renta entre el periodo 1982-1985 y el periodo 1960-1981, manteniendo el precio constante.

   $\hat{\delta}_9$ = 4,25135: diferencia en el incremento porcentual estimado en el consumo ante incrementos de la renta de un 1 % entre el periodo 1986-1988 y el periodo 1982-1985, manteniendo el precio constante.
   = diferencia en la elasticidad renta entre el periodo 1986-1988 y el periodo 1982-1985, manteniendo el precio constante.

5. Para elegir entre ambos modelos hay que saber si existe evidencia en la muestra a favor de la hipótesis de cambio estructural que recoge el Modelo P11.3:

$$H_0: \delta_6 = \delta_7 = \delta_8 = \delta_9 = 0$$
$$H_a: \delta_6 \neq 0 \text{ y/o } \delta_7 \neq 0 \text{ y/o } \delta_8 \neq 0 \text{ y/o } \delta_9 \neq 0$$
$$F = \frac{(SCR_R - SCR_{NR})/q}{SCR_{NR}/(N-k)} \overset{H_0}{\sim} \mathcal{F}(q, N-k)$$

donde $SCR_{NR}$ es la suma de cuadrados de residuos del modelo no restringido, Modelo P11.3, $SCR_R$ es la suma de cuadrados de residuos del modelo restringido, Modelo P11.2, y $q = 4$ es el número de restricciones.

$$F = \frac{(0,040520 - 0,016601)/4}{0,016601/(29-9)} = 7,20384 \; > \; 2,86608 = \mathcal{F}_{0,05}(4,20)$$

Se rechaza la hipótesis nula con un nivel de significación del 5 %, es decir, se puede concluir que existe cambio estructural.

Dado este resultado, se elige el Modelo P11.3, porque los estimadores MCO del Modelo P11.2 al omitir este efecto relevante estarán sesgados y la inferencia no es válida.

# Solución PRÁCTICA P12.

## PRIMERA PARTE.

1. MRLG:

$$V_t = \beta_1 + \beta_2 \, P_t + \beta_3 \, C_t + \beta_4 \, GP_t + u_t \qquad u_t \sim NID(0, \sigma_u^2) \quad t = 1989{:}1, \ldots, 2010{:}4$$

2. FRM:

$$\underset{(\hat{\sigma}_{\hat{\beta}})}{\widehat{V}_t} = \underset{(173{,}557)}{304,430} - \underset{(0{,}983204)}{1,41396} \; P_t + \underset{(0{,}798534)}{2,56275} \; C_t + \underset{(0{,}0757153)}{0,743241} \; GP_t$$

$$R^2 = 0,994851 \qquad\qquad SCR = 96083,63$$

3. $\hat{\beta}_2$ = -1,41396. Se estima que si el precio de la botella aumenta en un euro, el número de botellas vendidas disminuye en 1,4139, manteniendo las variables precio de los competidores y gastos en publicidad constantes.

   $\hat{\beta}_3$ = 2,56275. Si el precio de la botella de las empresas competidoras aumenta en un euro, se estima que el número de botellas vendidas aumenta en 2,56275, manteniendo las variables precio de la empresa y gastos en publicidad constantes.

   Ambos coeficientes tienen los signos esperados. El modelo con el que estamos trabajando es un modelo de demanda y se espera, por lo tanto, que la cantidad vendida esté relacionada inversamente (signo negativo) con su precio de venta y directamente (signo positivo) con el precio de los competidores.

4. $R^2 = 1 - \dfrac{SCR}{SCT} = 1 - \dfrac{\sum \hat{u}_t^2}{\sum (V_t - \bar{V})^2} = 0,994851$

   Interpretación: el 99,4851 % de la variabilidad de las ventas de botellas de txakoli en la muestra está explicada por la variabilidad de las variables explicativas: precio de la empresa, precio de las empresas competidoras y gastos en publicidad.

5. Valor observado: $V_{90:2} = 1863,5$ botellas.

   Valor estimado:

   $\widehat{V}_{90:2} = 304,43 - 1,41396 \times 39,4 + 2,56275 \times 23,7 + 0,7432 \times 2098 = 1868,8$ botellas.

   Diferencia: $V_{1990:2} - \widehat{V}_{1990:2} = 1863,5 - 1868,8 = -5,3$ botellas.

   A esta diferencia se le denomina residuo.

6. Incremento estimado:

$$\begin{aligned}
\widehat{\Delta V}_t &= \hat{V}_t - \hat{V}_{t-1} = -1,41396 \times \Delta P_t + 2,56275 \times \Delta C_t + 0,743241 \times \Delta GP_t = \\
&= -1,41396 \times \Delta P_t + 2,56275 \times \Delta C_t + 0,743241 \times 200 = \\
&= (-1,41396 \times \Delta P_t + 2,56275 \times \Delta C_t + 148,6482) \text{ botellas}
\end{aligned}$$

   Si las demás variables, precio de la empresa y precio de los competidores, se mantienen constantes:

$$\widehat{\Delta V}_t = 0,743241 \times 200 = 148,6482 \text{ botellas.}$$

7. Interpretación del parámetro $\hat{\beta}_4 = 0{,}743241$. Si los gastos en publicidad aumentan en un euro manteniendo el resto de las variables constantes, se estima un incremento en las ventas de 0,743241 botellas.

   Por lo tanto, si el gasto en publicidad aumentara en 100 euros se estimaría un aumento en las ventas de $100 \times 0,743241 = 74,3241$ botellas.

   Este resultado se basa en una estimación por punto del coeficiente $\beta_4$. Para contar con más información se puede estimar dicho coeficiente por intervalo, lo que proporcionaría un límite superior y un límite inferior de la variación de la ventas ante cambios en $GP$ con una probabilidad determinada.

   En este caso en particular, hay que construir un intervalo para $100\,\beta_4$:

$$\begin{aligned}
IC(100\,\beta_4)_{95\,\%} &= \left[100\,\hat{\beta}_4 \pm t(T-k)_{0,05/2}\,100\,\hat{\sigma}_{\hat{\beta}_4}\right] = \\
&= 100\left[\hat{\beta}_4 \pm t(T-k)_{0,05/2}\,\hat{\sigma}_{\hat{\beta}_4}\right] =
\end{aligned}$$

$$
\begin{aligned}
&= 100\,\left[0,743241 \pm t(88-4)_{0,05/2}\; 0,0757153\right] = \\
&= 100\,[0,743241 \;\pm\; 1,98861 \;\times\; 0,0757153] = \\
&= 100\,[0,743241 \;\pm\; 0,150568202] = \\
&= 100\,[0,592672797 \quad 0,893809202] = \\
&= [59,2672797 \quad 89,3809202] \text{ botellas}
\end{aligned}
$$

La variación mínima en las ventas es de aproximadamente 60 botellas y la máxima está alrededor de 90 botellas.

8. Contraste de significatividad individual de las variables explicativas relativas a los precios.

$$
\begin{array}{l} H_0 : \beta_i = 0 \\ H_a : \beta_i \neq 0 \end{array} \qquad t = \frac{\hat{\beta}_i - 0}{\hat{\sigma}_{\hat{\beta}_i}} \overset{H_0}{\sim} t(T-k) \qquad i = 2,3.
$$

Precio botella ($P$) $\quad |t| = \left| \dfrac{-1,41396 - 0}{0,9832} \right| = 1,438 \;<\; 1,9886 = t(88-4)_{0,05/2}$

Precio competidor ($C$) $\quad |t| = \left| \dfrac{2,56275 - 0}{0,79853} \right| = 3,2093 \;>\; 1,9886 = t(88-4)_{0,05/2}$

Por lo tanto, se puede concluir que, para un nivel de significación del 5 %, la variable precio del marisco no es individualmente significativa y la variable precio de los competidores sí es individualmente significativa.

9. Contraste de significatividad conjunta.

$$
\begin{array}{l} H_0 : \beta_2 = \beta_3 = \beta_4 = 0 \\ H_a : \beta_2 \neq 0 \text{ y/o } \beta_3 \neq 0 \text{ y/o } \beta_4 \neq 0 \end{array} \qquad F = \frac{R^2/q}{(1-R^2)/(T-k)} \overset{H_0}{\sim} \mathcal{F}(q, T-k)
$$

$$
F = \frac{0,994851/3}{(1-0,994851)/(88-4)} = 5409,541 \;>\; 2,71323 = \mathcal{F}(3,84)_{0,05}
$$

Se rechaza la hipótesis nula con un nivel de significación del 5 %, es decir, se puede concluir que las variables explicativas son conjuntamente significativas.

En principio, no existe ningún problema muestral, dado que en el apartado anterior las variables precio de la competencia y renta son individualmente significativas.

10. Se observa que los residuos oscilan en torno a cero con una variabilidad homogénea. Ahora bien, el gráfico muestra largas rachas de residuos positivos y de residuos negativos. Este hecho puede estar indicando que las perturbaciones están autocorrelacionadas en el tiempo, es decir, que la perturbación en el momento $t$ no es independiente de las perturbaciones pasadas, $u_{t-1}, \ldots$. Luego podría estar incumpliéndose el supuesto de ausencia de autocorrelación de las perturbaciones.

## SEGUNDA PARTE.

1. Modelo P12.2:

$$V_t = \beta_1 + \beta_2 P_t + \beta_3 C_t + \beta_4 GP_t + \beta_5 \text{Clima}_t + u_t \quad t = 1989{:}1, \ldots, 2004{:}4$$

   La diferencia con el Modelo P12.1 es que se ha incluído la variable climatología de forma cualitativa con dos categorías: buena climatología y mala climatología. En el modelo se ha mantenido el término constante y se ha incluido una de las categorías a través de la variable ficticia Clima que da el valor 1 para el periodo de mala climatología y 0 en otro caso.

2. $\hat{\beta}_5$ = -55,9927: diferencia estimada en las ventas de botellas de txakoli entre un periodo de mala climatología para las uvas y un periodo de buena climatología, a igualdad de condiciones de precios (propios y de los competidores) y gastos en publicidad. Es de esperar que las ventas disminuyan en épocas de mala climatología por lo que el coeficiente estimado tiene el signo esperado.

3. Contraste a una cola.

$$\begin{array}{l} H_0 : \beta_5 = 0 \\ H_a : \beta_5 < 0 \end{array} \qquad t = \frac{\hat{\beta}_5 - 0}{\hat{\sigma}_{\hat{\beta}_5}} \overset{H_0}{\sim} t(T-k)$$

$$t = \frac{-55,9927 - 0}{6,59108} = -8,4952 \; < \; -1,66342 = t(88-5)_{0,05}$$

   Se rechaza la hipótesis nula con un nivel de significación del 5 %, es decir, se puede concluir que existe evidencia de que la mala climatología ha disminuido las ventas.

4. En este gráfico los residuos oscilan en torno a cero con una variabilidad homogénea. Comparando estos residuos con los del Modelo P12.1, parece que se comportan de una forma más aleatoria, con rachas más cortas de residuos positivos y negativos, por lo que la introducción del factor explicativo climatología parece haber paliado el problema que presentaba el Modelo P12.1.

5. Contraste de significatividad individual de las variables explicativas relativas a los precios.

$$\begin{array}{l} H_0 : \beta_i = 0 \\ H_a : \beta_i \neq 0 \end{array} \qquad t = \frac{\hat{\beta}_i - 0}{\hat{\sigma}_{\hat{\beta}_i}} \overset{H_0}{\sim} t(T-k) \qquad i = 2, 3.$$

Precio botella ($P$) $\qquad |t| = \left| \dfrac{-2,883 - 0}{0,743793} \right| = 3,8764 > 1,989 = t(88-5)_{0,05/2}$

Precio competidor ($C$) $\quad |t| = \left| \dfrac{3,02811 - 0}{0,590079} \right| = 5,1317 > 1,989 = t(88-5)_{0,05/2}$

Por lo tanto, se puede concluir que, para un nivel de significación del 5 %, las variables precio del marisco y precio de los competidores son individualmente significativas.

Como la climatología es una variable explicativa significativa, el Modelo P12.1 es un modelo que no cumple todas las hipótesis básicas del MRLG ya que omite una variable relevante. Esto implica que los estimadores MCO del Modelo P12.1 son sesgados, el estadístico $t$ no sigue la distribución $t$ de Student y la inferencia realizada en base a esta distribución como la que se ha llevado a cabo en el apartado 8 de la Primera Parte puede conducir a conclusiones erróneas.

## TERCERA PARTE.

1. MRLG:

$$\begin{aligned} V_t &= \beta_1 + \beta_2\, P_t + \beta_3\, C_t + \beta_4\, GP_t + \beta_5\,\text{Clima}_t + \beta_6\, dq1_t + \\ &+ \beta_7\, dq2_t + \beta_8\, dq3_t + u_t \qquad u_t \sim NID(0, \sigma_u^2) \quad t = 1989{:}1, \ldots, 2010{:}4. \end{aligned}$$

2. FRM:

$$\begin{aligned} \underset{(\hat{\sigma}_{\hat{\beta}})}{\widehat{V}_t} &= \underset{(131{,}504)}{458,366} - \underset{(0{,}755292)}{2,84476}\, P_t + \underset{(0{,}599282)}{2,99899}\, C_t + \underset{(0{,}0569771)}{0,690514}\, GP_t - \\ &- \underset{(6{,}68217)}{55,9285}\,\text{Clima} + \underset{(7{,}61884)}{1,93326}\, dq1_t + \underset{(7{,}61273)}{1,40220}\, dq2_t + \underset{(7{,}62133)}{6,30854}\, dq3_t \end{aligned}$$

$$R^2 = 0,997272 \qquad SCR = 50907,59$$

3. Estimación de las ventas:

$\hat{\beta}_1$ = 458,366: venta estimada de botellas de txakoli en los cuartos trimestres con buena climatología cuando los precios (propios y de los competidores) y los gastos en publicidad son cero.

$\hat{\beta}_7$ = 1,40220: diferencia estimada en la venta de botellas de txakoli entre el segundo trimestre y el cuarto trimestre a igualdad de condiciones climáticas, de precios (propios y de los competidores) y gastos en publicidad.

4. Basándonos en los datos del fichero practicaP12.gdt:

$$\begin{aligned}\widehat{V}_{03:1} &= 458,366 - 2,845\,P_{03:1} + 2,999\,C_{03:1} + 0,691\,GP_{03:1} - 55,9285 + 1,93326 \\ &= 2712,3\,\text{botellas}\end{aligned}$$

$$\begin{aligned}\widehat{V}_{04:1} &= 458,366 - 2,845\,P_{04:1} + 2,999\,C_{04:1} + 0,691\,GP_{04:1} + 1,93326 \\ &= 2833,4\,\text{botellas}\end{aligned}$$

5. La variable estacionalidad no es significativa si, a igualdad de condiciones, no hay diferencias en el promedio de las ventas por estar en un trimestre o en otro.

$$H_0 : \beta_6 = \beta_7 = \beta_8 = 0$$
$$H_a : \beta_6 \neq 0 \text{ y/o } \beta_7 \neq 0 \text{ y/o } \beta_8 \neq 0$$

$$F = \frac{(SCR_R - SCR_{NR})/q}{SCR_{NR}/(T-k)} \overset{H_0}{\sim} \mathcal{F}(q, T-k)$$

donde $SCR_{NR}$ es la suma de cuadrados de residuos del modelo no restringido, Modelo P12.3, $SCR_R$ es la suma de cuadrados de residuos del modelo restringido, Modelo P12.2, y $q = 3$ es el número de restricciones.

$$F = \frac{(51395,28 - 50907,59)/3}{50907,59/(88-8)} = 0,255464 \;<\; 1,99006 = \mathcal{F}_{0,05}(3,80)$$

No se rechaza la hipótesis nula con un nivel de significación del 5 %, es decir, se puede concluir que la estacionalidad no es una variable significativa para explicar las ventas, una vez incluidas el resto de las variables.

6. El Modelo P12.2 porque:

- El Modelo P12.1 omite la variable relevante climatología por lo que los estimadores MCO están sesgados y la inferencia no es válida.
- El Modelo P12.3 incluye una variable irrelevante, la estacionalidad. Los estimadores MCO del Modelo P12.3 siguen siendo insesgados y la inferencia es válida. Pero el Modelo P12.2 es el Modelo P12.3 incluyendo la restricción cierta de que la variable estacionalidad no es significativa. Esto implica que los estimadores MCR (MCO en el Modelo P12.2) tienen menor varianza que los estimadores MCO en Modelo P12.3.

# Solución PRÁCTICA P13.

## PRIMERA PARTE.

1. Estimación MCO.

$$\hat{\beta}_2 = \frac{\sum_{i=1}^{N}(S_i - \bar{S})(E_i - \bar{E})}{\sum_{i=1}^{N}(E_i - \bar{E})^2} = \frac{30421,7551}{270,5306} = 112,452$$

$$\hat{\beta}_1 = \bar{S} - \hat{\beta}_2\,\bar{E} = \frac{89190}{49} - 112,4521777 \times \frac{305}{49} = 1120,25$$

Recta de regresión muestral: $\widehat{S}_i = 1120,25 + 112,452\,E_i \qquad i = 1, 2, \ldots, 49.$

2. $\hat{\beta}_2 = 112{,}452$: incremento estimado en el salario (en euros) si el nivel de educación por encima de los estudios básicos aumenta en un año.

3. $R^2 = 1 - \dfrac{SCR}{SCT} = 1 - \dfrac{\sum \hat{u}_i^2}{\sum(S_i - \bar{S})^2} = 1 - \dfrac{16751123,49}{20172111,96} = 0,16959$

dado que:

$$\begin{aligned}\sum \hat{u}_i^2 &= \sum S_i^2 - \hat{\beta}_1 \sum S_i - \hat{\beta}_2 \sum S_i E_i = \\ &= -1120,25 \times 89190 - 112,452 \times 585584 = 16751123,49\end{aligned}$$

y

$$\sum S_i^2 = \sum(S_i - \bar{S})^2 + T\bar{S}^2 = 20172111,96 + 49 \times (1820,204082)^2 = 182516114$$

Interpretación: el 16,959 % de la variabilidad de salario en la muestra viene explicada por la variabilidad de la variables explicativa educación.

4. Contraste de significatividad individual: educación.

$$\begin{array}{l} H_0 : \beta_2 = 0 \\ H_a : \beta_2 \neq 0 \end{array} \qquad t = \frac{\hat{\beta}_2 - 0}{\hat{\sigma}_{\hat{\beta}_2}} \overset{H_0}{\sim} t(N-k)$$

$$|t| = \left| \frac{112,452 - 0}{\sqrt{1317,436485}} \right| = 3,0981 > 2,01174 = t(49-2)_{0,05/2}$$

donde:

$$\hat{\sigma}^2_{\hat{\beta}_2} = \frac{\hat{\sigma}_u^2}{\sum(E_i - \bar{E})^2} = \frac{356406,612}{270,5306} = 1317,436485$$

y

$$\hat{\sigma}_u^2 = \frac{\sum \hat{u}_i^2}{N-k} = \frac{16751123,49}{49-2} = 356406,612$$

Se rechaza la hipótesis nula con un nivel de significación del 5 %, es decir, se puede concluir que la educación es una variable significativa para explicar el salario.

5. Si todos los individuos tuvieran estudios básicos, la variable educación tomaría el valor 0 para todas las observaciones de la muestra. En este caso, el rango de la matriz de datos $X$ no es completo y no se pueden estimar los coeficientes por MCO. Además, en una muestra de estas características no tiene sentido analizar la influencia de los años de educación sobre el salario.

## SEGUNDA PARTE.

1. MRLG:

$$S_i = \beta_1 + \beta_2 E_i + \beta_3 X_i + \beta_4 H_i + u_i \qquad u_i \sim NID(0, \sigma_u^2) \quad i = 1, 2, \ldots, 49.$$

En el Modelo P13.2 se han incluido dos variables explicativas nuevas: una cuantitativa, experiencia medida en años, y una cualitativa, género. Esta última tiene dos categorías, hombre y mujer, por lo que se definirían dos variables ficticias, una por categoría. El modelo se ha especificado introduciendo el término independiente y una sola variable ficticia la que corresponde a la categorías hombre.

2. $\hat{\beta}_1$ = 434,821: salario estimado en euros para un hombre con estudios básicos y sin experiencia en el mercado laboral.

   $\hat{\beta}_2$ = 133,551: incremento estimado en el salario en euros ante un incremento de un año en la educación (por encima de los estudios básicos) manteniendo constante la experiencia e independientemente del género.

   $\hat{\beta}_3$ = 34,4543: se estima que el salario aumenta en 34,4543 euros ante un incremento de un año en la experiencia laboral, manteniendo constante la educación e independientemente del género.

   $\hat{\beta}_4$ = 470,46: diferencia estimada en el salario entre un hombre y una mujer para los mismos años de educación y experiencia.

3. **A.** Contraste de significatividad conjunta.

$$\begin{array}{l} H_0 : \beta_2 = \beta_3 = \beta_4 = 0 \\ H_a : \beta_2 \neq 0 \text{ y/o } \beta_3 \neq 0 \text{ y/o } \beta_4 \neq 0 \end{array} \qquad F = \frac{R^2/q}{(1-R^2)/(N-k)} \overset{H_0}{\sim} \mathcal{F}(q, N-k)$$

$$F = \frac{0,450263/3}{(1-0,450263)/(49-4)} = 12,28578 \, > \, 2,81154 = \mathcal{F}(3, 49-4)_{0,05}$$

Se rechaza la hipótesis nula con un nivel de significación del 5 %, es decir, se puede concluir que las variables explicativas educación, experiencia y género son conjuntamente significativas.

**B.** Contraste de significatividad individual.

$$\begin{array}{l} H_0 : \beta_i = 0 \\ H_a : \beta_i \neq 0 \end{array} \qquad t = \frac{\hat{\beta}_i - 0}{\hat{\sigma}_{\hat{\beta}_i}} \overset{H_0}{\sim} t(N-k) \qquad i = 2, 3, 4.$$

$$\text{Educación } (E) \qquad |t| = \left| \frac{133,551 - 0}{31,5081} \right| = 4,2386 \; > \; 2,0141 = t(49-4)_{0,05/2}$$

$$\text{Experiencia } (X) \qquad |t| = \left| \frac{34,4543 - 0}{12,1316} \right| = 2,8401 \; > \; 2,0141 = t(49-4)_{0,05/2}$$

$$\text{Género } (H) \qquad |t| = \left| \frac{470,46 - 0}{144,87} \right| = 3,2475 \; > \; 2,0141 = t(49-4)_{0,05/2}$$

Por lo tanto, se puede concluir que, para un nivel de significación del 5 %, las variables educación, experiencia y género son individualmente significativas.

Las variables explicativas son significativas tanto individual como conjuntamente por lo que no hay evidencia de ningún problema muestral del tipo de alto grado de multicolinealidad entre los regresores.

4. Contraste a una cola.

$$\begin{array}{l} H_0 : \beta_4 = 0 \\ H_a : \beta_4 > 0 \end{array} \qquad t = \frac{\hat{\beta}_4 - 0}{\hat{\sigma}_{\hat{\beta}_4}} \overset{H_0}{\sim} t(N-k)$$

$$t = \frac{470,46 - 0}{144,87} = 3,2475 \; > \; 1,67943 = t(49-4)_{0,05}$$

Se rechaza la hipótesis nula con un nivel de significación del 5 %, es decir, se puede concluir que existe evidencia de que los hombres ganan más en media que las mujeres.

5. Salarios estimados:

$$\begin{aligned} (\hat{S}|E = 5, X = 0, H = 0) &= 434,821 + 133,551 \times 5 = 1102,576 \text{ euros} \\ (\hat{S}|E = 0, X = 5, H = 0) &= 434,821 + 34,4543 \times 5 = 607,0925 \text{ euros} \end{aligned}$$

6. Incremento estimado salario para una trabajadora: 34,4543 euros.

   Incremento estimado salario para un trabajador: 34,4543 euros.

   Estos resultados se obtienen porque el coeficiente $\hat{\beta}_3 = 34,4543$ mide la variación estimada en el salario si la experiencia aumenta en un año, manteniéndose constante la educación, tanto para un hombre como para una mujer.

## TERCERA PARTE.

1. MRLG:

$$S_i = \beta_1 + \beta_2 E_i + \beta_3 X_i + \beta_4 H_i + \beta_5 A_i + \beta_6 G_i + u_i \qquad u_i \sim NID(0, \sigma_u^2) \quad i = 1, \ldots, 49.$$

   En el Modelo P13.3 se ha incluido una nueva variable explicativa cualitativa, la localización de la sucursal de la empresa. Esta variable tiene tres categorías dependiendo de si la sucursal está en Alava, Bizkaia o Gipuzkoa, por lo que se definen tres variables ficticias:

$$A_i = \begin{cases} 1 & i \in \text{Alava} \\ 0 & \text{en otro caso} \end{cases} \qquad B_i = \begin{cases} 1 & i \in \text{Bizkaia} \\ 0 & \text{en otro caso} \end{cases} \qquad G_i = \begin{cases} 1 & i \in \text{Gipuzkoa} \\ 0 & \text{en otro caso} \end{cases}$$

   Para evitar problemas de colinealidad perfecta en la matriz de datos $X$, el modelo se ha especificado con el término constante e incluyendo tantas variables ficticias como categorías menos una. Se ha dejado fuera la categoría Bizkaia, que servirá de referencia.

2. La variable localización de la sucursal no es significativa si, a igualdad de condiciones, no hay diferencias en el salario promedio del empleado por trabajar en una provincia o en otra.

$$\begin{array}{l} H_0 : \beta_5 = \beta_6 = 0 \\ H_a : \beta_5 \neq 0 \text{ y/o } \beta_6 \neq 0 \end{array} \qquad F = \frac{(SCR_R - SCR_{NR})/q}{SCR_{NR}/(N-k)} \overset{H_0}{\sim} \mathcal{F}(q, N-k)$$

   donde $SCR_{NR}$ es la suma de cuadrados de residuos del modelo no restringido, Modelo P13.3, $SCR_R$ es la suma de cuadrados de residuos del modelo restringido, Modelo P13.2, y $q = 2$ es el número de restricciones.

$$F = \frac{(11089355 - 5689030)/2}{5689030/(49-6)} = 20,4089 \ > \ 3,21448 = \mathcal{F}(2, 49-6)_{0,05}$$

   Se rechaza la hipótesis nula con un nivel de significación del 5 %, es decir, se puede concluir que la localización de la sucursal es una variable significativa para explicar el salario, una vez incluidas el resto de las variables.

3. Diferencia salarial estimada entre hombres y mujeres si el resto de las características se mantienen iguales = 594,659 euros = $\hat{\beta}_4$.

   Contraste de significatividad individual: género.

$$\begin{array}{l} H_0 : \beta_4 = 0 \\ H_a : \beta_4 \neq 0 \end{array} \qquad t = \frac{\hat{\beta}_4 - 0}{\hat{\sigma}_{\hat{\beta}_4}} \overset{H_0}{\sim} t(N-k)$$

$$|t| = \left| \frac{594,659 - 0}{118,738} \right| = 5,0082 \; > \; 2,01669 = t(49-6)_{0,05/2}$$

   Se rechaza la hipótesis nula con un nivel de significación del 5 %, es decir, se puede concluir que el género es una variable significativa para explicar el salario.

4. Salario estimado para una mujer de la sucursal vizcaína con estudios básicos y sin experiencia:

   $(\hat{S}|E = 0, X = 0, H = 0, A = 0, G = 0) = \hat{\beta}_1 = 1580{,}34$ euros.

   Salario estimado para una mujer de la sucursal alavesa con estudios básicos y sin experiencia:

   $(\hat{S}|E = 0, X = 0, H = 0, A = 1, G = 0) = \hat{\beta}_1 + \hat{\beta}_5 =$ (1580,34 -1151,2) euros.

   Salario estimado para una mujer de la sucursal guipuzkoana con estudios básicos y sin experiencia:

   $(\hat{S}|E = 0, X = 0, H = 0, A = 0, G = 1) = \hat{\beta}_1 + \hat{\beta}_6 =$ (1580,34 - 800,65) euros.

5. (1580,34 + 594,659) $= \hat{\beta}_1 + \hat{\beta}_4 = (\hat{S}|E = 0, X = 0, H = 1, A = 0, G = 0)$

   Salario estimado para un hombre de la sucursal vizcaína con estudios básicos y sin experiencia.

## CUARTA PARTE.

1. Restricción en los coeficientes del Modelo P13.3 que conduce a la especificación del Modelo P13.4: $\beta_5 = \beta_6$.

$$\begin{aligned} S_i &= \beta_1 + \beta_2\, E_i + \beta_3\, X_i + \beta_4\, H_i + \beta_5\, A_i + \beta_6\, G_i + u_i \quad \rightarrow \\ &= \beta_1 + \beta_2\, E_i + \beta_3\, X_i + \beta_4\, H_i + \beta_5\, A_i + \beta_5\, G_i + u_i \quad \rightarrow \\ &= \beta_1 + \beta_2\, E_i + \beta_3\, X_i + \beta_4\, H_i + \beta_5\, (A+G)_i + u_i \end{aligned}$$

   Esta restricción significa que, a igualdad de condiciones, no existe diferencia salarial entre los empleados de las sucursales de Gipuzkoa y Alava, pero sí respecto de los de Bizkaia.

2. Salario estimado para una mujer de la sucursal vizcaína con estudios básicos y sin experiencia:

   $(\hat{S}|E = 0, X = 0, H = 0, A = 0, G = 0) = \hat{\beta}_1 = 1580{,}34$ euros.

   Salario estimado para una mujer de la sucursal alavesa con estudios básicos y sin experiencia:

   $(\hat{S}|E = 0, X = 0, H = 0, A = 1, G = 0) = \hat{\beta}_1 + \hat{\beta}_5 =$ (1503,49 - 857,85) euros.

   Salario estimado para una mujer de la sucursal guipuzkoana con estudios básicos y sin experiencia:

   $(\hat{S}|E = 0, X = 0, H = 0, A = 0, G = 1) = \hat{\beta}_1 + \hat{\beta}_5 =$ (1503,49 - 857,85) euros.

   Como se puede observar existen diferencias con los resultados del apartado 4. de la Tercera Parte. Estas diferencias provienen, por un lado, de que las estimaciones de los coeficientes son diferentes en el Modelo P13.4 y en el Modelo P13.3 y, por otro lado, de que la especificación de los dos modelos es también diferente:

   En el apartado 4 de la Tercera Parte el salario estimado de una trabajadora con estudios básicos y sin experiencia es diferente para las trabajadoras de las tres sucursales.

   Sin embargo, según el Modelo P13.4, el salario estimado de una trabajadora con estudios básicos y sin experiencia es el mismo para las trabajadoras de Alava y Gipuzkoa y sólo es distinto para las de Bizkaia.

   Notése que se obtendría el mismo tipo de resultado para el caso de un hombre con las mismas características.

3. El Modelo P13.3 porque:

   En el apartado 2 de la Tercera Parte se ha comprobado que la variable localización de la sucursal es significativa. Además, analizando el resto de los resultados de la estimación del Modelo P13.3, se puede concluir que las variables educación, experiencia y género son también individualmente significativas ya que sus estadísticos t superan el valor crítico para un nivel de significación del 5 %.

   En consecuencia el Modelo P13.1 y el Modelo P13.2 quedan descartados porque omiten variables relevantes lo que supone que los estimadores MCO son sesgados y la inferencia no es válida porque los estadísticos t y F no siguen las distribuciones habituales.

   En lo que se refiere al Modelo P13.4, éste no es más que una versión restringida del Modelo P13.3. Elegiremos el Modelo P13.4 si la restricción impuesta ($\beta_5 = \beta_6$) es cierta porque, en este caso, los estimadores MCR son insesgados y su varianza es menor que el estimador MCO del Modelo P13.3. Por el contrario, si la restricción no es cierta, se elegirá el Modelo P13.3 porque sabemos que, en este caso, los estimadores MCR están sesgados.

Por lo tanto, la elección depende del resultado del contraste de la restricción $\beta_5 = \beta_6$ en el Modelo P13.3.

$$H_0 : \beta_5 = \beta_6 \ (\beta_5 - \beta_6 = 0) \qquad F = \frac{(SCR_R - SCR_{NR})/q}{SCR_{NR}/(N-k)} \overset{H_0}{\sim} \mathcal{F}(q, N-k)$$
$$H_a : \beta_5 \neq \beta_6 \ (\beta_5 - \beta_6 \neq 0)$$

donde $SCR_{NR}$ es la suma de cuadrados de residuos del modelo no restringido, Modelo P13.3, $SCR_R$ es la suma de cuadrados de residuos del modelo restringido, Modelo P13.4, y $q = 1$ es el número de restricciones.

$$F = \frac{(6347297 - 5689030)/1}{5689030/(49-6)} = 4,97545 \ > \ 4,06705 = \mathcal{F}(1, 49-6)_{0,05}$$

Se rechaza la hipótesis nula con un nivel de significación del 5 %, es decir, se puede concluir que, a igualdad de condiciones, la diferencia salarial entre los trabajadores de las sucursales de Alava y Gipuzkoa respecto de la de Bizkaia no es la misma. Por lo tanto, elegimos el Modelo P13.3.

# Solución PRÁCTICA P14.

## PRIMERA PARTE.

1. Función de producción:

$$\log Q_t = \beta_1 + \beta_2 \log K_t + \beta_3 \log L_t + \beta_4 \log E_t + \beta_5 \log H_t + \beta_6 \log F_t + \beta_7 \log S_t + u_t$$

$$u_t \sim NID(0, \sigma_u^2) \qquad t = 1948, 1949, \ldots, 1993.$$

La estimación de este modelo por Mínimos Cuadrados Ordinarios proporciona los siguientes resultados:

Modelo P14.1: estimaciones MCO utilizando las 46 observaciones 1948–1993
Variable dependiente: $l\_Q$

| Variable | Coeficiente | Desv. típica | Estadístico $t$ | valor p |
|---|---|---|---|---|
| const | 2,62172 | 1,48327 | 1,7675 | 0,0850 |
| $l\_K$ | -0,340100 | 0,0886314 | -3,8372 | 0,0004 |
| $l\_L$ | -0,666346 | 0,0796336 | -8,3677 | 0,0000 |
| $l\_E$ | 0,208275 | 0,105148 | 1,9808 | 0,0547 |
| $l\_H$ | 1,07952 | 0,384756 | 2,8057 | 0,0078 |
| $l\_F$ | -0,00796152 | 0,0731634 | -0,1088 | 0,9139 |
| $l\_S$ | 0,141183 | 0,125934 | 1,1211 | 0,2691 |

Suma cuadrados residuos 0,0609414 $R^2$ 0,977494

Analizando los resultados de la tabla, se puede observar que los factores de producción, capital, trabajo, energía y superficie cultivada son individualmente significativos mientras que los factores de producción fertilizantes y semillas no lo son ni al 5 % ni al 10 %.

Comprobemos si los factores de producción fertilizantes y semillas son conjuntamente significativos.

$$H_0 : \beta_6 = \beta_7 = 0 \qquad F = \frac{(SCR_R - SCR_{NR})/q}{SCR_{NR}/(T-k)} \overset{H_0}{\sim} \mathcal{F}(q, T-k)$$
$$H_a : \beta_6 \neq 0 \text{ y/o } \beta_7 \neq 0$$

donde $q = 2$ es el número de restricciones, $SCR_{NR}$ es la suma de cuadrados de residuos del modelo no restringido, Modelo P14.1 y $SCR_R$ es la suma de cuadrados de residuos del siguiente modelo restringido:

$$\log Q_t \;=\; \beta_1 \;+\; \beta_2 \, \log K_t \;+\; \beta_3 \, \log L_t \;+\; \beta_4 \, \log E_t \;+\; \beta_5 \, \log H_t \;+ u_t$$

$$F = \frac{(0,062971 - 0,0609414)/2}{0,0609414/(46-7)} = 0,649457 \;<\; 3,2381 = \mathcal{F}(2,39)_{0,05}$$

No se rechaza la hipótesis nula con un nivel de significación del 5 %, es decir, se puede concluir que los factores de producción fertilizantes y semillas no son conjuntamente significativos.

Al no ser los factores de producción fertilizantes y semillas significativos, ni individual ni conjuntamente, se eliminan de la especificación del modelo estimando la siguiente función de producción:

$$\log Q_t \;=\; \beta_1 \;+\; \beta_2 \, \log K_t \;+\; \beta_3 \, \log L_t \;+\; \beta_4 \, \log E_t \;+\; \beta_5 \, \log H_t \;+\; u_t$$
$$u_t \;\sim\; NID(0, \sigma_u^2) \qquad t = 1948, 1949, \ldots, 1993.$$

Modelo P14.2: estimaciones MCO utilizando las 46 observaciones 1948–1993
Variable dependiente: $l\_Q$

| Variable | Coeficiente | Desv. típica | Estadístico $t$ | valor p |
|---|---|---|---|---|
| const | 3,39123 | 1,29215 | 2,6245 | 0,0121 |
| $l\_K$ | -0,315200 | 0,0850853 | -3,7045 | 0,0006 |
| $l\_L$ | -0,720098 | 0,0388490 | -18,5358 | 0,0000 |
| $l\_E$ | 0,191347 | 0,0949839 | 2,0145 | 0,0505 |
| $l\_H$ | 1,08932 | 0,364861 | 2,9856 | 0,0048 |

Suma cuadrados residuos 0,0629711 $R^2$ 0,976745

FRM:

$$\underset{(\hat{\sigma}_{\hat{\beta}})}{\widehat{\log Q_t}} = \underset{(1{,}29215)}{3,39123} - \underset{(0{,}0850853)}{0,3152} \log K_t - \underset{(0{,}0388490)}{0,720098} \log L_t + \underset{(0{,}0949839)}{0,191347} \log E_t +$$

$$+ \underset{(0{,}364861)}{1,08932} \log H_t \qquad R^2 = 0,976745 \qquad SCR = 0,0629711$$

**A.** Interpretación de los coeficientes.

$\hat{\beta}_1$ = 3,39123: producción estimada (en logaritmos) cuando todas las variables explicativas (en logaritmos) toman el valor cero, es decir, cuando se emplea una unidad de cada factor.

$\hat{\beta}_2$ = -0,3152: se estima que la producción disminuye un 0,3152 % cuando el factor capital aumenta un 1 %, manteniéndose el resto de los factores constantes.

$\hat{\beta}_3$ = -0,720098: se estima que la producción disminuye un 0,720098 % cuando el factor trabajo aumenta un 1 %, manteniéndose el resto de los factores constantes.

$\hat{\beta}_4$ = 0,191347 : se estima que la producción aumenta un 0,191347 % cuando el factor energía aumenta un 1 %, manteniéndose el resto de los factores constantes.

$\hat{\beta}_5$ = 1,08932 : se estima que la producción aumenta un 1,08932 % cuando el factor superficie cultivada aumenta un 1 %, manteniéndose el resto de los factores constantes.

Dado que el Modelo P14.2 es doble logarítmico, los coeficientes $\beta_2, \beta_3, \beta_4$ y $\beta_5$ son las elasticidades capital, trabajo, energía y superfice cultivada, respectivamente.

**B.** Bondad del ajuste. $R^2 = 0{,}976745$.

El 97,67 % de la variabilidad de la producción en la muestra (en logaritmos) está explicada por la variabilidad de las variables explicativas, capital, trabajo, energía y superfice cultivada (en logaritmos).

**C.** Significatividad de las variables explicativas.

C.1. Contraste de significatividad conjunta.

$$\begin{array}{l} H_0 : \beta_2 = \beta_3 = \beta_4 = \beta_5 = 0 \\ H_a : \beta_2 \neq 0 \text{ y/o } \beta_3 \neq 0 \text{ y/o } \beta_4 \neq 0 \text{ y/o } \beta_5 \neq 0 \end{array}$$

$$F = \frac{R^2/q}{(1-R^2)/(T-k)} \overset{H_0}{\sim} \mathcal{F}(q, T-k)$$

$$F = \frac{0,0,976745/4}{(1-0,976745)/(46-5)} = 430,506 \, > \, 2,59997 = \mathcal{F}(4,41)_{0{,}05}$$

Se rechaza la hipótesis nula con un nivel de significación del 5 %, es decir, se puede concluir que las variables explicativas capital, trabajo, energía y superfice cultivada son conjuntamente significativas.

C.2. Contraste de significatividad individual.

$$\begin{array}{l} H_0 : \beta_i = 0 \\ H_a : \beta_i \neq 0 \end{array} \qquad t = \frac{\hat{\beta}_i - 0}{\hat{\sigma}_{\hat{\beta}_i}} \overset{H_0}{\sim} t(T-k) \qquad i = 2,3,4,5.$$

Capital ($K$) $$|t| = \left| \frac{-0,3152-0}{0,0850853} \right| = 3,705 \, > \, 2,01954 = t(46-5)_{0{,}05/2}$$

Trabajo ($L$) $$|t| = \left| \frac{-0,720098-0}{0,038849} \right| = 18,54 \, > \, 2,01954 = t(41)_{0{,}05/2}$$

Energía ($E$) $$|t| = \left| \frac{0,191347-0}{0,0949839} \right| = 2,015 \, \approx \, 2,01954 = t(41)_{0{,}05/2}$$

Superficie ($H$) $$|t| = \left| \frac{1,08932-0}{0,364861} \right| = 2,986 \, > \, 2,01954 = t(41)_{0{,}05/2}$$

Por lo tanto, se puede concluir que, para un nivel de significación del 5 %, las variables explicativas capital, trabajo, energía y superfice cultivada son individualmente significativas.

2. Contraste de hipótesis de rendimientos constantes a escala.

$$H_0 : \beta_2 + \beta_3 + \beta_4 + \beta_5 = 1 \qquad F = \frac{(SCR_R - SCR_{NR})/q}{SCR_{NR}/(T-k)} \overset{H_0}{\sim} \mathcal{F}(q, T-k)$$
$$H_a : \beta_2 + \beta_3 + \beta_4 + \beta_5 \neq 1$$

donde $q = 1$ es el número de restricciones, $SCR_{NR}$ es la suma de cuadrados de residuos del modelo no restringido, Modelo P14.2 y $SCR_R$ es la suma de cuadrados de residuos del siguiente modelo restringido,

$$\log Q_t = \beta_1 + \beta_2 \log K_t + \beta_3 \log L_t + \beta_4 \log E_t + \beta_5 \log H_t + u_t$$

$$\begin{aligned} \log Q_t &= \beta_1 + (1 - \beta_3 - \beta_4 - \beta_5) \log K_t + \beta_3 \log L_t + \beta_4 \log E_t + \\ &+ \beta_5 \log H_t + u_t \end{aligned}$$

$$\begin{aligned} \log Q_t - \log K_t &= \beta_1 + \beta_3 (\log L_t - \log K_t) + \beta_4 (\log E_t - \log K_t) + \\ &+ \beta_5 (\log H_t - \log K_t) + u_t \end{aligned}$$

$$F = \frac{(0,074059 - 0,0629711)/1}{0,0629711/(46-5)} = 7,21936 > 4,07855 = \mathcal{F}_{0,05}(1, 46-5)$$

Se rechaza la hipótesis nula con un nivel de significación del 5 %, es decir, se puede concluir que la función de producción no tiene rendimientos constantes a escala.

## SEGUNDA PARTE.

1. Hipótesis de trabajo:

> *Para los mismos niveles de los factores de producción, el cambio tecnológico experimentado a partir de 1976 ha hecho que el nivel de producción sea mayor que antes de 1976.*

Para introducir este efecto se define la variable ficticia $A76$ que toma el valor 1 para las observaciones desde 1976 en adelante y 0 en otro caso.

La especificación del modelo es la siguiente:

$$\log Q_t = \beta_1 + \beta_2 \log K_t + \beta_3 \log L_t + \beta_4 \log E_t + \beta_5 \log H_t + \beta_6 A76_t + u_t$$
$$u_t \sim NID(0, \sigma_u^2) \qquad t = 1948, 1949, \ldots, 1993.$$

Modelo P14.3: estimaciones MCO utilizando las 46 observaciones 1948–1993
Variable dependiente: $l\_Q$

| Variable | Coeficiente | Desv. típica | Estadístico $t$ | valor p |
|---|---|---|---|---|
| const | 3,98004 | 1,19167 | 3,3399 | 0,0018 |
| $l\_K$ | -0,231053 | 0,0821327 | -2,8132 | 0,0076 |
| $l\_L$ | -0,646251 | 0,0427303 | -15,1239 | 0,0000 |
| $l\_E$ | 0,0844008 | 0,0931791 | 0,9058 | 0,3705 |
| $l\_H$ | 0,901690 | 0,337673 | 2,6703 | 0,0109 |
| $A76$ | 0,0632216 | 0,0205339 | 3,0789 | 0,0037 |

Suma cuadrados residuos 0,0509067 $R^2$ 0,981200

FRM:

$$\underset{(\hat{\sigma}_{\hat{\beta}})}{\widehat{\log Q_t}} = \underset{(\,1{,}19167)}{3,98004} - \underset{(0{,}0821327)}{0,231053} \log K_t - \underset{(0{,}0427303)}{0,646251} \log L_t + \underset{(0{,}0931791)}{0,0844008} \log E_t +$$

$$+ \underset{(\,0{,}337673)}{0,90169} \log H_t + \underset{(0{,}0205339)}{0,0632216}\, A76_t \quad R^2 = 0,9812 \quad SCR = 0,0509067$$

El coeficiente estimado $\hat{\beta}_6 = 0,0632216$ significa que, a igualdad de nivel de los factores de producción, hay una diferencia en la producción estimada (en logaritmos) de 0,0632212 toneladas a partir de 1976 con respecto al periodo anterior.

Se puede comprobar que este efecto es significativo:

$$\begin{array}{l} H_0 : \beta_6 = 0 \\ H_a : \beta_6 \neq 0 \end{array} \qquad t = \frac{\hat{\beta}_6 - 0}{\hat{\sigma}_{\hat{\beta}_6}} \overset{H_0}{\sim} t(T-k)$$

$$|t| = \left| \frac{0,0632216 - 0}{0,0205339} \right| = 3,0789 > 2,02108 = t(46-6)_{0{,}05/2}$$

Se rechaza la hipótesis nula con un nivel de significación del 5 %, es decir, se puede concluir que, para el mismo nivel de los factores, la producción media es diferente antes y después de 1976.

2. Hipótesis de trabajo.

*La función de producción es distinta antes y después de 1976, es decir, el efecto sobre la cantidad producida de maíz de un incremento en los factores de producción es diferente antes y después de 1976.*

Esta hipótesis implica la existencia de un cambio estructural en la función de producción a partir de 1976. Para poder contrastar esta hipótesis es preciso introducir en el modelo los términos de interacción entre cada uno de los factores de producción y la variable ficticia $A76$:

$$(\log K \times A76)_t = \begin{cases} 0 & t \in [1948 - 1975] \\ K_t & t \in [1976 - 1993] \end{cases} \qquad (\log L \times A76)_t = \begin{cases} 0 & t \in [1948 - 1975] \\ L_t & t \in [1976 - 1993] \end{cases}$$

$$(\log E \times A76)_t = \begin{cases} 0 & t \in [1948 - 1975] \\ E_t & t \in [1976 - 1993] \end{cases} \qquad (\log H \times A76)_t = \begin{cases} 0 & t \in [1948 - 1975] \\ H_t & t \in [1976 - 1993] \end{cases}$$

El modelo de regresión lineal general queda como sigue:

$$\begin{aligned} \log Q_t &= \beta_1 + \beta_2 \log K_t + \beta_3 \log L_t + \beta_4 \log E_t + \beta_5 \log H_t + \beta_6 A76_t + \\ &+ \beta_7 (\log K \times A76)_t + \beta_8 (\log L \times A76)_t + \beta_9 (\log E \times A76)_t + \\ &+ \beta_{10} (\log H \times A76)_t + u_t \quad u_t \sim NID(0, \sigma_u^2) \quad t = 1948, \ldots, 1993. \end{aligned}$$

La estimación MCO de este modelo proporciona los siguientes resultados:

Modelo P14.4: estimaciones MCO utilizando las 46 observaciones 1948–1993
Variable dependiente: $l\_Q$

| Variable | Coeficiente | Desv. típica | Estadístico $t$ | valor p |
|---|---|---|---|---|
| const | 6,35278 | 2,13003 | 2,9825 | 0,0051 |
| $l\_K$ | -0,0432245 | 0,154658 | -0,2795 | 0,7815 |
| $l\_L$ | -0,536663 | 0,0651691 | -8,2349 | 0,0000 |
| $l\_E$ | 0,0943920 | 0,240928 | 0,3918 | 0,6975 |
| $l\_H$ | 0,0770934 | 0,470323 | 0,1639 | 0,8707 |
| $A76$ | -5,8641 | 2,68453 | -2,1844 | 0,0355 |
| $l\_KA76$ | -0,320497 | 0,187181 | -1,7122 | 0,0955 |
| $l\_LA76$ | -0,414848 | 0,241894 | -1,7150 | 0,0949 |
| $l\_EA76$ | 0,0754497 | 0,303419 | 0,2487 | 0,8050 |
| $l\_HA76$ | 1,95036 | 0,676819 | 2,8817 | 0,0066 |

Suma cuadrados residuos 0,0400310 $R^2$ 0,985216

FRM:

$$\underset{(\hat\sigma_{\hat\beta})}{\widehat{\log Q_t}} = \underset{(2{,}13003)}{6,35278} - \underset{(0{,}154658)}{0,0432245} \log K_t - \underset{(0{,}0651691)}{0,536663} \log L_t + \underset{(0{,}240928)}{0,0943920} \log E_t +$$

$$+ \underset{(0{,}470323)}{0,0770934} \log H_t - \underset{(2{,}68453)}{5,8641} A76_t - \underset{(0{,}187181)}{0,320497} (\log K \times A76)_t -$$

$$- \underset{(\ 0{,}241894)}{0,414848} (\log L \times A76)_t + \underset{(0{,}303419)}{0,0754497} (\log E \times A76)_t +$$

$$+ \underset{(0{,}676819)}{1,95036} (\log H \times A76)_t \qquad R^2 = 0,985216 \quad SCR = 0,0400310$$

Contraste de la hipótesis de cambio estructural.

$$H_0 : \beta_6 = \beta_7 = \beta_8 = \beta_9 = \beta_{10} = 0$$
$$H_a : \beta_6 \neq 0 \text{ y/o } \beta_7 \neq 0 \text{ y/o } \beta_8 \neq 0 \text{ y/o } \beta_9 \neq 0 \text{ y/o } \beta_{10} \neq 0$$

$$F = \frac{(SCR_R - SCR_{NR})/q}{SCR_{NR}/(T-k)} \overset{H_0}{\sim} \mathcal{F}(q, T-k)$$

donde $SCR_{NR}$ es la suma de cuadrados de residuos del modelo no restringido, Modelo P14.4, $SCR_R$ es la suma de cuadrados de residuos del modelo restringido, Modelo P14.2, y $q = 5$ es el número de restricciones.

$$F = \frac{(0,062971 - 0,0400310)/5}{0,0400310/(46-10)} = 4,12602 > 2,47717 = \mathcal{F}(5, 46-10)_{0,05}$$

Se rechaza la hipótesis nula con un nivel de significación del 5 %, es decir, se puede concluir que hay cambio estructural en la función de producción a partir de 1976.

## TERCERA PARTE.

1. Dados los resultados de la Segunda Parte donde se ha comprobado que hay cambio estructural, el modelo más adecuado para la función de producción es el Modelo P14.4. El problema común a los modelos anteriores, Modelo P14.2 y Modelo P14.3, es que se supone que los coeficientes son constantes para toda la muestra, es decir, que las elasticidades capital, trabajo, energía y superficie son constantes desde 1948 hasta 1983. Se ha comprobado que esta hipótesis no se cumple, lo que implica que los estimadores MCO del Modelo P14.3 estarán sesgados, también la estimación de la varianza de las perturbaciones estará sesgada y los estadísticos $t$ y $F$ no seguirán las distribuciones habituales por lo que la inferencia no será válida.

2. Como el modelo es doble logarítmico, los resultados se van a proporcionar en logaritmos.

**Predicción por punto** :

$$
\begin{aligned}
\widehat{Q}_{1994} &= 6,35278 - 0,0432245 \log K_{1994} - 0,536663 \log L_{1994} + \\
&+ 0,0943920 \log E_{1994} + 0,0770934 \log H_{1994} - 5,8641\, A76_{1994} - \\
&- 0,320497\, (\log K \times A76)_{1994} - 0,414848\, (\log L \times A76)_{1994} \\
&+ +0,0754497\, (\log E \times A76)_{1994} + 1,95036\, (\log H \times A76)_{1994} \\
&= 6,35278 - 0,0432245 \log K_{1994} - 0,536663 \log L_{1994} + \\
&+ 0,0943920 \log E_{1994} + 0,0770934 \log H_{1994} - 5,8641 - \\
&- 0,320497 \log K_{1994} - 0,414848 \log L_{1994} + \\
&+ +0,0754497 \log E_{1994} + 1,95036 \log H_{1994} \\
&= (6,35278 - 5,8641) - (0,0432245 \ + \ 0,320497) \log K_{1994} - \\
&- (0,536663 \ + \ 0,414848) \log L_{1994} + \\
&+ (0,0943920 \ + \ 0,0754497) \log E_{1994} + (0,0770934 \ + \ 1,95036) \log H_{1994} \\
&= (6,35278 - 5,8641) - (0,0432245 \ + \ 0,320497) \log 59 - \\
&- (0,536663 \ + \ 0,414848) \log 79 + (0,0943920 \ + \ 0,0754497) \log 92 + \\
&+ (0,0770934 \ + \ 1,95036) \log 87 = 4,670392
\end{aligned}
$$

**Predicción por intervalo**

$$
IC(Q_{1994})_{95\,\%} \;=\; \left[\widehat{Q}_{1994} \pm t(46-10)_{0,05/2}\ \hat{\sigma}_{1994}\right] = [4,590395 \quad 4,75039]
$$

dado que

$$
\hat{\sigma}^2_{1994} = \hat{\sigma}^2_u\,[1 + X'_{1994}(X'X)^{-1}X'_{1994}] = 0,001556
$$

con

$$
X'_{1994} = (1,\ \log 59,\ \log 79,\ \log 92,\ \log 87) = (1\ 4,077\ 4,369\ 4,5217\ 4,4659)
$$

# Solución PRÁCTICA P15.

## PRIMERA PARTE.

1. MRLG:

$$V_t = \beta_1 + \beta_2\, PRAL_t + \beta_3\, PRIS_t + \beta_4\, PRCA_t + u_t \quad u_t \sim NID(0, \sigma_u^2) \quad t = 1, \ldots, 52.$$

Modelo P15.1: estimaciones MCO utilizando las 52 observaciones 1–52 Variable dependiente: V

| Variable | Coeficiente | Desv. típica | Estadístico $t$ | valor p |
|---|---|---|---|---|
| const | 22963,4 | 9806,55 | 2,3416 | 0,0234 |
| $PRAL$ | -47084, | 7957,79 | -5,9168 | 0,0000 |
| $PRIS$ | 9299,00 | 7001,26 | 1,3282 | 0,1904 |
| $PRCA$ | 16511,3 | 9367,03 | 1,7627 | 0,0843 |

| | | | |
|---|---|---|---|
| Suma cuadrados residuos | 1,35872e+09 | $R^2$ | 0,442860 |

FRM:

$$\underset{(\hat\sigma_{\hat\beta})}{\widehat{V}_t} = \underset{(9806{,}55)}{22963,4} - \underset{(7957{,}79)}{47084}\; PRAL_t + \underset{(7001{,}26)}{9299,00}\; PRIS_t + \underset{(9367{,}03)}{16511,3}\; PRCA_t$$

$$R^2 = 0,442860 \quad SCR = 1358720000$$

**A.** Interpretación de los coeficientes estimados:

$\hat\beta_2$ = -47084. Si el precio de la marca AL aumenta un euro, manteniéndose fijos los precios de las marcas competidoras, se estima que las ventas disminuyen en 47084 latas.

$\hat\beta_3$ = 9299,00. Si el precio de la marca competidora IS aumenta un euro, manteniéndose fijos los precios de las marcas AL y CA, se estima que las ventas aumentan en 9299 latas.

$\hat\beta_4$ = 16511,3. Si el precio de la marca competidora CA aumenta un euro, manteniéndose fijos los precios de las marcas AL e IS, se estima que las ventas aumentan en 16511,3 latas.

**B.** Coeficiente de determinación: $R^2$ = 0,4486.

Interpretación: el 44,86 % de la variabilidad de las ventas en la muestra está explicada por la variabilidad de los precios, tanto de la propia marca, AL, como de las competidoras, IS y CA.

2. Hipótesis: el efecto de las políticas de precios de las dos marcas competidoras es el mismo sobre las ventas medias de la marca AL.

$$\begin{array}{l} H_0 : \beta_3 = \beta_4 \\ H_a : \beta_3 \neq \beta_4 \end{array} \qquad F = \frac{(SCR_R - SCR_{NR})/q}{SCR_{NR}/(T-k)} \overset{H_0}{\sim} \mathcal{F}(q, T-k)$$

donde $q = 1$ es el número de restricciones, $SCR_{NR}$ es la suma de cuadrados de residuos del modelo no restringido, Modelo P15.1, $SCR_R$ es la suma de cuadrados de residuos del siguiente modelo restringido:

$$\begin{aligned} V_t &= \beta_1 + \beta_2\, PRAL_t + \beta_3\, PRIS_t + \beta_3\, PRCA_t + u_t \quad \rightarrow \\ V_t &= \beta_1 + \beta_2\, PRAL_t + \beta_3\, (PRIS_t + PRCA_t) + u_t \end{aligned} \qquad \text{(P15.2)}$$

El estadístico muestral es:

$$F = \frac{(1368430000 - 1358720000)/1}{1358720000/(52-4)} = 0,343166 \; < \; 4,04265 = \mathcal{F}(1,48)_{0,05}$$

No se rechaza la hipótesis nula con un nivel de significación del 5 %, es decir, se puede concluir que el efecto de las políticas de precios de las dos marcas competidoras es el mismo sobre las ventas medias de la marca AL.

3. Dado el resultado del constrate del apartado anterior, se eligiría el Modelo P15.2 porque los estimadores de Mínimos Cuadrados Restringidos son insesgados y de varianza más pequeña que los estimadores MCO del Modelo P15.1.

   Ahora bien, este resultado desde un punto de vista económico se puede interpretar como que las ventas de la marca AL se ven afectadas por su propio precio y por el precio conjunto promedio de los competidores, sin distinguir entre ellos. En vez de utilizar como medida general de los precios de los competidores la suma de los precios $PRIS$ y $PRCA$ como en el Modelo P15.2, vamos a utilizar un índicador de precios de los competidores que definimos como el precio promedio de las dos marcas, IS y CA:

$$PRCOMP_t = \frac{(PRIS_t + PRCA_t)}{2}$$

   Estimamos por MCO el modelo de regresión lineal:

$$V_t = \beta_1 + \beta_2\, PRAL_t + \beta_3\, PRCOMP_t + u_t \qquad \text{(P15.3)}$$

   con los siguientes resultados:

Modelo 3: MCO, usando las observaciones 1–52
Variable dependiente: V

| | Coeficiente | Desv. Típica | Estadístico $t$ | Valor p |
|---|---|---|---|---|
| const | 25002,5 | 9106,37 | 2,7456 | 0,0084 |
| PRAL | −47863,7 | 7793,05 | −6,1419 | 0,0000 |
| PRCOMP | 23967,8 | 10514,0 | 2,2796 | 0,0270 |

Suma cuadrados residuos 1368430000 $R^2$ 0,438877

FRM:

$$\underset{(\hat{\sigma}_{\hat{\beta}})}{\widehat{V}_t} = \underset{(9106{,}37)}{25002,5} - \underset{(7793{,}05)}{47863,7}\, PRAL_t + \underset{(10514{,}0)}{23967,8}\, PRCOMP_t$$

$$R^2 = 0,438877 \qquad SCR = 1368430000$$

**A.** Interpretación de los coeficientes estimados:

$\hat{\beta}_2$ = -47863,7. Si el precio de la marca AL aumenta un euro, manteniéndose fijo el índicador de precios de los competidores, se estima que las ventas disminuyen en 47863,7 latas.

$\hat{\beta}_3$ = 23967,8. Si el indicador de precios de los competidores aumentan un euro, manteniéndose fijo el precio de la marca Al, se estima que las ventas aumentan en 23967,8 latas.

**B.** Coeficiente de determinación: $R^2$ = 0,438877.

Interpretación: el 43,89 % de la variabilidad de las ventas en la muestra está explicada por la variabilidad de los precios, tanto de la propia marca como de las competidoras.

**C.** Significatividad de las variables explicativas.

C.1. Contraste de significatividad conjunta.

$$\begin{array}{l} H_0 : \beta_2 = \beta_3 = 0 \\ H_a : \beta_2 \neq 0 \text{ y/o } \beta_3 \neq 0 \end{array} \qquad F = \frac{R^2/q}{(1-R^2)/(T-k)} \overset{H_0}{\sim} \mathcal{F}(q, T-k)$$

$$F = \frac{0,438877/2}{(1-0,438877)/(52-3)} = 19,1624 \; > \; 3,18658 = \mathcal{F}(2,49)_{0{,}05}$$

Se rechaza la hipótesis nula con un nivel de significación del 5 %, es decir, se puede concluir que las variables explicativas precio de la marca y el indicador de precios de los competidores son conjuntamente significativas.

C.2. Contraste significatividad individual.

$$\begin{array}{l} H_0 : \beta_i = 0 \\ H_a : \beta_i \neq 0 \end{array} \qquad t = \frac{\hat{\beta}_i - 0}{\hat{\sigma}_{\hat{\beta}_i}} \overset{H_0}{\sim} t(T-k) \qquad i = 2, 3.$$

$$\text{Precio AL} \qquad |t| = \left| \frac{-47863,7 - 0}{7793,05} \right| = 6,142 \; > \; 2,00958 = t(52-3)_{0,05/2}$$

$$\text{Precio PRCOMP} \quad |t| = \left| \frac{23967,8 - 0}{10514,0} \right| = 2,28 \; > \; 2,00958 = t(52-3)_{0,05/2}$$

Por lo tanto, se puede concluir que, para un nivel de significación del 5 %, las variables precio de la marca y indicador de precios de los competidores son individualmente significativas.

4. Predicción por intervalo de las ventas en la semana 53.

$$IC(V_{53})_{95\,\%} = \left[ \widehat{V}_{53} \; \pm \; t(52-3)_{0,05/2} \; \hat{\sigma}_u \sqrt{1 + X'_{53}(X'X)^{-1}X_{53}} \right]$$

donde:

$$\begin{array}{rcl} X'_{53} & = & (1 \quad 0,82 \quad 0,965) \\ \widehat{V}_{53} & = & 25002,5 - 47863,7 \times 0,82 + 11983,9 \times 0,965 = 8883,11 \text{ latas.} \end{array}$$

$$IC(V_{53})_{95\,\%} = [8883,11 \; \pm \; 2,00958 \times 5595,863] = [-2362,19 \quad 20128,42]$$

Las ventas en la semana 52 fueron de 6668 latas. Si el gerente piensa que las ventas pueden aumentar en 50 unidades, quiere decir que en la semana 53 podrían ascender a 6718 unidades. Como se puede observar $6718 \in IC(V_{53})_{95\,\%}$ por lo tanto, con una confianza del 95 %, existe evidencia en la muestra a favor de la hipótesis del gerente.

## SEGUNDA PARTE.

1. MRLG:

$$V_t = \beta_1 + \beta_2\, PRAL_t + \beta_3\, PRCOMP_t + \beta_4\, CA_t + \beta_5\, CB_t + u_t \qquad t = 1, \ldots, 52.$$

Cada campaña publicitaria es independiente de la otra por lo que cada una de ellas se puede considerar como una variable explicativa cualitativa. Cada una de estas variables explicativas tiene dos categorías: campaña y no campaña, por lo que habría que definir dos variables ficticias por cada una de ellas:

Campaña A $\quad CA_t = \begin{cases} 1 & i \in \text{campaña A} \\ 0 & \text{en otro caso} \end{cases} \quad NOCA_t = \begin{cases} 1 & i \in \text{no campaña A} \\ 0 & \text{en otro caso} \end{cases}$

Campaña B $\quad CB_t = \begin{cases} 1 & i \in \text{campaña B} \\ 0 & \text{en otro caso} \end{cases} \quad NOCB_t = \begin{cases} 1 & i \in \text{no campaña B} \\ 0 & \text{en otro caso} \end{cases}$

Para evitar problemas de colinealidad perfecta en la matriz de datos $X$, el modelo se ha especificado con término constante y 1 categoría sólo, la correspondiente al periodo de campaña, para cada una de las variables explicativas cualitativas.

Los resultados de la estimación mínimo-cuadrática ordinaria del modelo de regresión son los siguientes:

Modelo P15.4: estimaciones MCO utilizando las 52 observaciones 1–52
Variable dependiente: V

| Variable | Coeficiente | Desv. típica | Estadístico $t$ | valor p |
|---|---|---|---|---|
| const | 27718,3 | 8325,55 | 3,3293 | 0,0017 |
| $PRAL$ | -50671,0 | 7221,18 | -7,0171 | 0,0000 |
| $PRCOMP$ | 19583,0 | 9448,74 | 2,0726 | 0,0437 |
| $CA$ | 5496,27 | 1454,55 | 3,7787 | 0,0004 |
| $CB$ | 3622,38 | 1988,17 | 1,8220 | 0,0748 |

Suma cuadrados residuos 1,04429e+09 $\quad R^2$ 0,571792

FRM:

$$\underset{(\hat{\sigma}_{\hat{\beta}})}{\widehat{V}_t} = \underset{(8325{,}55)}{27718,3} - \underset{(7221{,}18)}{50671,0}\, PRAL_t + \underset{(9448{,}74)}{19583,0}\, PRCOMP_t + \underset{(1454{,}55)}{5496,27}\, CA_t +$$

$$+ \underset{(1988{,}17)}{3622,38}\, CB_t \qquad R^2 = 0,571792 \qquad SCR = 1044290000$$

**A.** Interpretación de los coeficientes estimados:

$\hat{\beta}_1$ = 27718,3. Ventas estimadas cuando los precios de la propia marca y de los competidores son cero y no se está llevando a cabo ninguna campaña publicitaria.

$\hat{\beta}_2$ = -50671. Se estima que las ventas disminuyen en 50671 latas, si el precio de la marca AL aumenta un euro, manteniéndose fijos el indicador de los precios de los competidores, e independientemente de si están en marcha las campañas publicitarias A y/o B.

$\hat{\beta}_3$ = 9791,51. Se estima que las ventas aumentan en 9791,51 latas si los precios de las marcas competidoras, IS y CA, aumentan un euro en total, manteniéndose fijo el precio de la marca Al, e independientemente de si están en marcha las campañas publicitarias A y/o B.

$\hat{\beta}_4$ = 5496,27. Diferencia estimada en las ventas en el periodo que hay campaña A y en el que no la hay a igualdad de condiciones de precios e independientemente de si está en marcha la campaña publicitaria B.

$\hat{\beta}_5$ = 3622,38. Diferencia estimada en las ventas en el periodo que hay campaña B y en el que no la hay a igualdad de condiciones de precios e independientemente de si está en marcha la campaña publicitaria A.

**B.** Coeficiente de determinación: $R^2 = 0{,}571792$.

Interpretación: el 51,18 % de la variabilidad de las ventas en la muestra está explicada por la variabilidad de los precios, tanto de la propia marca como de los competidores, y las campañas publicitarias, A y B.

**C.** Significatividad de las variables explicativas.

C.1. Contraste de significatividad conjunta.

$$H_0 : \beta_2 = \beta_3 = \beta_4 = \beta_5 = 0$$
$$H_a : \beta_2 \neq 0 \text{ y/o } \beta_3 \neq 0 \text{ y/o } \beta_4 \neq 0 \text{ y/o } \beta_5 \neq 0$$

$$F = \frac{R^2/q}{(1-R^2)/(T-k)} \overset{H_0}{\sim} \mathcal{F}(q, T-k)$$

$$F = \frac{0,571792/4}{(1-0,571792)/(52-5)} = 15,69 \; > \; 2,56954 = \mathcal{F}(4,47)_{0{,}05}$$

Se rechaza la hipótesis nula con un nivel de significación del 5 %, es decir, se puede concluir que las variables explicativas precio de la marca, indicador de precios de los competidores, campaña A y campaña B son conjuntamente significativas.

C.2. Contraste de significatividad individual.

$$\begin{array}{l} H_0 : \beta_i = 0 \\ H_a : \beta_i \neq 0 \end{array} \qquad t = \frac{\hat{\beta}_i - 0}{\hat{\sigma}_{\hat{\beta}_i}} \overset{H_0}{\sim} t(T-k) \qquad i = 2,3,4,5.$$

$$\text{Precio AL} \qquad |t| = \left| \frac{-50671 - 0}{7221,18} \right| = 7,0171 \ > \ 2,00958 = t(52-3)_{0,05/2}$$

$$\text{PRCOMP} \qquad |t| = \left| \frac{19583,0 - 0}{9448,740} \right| = 2,0726 \ > \ 2,00958 = t(52-3)_{0,05/2}$$

$$\text{Campaña A} \qquad |t| = \left| \frac{5496,27 - 0}{1454,55} \right| = 3,7787 \ > \ 2,00958 = t(52-3)_{0,05/2}$$

$$\text{Campaña B} \qquad |t| = \left| \frac{3622,38 - 0}{1988,17} \right| = 1,8220 \ < \ 2,00958 = t(52-3)_{0,05/2}$$

Por lo tanto, se puede concluir que, para un nivel de significación del 5 %, las variables precio de la marca, indicador de precios de los competidores y la campaña publicitaria A son individualmente significativas. Notése que la campaña publicitaria B no es una variable individualmente significativa para un nivel de significación del 5 % aunque sí lo es al 10 %.

2. Vistos los resultados del apartado anterior, le aconsejaría al gerente que llevara a cabo la campaña publicitaria A por dos razones:
   - La campaña A sí es significativa con un nivel de significación del 5 %, mientras que la campaña B no lo es, aunque sí lo es para un nivel de significación del 10 %.
   - El incremento de las ventas que genera la campaña A es, a igualdad de condiciones, superior al de la campaña B.

**TERCERA PARTE.**

1. El Modelo P15.4 porque en el modelo Modelo P15.3 no se ha incluido el efecto de las campañas publicitarias sobre las ventas, efecto que es estadísticamente significativo. Por lo tanto, los estimadores MCO de los coeficientes del Modelo P15.3 están sesgados, y la inferencia realizada de la forma habitual no es válida porque el estimador de la varianza de las perturbaciones también está sesgado y los estadísticos $t$ y $F$ no siguen las distribuciones habituales.

2. Sobre el tema de las políticas de marketing, por un lado, hay que resaltar que son efectivas, por lo que son una buena medida para aumentar las ventas, y además las más efectivas son las políticas del tipo de la campaña A.

   En lo que se refiere a políticas de precio, sabemos que el efecto de los precios de las empresas competidoras sobre las ventas de la marca AL es el mismo. Pero interesaría saber también si, por ejemplo, la empresa puede contrarrestar una bajada del indicador de precios de los competidores con una bajada de la misma cuantía en su propio precio. El contraste se lleva a cabo en el Modelo P15.4.

$$\begin{array}{ll} H_0 : \beta_2 = -\beta_3 & (\beta_2 + \beta_3 = 0) \\ H_a : \beta_2 > -\beta_3 & (\beta_2 + \beta_3 > 0) \end{array} \qquad t = \frac{\hat{\beta}_2 + \hat{\beta}_3 - 0}{\sqrt{\widehat{var}(\hat{\beta}_2 + \hat{\beta}_3)}} \overset{H_0}{\sim} t(T-k)$$

donde:

$$\widehat{var}(\hat{\beta}_2 + \hat{\beta}_3) = \widehat{var}(\hat{\beta}_2) + \widehat{var}(\hat{\beta}_3) + 2\widehat{cov}(\hat{\beta}_2, \hat{\beta}_3) = 111083445{,}5$$

$$t = \frac{-50671 + 19583 - 0}{\sqrt{111083445{,}5}} = -2,9496 \; < \; 1,67793 = t(47)_{0{,}05}$$

No se rechaza la hipótesis nula con un nivel de significación del 5 %, es decir, se puede concluir que la empresa puede compensar la bajada del indicador de precios de los competidores con una bajada similar en su propio precio.

# Bibliografía


[1] A. Cottrell and R.J. Lucchetti. *Gretl*. http://gretl.sourceforge.net/.

[2] D.N. Gujarati. *Basic Econometrics*. McGraw-Hill, United States, 2003.

[3] R.C. Hill, W.E. Griffiths, and G.C. Lim. *Principles of Econometrics*. Wiley, United States, 2008.

[4] G. Koop. *Analysis of Economic Data*. Wiley, England, 2002.

[5] R. Ramanathan. *Introductory Econometrics with Applications*. South-Western, Thomson Learning, United States, 2002.

[6] J.M. Wooldridge. *Introductory Econometrics.A modern approach.* South-Western, Thomson Learning, United States, 2003.