Hoja de Ejercicios 3
El modelo de regresión lineal múltiple

Nota: En aquellos ejercicios en los que se incluyen estimaciones y referencia al archivo de datos utilizado, el estudiante debería comprobar los resultados obtenidos en Gretl.

1. Sea el modelo

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon,$$

donde $E(\varepsilon|X_1, X_2) = 0$, y suponga que disponemos de una muestra de tamaño n.

a) Derive las condiciones de primer orden de los estimadores MCO $\widehat{\beta}_0$, $\widehat{\beta}_1$, $\widehat{\beta}_2$ de los coeficientes β.

b) Muestre que

$$\begin{aligned} \widehat{\beta}_0 &= \overline{Y} - \widehat{\beta}_1 \overline{X}_1 - \widehat{\beta}_2 \overline{X}_2 \\ \widehat{\beta}_1 &= \frac{1}{D}(s_{22}s_{1y} - s_{12}s_{2y}) \\ \widehat{\beta}_2 &= \frac{1}{D}(s_{11}s_{2y} - s_{12}s_{1y}) \end{aligned}$$

donde

$$\begin{array}{lll} s_{11} = \frac{1}{n}\sum_i x_{1i}^2, & s_{12} = \frac{1}{n}\sum_i x_{1i}x_{2i} = s_{21}, & s_{1y} = \frac{1}{n}\sum_i x_{1i}y_i \\ s_{22} = \frac{1}{n}\sum_i x_{2i}^2, & & s_{2y} = \frac{1}{n}\sum_i x_{2i}y_i \\ D = s_{11}s_{22} - s_{12}^2. & & \end{array}$$

siendo $y_i = Y_i - \overline{Y}$, $x_{1i} = X_{1i} - \overline{X}_1$, $x_{2i} = X_{2i} - \overline{X}_2$.

c) ¿Qué ocurre cuando $D = 0$? Interprete el resultado.

d) Muestre que, cuando $s_{12} = 0$, el estimador $\widehat{\beta}_1$ coincide con el estimador $\widehat{\gamma}_1$ en la regresión simple

$$\widehat{Y} = \widehat{\gamma}_0 + \widehat{\gamma}_1 X_1,$$

e interprete el resultado.

2. ¿Cuál de las siguientes situaciones, si alguna, incumpliría los supuestos del modelo de regresión clásico?

a) La variable X_2 es el recíproco de la variable X_1.

b) La variable X_2 es el cuadrado de la variable X_1.

c) La variable X_1 es una variable artificial que toma el valor 1 para mujeres y 0 para hombres, y la variable X_2 es una variable artificial que toma el valor 1 para hombres, y 0 para mujeres.

3. Aunque el vino es un bien de consumo, dadas las características de los grandes vinos de reserva puede tener sentido considerarlos como una inversión. En particular, disponemos de datos de los precios de subasta de miles de vinos tintos de reserva de Burdeos de añadas entre 1952 y 1980. Estos vinos se almacenan durante bastante tiempo antes de ser consumidos, lo que incrementa su precio dado el coste de almacenaje, que supone un coste de oportunidad dada la posibilidad de inversiones alternativas. Nuestro fichero de datos `BORDEAUX.GDT` proporciona, para distintas añadas, información sobre LPR (logaritmo neperiano del precio del vino), *lluvinv* (Cantidad de lluvia caída en el invierno anterior a la cosecha), *tempmed* (Temperatura media durante el período de maduración de la uva), *lluvcos* (Cantidad de lluvia caída durante el período de maduración de la uva), *edad* (Número de años transcurridos desde la cosecha).

   a) Estime la proyección lineal de *lpr* sobre *edad*. Dados los resultados, ¿cuál sería la tasa de rentabilidad anual derivada de conservar el vino?

   b) Efectúe la regresión múltiple de *lpr* sobre *edad*, *lluvinv*, *lluvcos*, *tempmed*. ¿Cómo cambia la estimación de la tasa de rentabilidad de conservar el vino? ¿Cómo explica las diferencias? Explique qué (**Pista**: ¿Qué factores pueden afectar a la calidad del vino?).

4. Suponga que $\ln(wage)$ mide el logaritmo del salario mensual y *educ* el número de años de educación. Considere el modelo lineal

$$\ln(wage) = \gamma_0 + \gamma_1 educ + v.$$

   El archivo `WAGE2.GDT`, de Blackburn and Neumark (1992), contiene información sobre el salario mensual, la educación, características socieconómicas e IQ (coeficiente de inteligencia) de 935 varones estadounidenses en 1980.

   a) ¿Es razonable suponer que $E(v|educ) = 0$? ¿Qué otros factores estarían incluidos en v?

   b) Considere el modelo

$$\ln(wage) = \beta_0 + \beta_1 educ + \beta_2 abil + u,$$

   donde *abil* es el coeficiente de inteligencia (IQ) y u satisface el supuesto $E(u|educ, abil) = 0$. Interprete el coeficiente β_1.

   c) Dado que la muestra de la que inicialmente se dispone no contiene información sobre el coeficiente de inteligencia de los trabajadores, se estima la proyección lineal

$$\begin{aligned} \ln(\widehat{wage}) &= 5{,}97 + \underset{(0{,}006)}{0{,}06}\, educ \\ n &= 935, R^2 = 0{,}097, \sum\nolimits_i \widehat{v}_i^2 = 149{,}52. \end{aligned}$$

   Explique bajo qué condiciones la estimación MCO de la pendiente de dicha proyección lineal es un estimador insesgado de β_1. Proporcione un intervalo de confianza al 95 % para la pendiente del modelo estimado.

   d) Suponga ahora que se consigue información sobre el coeficiente de inteligencia de los trabajadores de la muestra, obteniéndose la siguiente estimación,

$$\begin{aligned} \ln(\widehat{wage}) &= 5{,}66 + \underset{(0{,}007)}{0{,}04}\, educ + \underset{s_{\hat{\beta}_2}}{0{,}0059} abil \\ n &= 935, R^2 = 0{,}130 \end{aligned}$$

   Contraste al nivel de significación del 5 % que la inteligencia no afecta el salario.

e) Obtenga la covarianza muestral entre el nivel educativo y el coeficiente de inteligencia. A la vista de los resultados, interprete el parámetro asociado a la educación en las dos regresiones.

f) Si $E(u^2|educ, abil) = \sigma^2 (educ)$, donde $\sigma^2 (educ)$ es una función positiva de *educ*, ¿cambiaría alguna de sus respuestas? ¿Cuáles serían las consecuencias del incumplimiento del supuesto de homocedasticidad?

5. La variable *rdintens* son los gastos en investigación y desarrollo (R&D) en porcentaje sobre las ventas. Las ventas, *sales*, se miden en millones de dólares. La variable *profmarg* son los beneficios en porcentaje sobre las ventas. Usando el archivo de datos `RDCHEM.GDT`, que contiene 32 empresas del sector químico en EE.UU., se ha obtenido la siguiente ecuación

$$\begin{aligned} \widehat{rdintens} &= 0{,}472 + \underset{(0{,}216)}{0{,}321} \ln(sales) + \underset{(0{,}046)}{\hat{\beta}_2} \; profmarg \\ n &= 32, \quad R^2 = 0{,}098 \end{aligned}$$

a) Interprete el coeficiente de $\ln(sales)$: si las ventas se incrementan en un 10 %, ¿cuál es el cambio promedio estimado, *ceteris paribus*, en *rdintens*? ¿Es un efecto económicamente importante? ¿Es estadísticamente significativo?

b) Se ha estimado el siguiente modelo alternativo utilizando la misma base de datos:

$$\begin{aligned} \widehat{rdintens} &= 1{,}104 + \underset{(0{,}216)}{0{,}302} \ln(sales) \\ n &= 32, \quad R^2 = 0{,}061 \end{aligned}$$

¿Es el coeficiente de *profmarg* en el modelo del apartado (a) significativo? Si sabemos que $\hat{\beta}_2$ es positivo, ¿cuál es su valor?

c) Considere este modelo que relaciona los beneficios con las ventas,

$$\begin{aligned} \widehat{profm}\arg &= 7{,}34 + \underset{s_{\hat{\gamma}_1}}{\hat{\gamma}_1} \ln(sales)\,. \\ n &= 32, \quad R^2 = 0{,}0069 \end{aligned}$$

¿Cuál es el signo y el valor de $\hat{\gamma}_1$? ¿Es significativo? Obtenga $s_{\hat{\gamma}_1}$.

6. Para estudiar el efecto de la asistencia a clase sobre las notas finales de una asignatura, se ha utilizado el archivo de datos `ATTEND.GDT` con 680 alumnos de Introducción a Microeconomía en una universidad americana para ajustar el siguiente modelo de regresión:

$$\begin{aligned} stndfnl &= \beta_0 + \beta_1 atndrte + \beta_2 priGPA + \beta_3 priGPA^2 \\ &\quad + \beta_4 ACT + \beta_5 ACT^2 + \beta_6 (priGPA \times atndrte) + U, \end{aligned} \tag{1}$$

donde *stndfnl* es el resultado de un examen final estandarizado, *atndrte* es el porcentaje de asistencia a clase, *priGPA* es la nota media obtenida en los cursos anteriores y *ACT* es la nota de acceso a la Universidad. Los resultados de la estimación son los siguientes:

OLS, using observations 1–680
Dependent variable: *stndfnl*

	Coefficient	Std. Error	t-ratio	p-value
const	2,0503	1.3603	1,5072	0.1322
$atndrte$	−0,0067	0.0102	−0,6561	0.5120
$priGPA$	−1,6285	0.4810	−3,3857	0.0008
$priGPA^2$	0,2959	0.1010	2,9283	0.0035
ACT	−0,1280	0.0985	−1,3000	0.1940
ACT^2	0,0045	0.0022	2,0829	0.0376
$priGPA \times atndrte$	0,0056	0.0043	1,2938	0.1962

Mean dependent var	0.029659	S.D. dependent var	0.989461
Sum squared resid	512.7624	S.E. of regression	0.872872
R^2	0.228654	Adjusted R^2	0.221777
$F(6, 673)$	33.25018	P-value(F)	3.49e–35

a) ¿Cuál es el efecto parcial estimado de la asistencia a clase sobre el examen final? ¿Se puede concluir que el efecto es significativo para un alumno cuya nota media en los cursos anteriores era igual a 3 si $\widehat{Cov}\left(\hat{\beta}_1, \hat{\beta}_6\right) = 0{,}0001$?

b) Si añadimos el término $\beta_7 \ \left(ACT \times atndrte^2\right)$ a la ecuación (1), ¿Cuál sería el efecto parcial de asistir a clase en el modelo en términos de los parámetros desconocidos?

c) Explique cómo contrastaría que el modelo es lineal en todas las variables. Exprese con claridad todos los pasos que ha de seguir para realizar el contraste, es decir: hipótesis nula y alternativa, estadístico a utilizar (explicando sus componentes con claridad y regla de decisión.

7. Utilizando las 526 observaciones de `WAGE1.GDT` sobre asalariados en EE.UU. en 1976, consideramos la siguiente especificación:

$$\ln(wage) = \beta_0 + \beta_1 educ + \beta_2 exper + \beta_3 tenure + \varepsilon$$

donde *wage* es el salario-hora, *educ* denota años de educación, *exper* es la experiencia profesional (en años) y *tenure* es la antigüedad en el puesto de trabajo actual (en años). Además, suponga que se verifica que $E\left(\varepsilon | \, educ, exper, tenure\right) = 0$.

OLS, using observations 1–526
Dependent variable: *ln(wage)*

	Coefficient	Std. Error	t-ratio	p-value
const	0,2844	0.1042	2,7292	0.0066
educ	0,0920	0.0073	12,5552	0.0000
exper	0,0041	0.0017	2,3914	0.0171
tenure	0,0221	0.0031	7,1331	0.0000

Mean dependent var	1.623268	S.D. dependent var	0.531538
Sum squared resid	101.4556	S.E. of regression	0.440862
R^2	0.316013	Adjusted R^2	0.312082
$F(3, 522)$	80.39092	P-value(F)	9.13e–43

a) ¿Cuál es la interpretación de β_1? Dadas las estimaciones, ¿cuál es la diferencia salarial media entre dos personas con igual experiencia y antigüedad que difieren en un año de educación? ¿Tiene una interpretación ceteris paribus?

b) Las pendientes del modelo de regresión múltiple miden el efecto de la variable correspondiente manteniéndose constantes las demás. Vamos a comprobarlo para el coeficiente de la educación, realizando los siguientes pasos:

(i) Estime la proyección lineal de *educ* sobre *exper* y *tenure*.
Después, guarde los residuos en una nueva variable, *res_ educ*. Para ello, en el Menú superior donde aparecen las estimaciones, escoja Save –> Residuals. Aparecerá una ventana, en la que puede cambiar el nombre de la nueva variable.
¿Cómo son los coeficientes? ¿En qué dirección varían *exper* y *educ* cuando varía *educ*?
¿Qué información contienen los residuos? (**Pista**: recuerde que la proyecciópn lineal descompone aditivamente la variable dependiente en la parte explicada y la parte no explicada por dicha proyección lineal).

(ii) Estime la proyección lineal de *ln(wage)* sobre *res_ educ*.
Compare el coeficiente de *res_ educ* en esta última estimación con el de *educ* en la regresión de $\ln(wage)$ sobre *educ*, *exper* y *tenure*.
¿Qué conclusión podemos obtener?

8. El análisis de regresión puede utilizarse para contrastar si el mercado hace un uso eficiente de la información a la hora de valorar las acciones. En concreto, sea *return* el rendimiento total de las acciones de una empresa a lo largo de un periodo de cuatro años, desde finales de 1990 hasta finales de 1994. La hipótesis de eficiencia del mercado dice que este rendimiento no debería estar relacionado de manera sistemática con la información conocida en 1990. Si las características de la empresa conocidas al principio del periodo fuesen de ayuda para predecir el rendimiento del mercado, entonces podríamos usar esta información para seleccionar unas acciones u otras.

Para 1990, sea *dkr* el cociente del endeudamiento de la empresa en relación a su capital, sean *eps* las ganancias por acción, *netinc* la renta neta y *salary* la remuneración total del director general.

a) Usando los datos de `RETURN.GDT`, estime la siguiente ecuación:

$$return = \beta_0 + \beta_1 dkr + \beta_2 eps + \beta_3 netinc + \beta_4 salary + u.$$

Contraste si las variables explicativas son conjuntamente significativas al 5 %. Contraste si *netinc* y *salary* son conjuntamente significativas. ¿Hay alguna variable explicativa que sea individualmente significativa?

b) Reestime ahora el modelo que aplica logaritmos a *netinc* y *salary*,

$$return = \beta_0 + \beta_1 dkr + \beta_2 eps + \beta_3 \log(netinc) + \beta_4 \log(salary) + u.$$

¿Cómo cambian las conclusiones del apartado *(a)*?

c) ¿Por qué no se aplican logaritmos a *dkr* y *eps* en el apartado *(b)*?

d) En términos generales, ¿la evidencia a favor de la predictibilidad del rendimiento de las acciones es fuerte o débil?

9. Considere el modelo de regresión lineal

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} + \beta_5 X_{5i} + \varepsilon_i.$$

Explique exactamente cómo contrastaría las siguientes hipótesis:

a) $\beta_1 = 0$.

b) $\beta_1 = 0$ y $\beta_4 = \beta_5$

c) $\beta_1 = 0$, $\beta_3 = 2$, y $\beta_4 = \beta_5$.

10. Considere el modelo de regresión múltiple

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon, \tag{2}$$

que verifica los supuestos **1** a **4**. Se quiere contrastar la hipótesis nula $H_0 : \beta_1 - 3\beta_2 = 1$.

a) Sean $\widehat{\beta}_1$ y $\widehat{\beta}_2$ los estimadores MCO respectivos de β_1 y β_2. Exprese $V\left(\widehat{\beta}_1 - 3\widehat{\beta}_2\right)$ en términos de $V\left(\widehat{\beta}_1\right)$, $V\left(\widehat{\beta}_2\right)$ y $C\left(\widehat{\beta}_1, \widehat{\beta}_2\right)$.

b) Escriba el estadístico t para contrastar $H_0 : \beta_1 - 3\beta_2 = 1$.

c) Definiendo $\theta = \beta_1 - 3\beta_2$ y su estimador correspondiente (basado en los estimadores MCO de β_1 y β_2), $\widehat{\theta} = \widehat{\beta}_1 - 3\widehat{\beta}_2$, escriba una especificación equivalente a (2) en la que aparezcan β_0, θ, β_2 y β_3 que permita obtener directamenta a partir de una muestra de datos, $\widehat{\theta}$ y su error estándar.

d) Explique una estrategia, alternativa al apartado (b), de contrastar $H_0 : \beta_1 - 3\beta_2 = 1$.