Hoja de Ejercicios 2
El modelo de regresión lineal simple

1. Para la función de masa de probabilidad conjunta del cuadro siguiente:

$P(Y,X)$	$X=1$	$X=2$	$X=3$
$Y=0$	0.15	0.10	0.15
$Y=1$	0.15	0.30	0.15

   a) Obtenga la función de esperanza condicional $E(Y \mid X)$.
   b) Obtenga el predictor lineal $L(Y \mid X)$.
   c) Presente un cuadro que contenga $E(Y \mid X = x_j)$ y $L(Y \mid X = x_j)$ para $x_j = 1, 2, 3$.
   d) Obtenga las distribuciones marginales de las variables aleatorias $E(Y \mid X)$ y de $E(X \mid Y)$.

2. Considere las distribuciones de probabilidad **condicionales** de Y dado X, $P(Y \mid X = x_j)$, $x_j = 0, 5, 10$.

$P(Y \mid X = x_j)$		X		
		0	**5**	**10**
Y	**5**	1/3	1/2	1/3
	10	1/3	1/2	1/3
	15	1/3	0	1/3

   y la distribución marginal de probabilidad de X, $P(X)$

$P(X)$	X		
	0	**5**	**10**
	3/10	4/10	3/10

   a) Obtenga la función de esperanza condicional $E(Y \mid X)$.
   b) Obtenga el predictor lineal $L(Y \mid X)$.
   c) Presente un cuadro que contenga $E(Y \mid X = x_j)$ y $L(Y \mid X = x_j)$ para $x_j = 0, 5, 10$.
   d) Obtenga las distribuciones marginales de las variables aleatorias $E(Y \mid X)$ y de $E(X \mid Y)$.

3. Suponga que $Y = X + U$, donde

$$E(X) = 100, \qquad E(U) = 0, \qquad V(X) = 600,$$
$$V(U) = 1000, \quad C(X,U) = 400.$$

   Se le indica que una persona tiene $Y = 110$. Prediga, de la mejor forma posible, su valor de X.

4. Suponga que las variables aleatorias X e Y son independientes. Demuestre que el producto esperado $E(XY)$ es igual al producto de las esperanzas, $E(X)E(Y)$. ¿Qué implicación tiene a su vez este resultado sobre $C(X,Y)$?

5. El archivo `DSTAR.GDT` contiene datos de 5743 alumnos que participaron en el experimento STAR realizado en Tennesse durante los 4 años consecutivos, siendo asignados de forma aleatoria desde el primer año a una clase de un determinado tamaño. Las variables incluidas en el archivo son *schidkn*, *classid* (identificadores de alumno y clase, respectivamente) *pscore* (calificación obtenida en una prueba objetiva denominada SAT) *size* (tamaño de la clase) *small* (que toma el valor 1 si la clase es pequeña -entre 13 y 17 alumnos- y 0 si es estándar -entre 22 y 25 alumnos-) *female* (que toma el valor 1 si el alumno es una niña y 0 en caso contrario) *nonwhite* (que toma el valor 0 si el alumno es de raza blanca y 1 en caso contrario). Debe destacarse que por diversas razones, hay clases cuyo tamaño difiere de las calificadas como pequeñas o estándar, y en tal caso la variable *small* no tiene un valor disponible para dichas observaciones.

   a) Utilizando el programa Gretl, realice un análisis de varianza para la calificación *pscore* en función de la variable *small*, que indica si la clase es pequeña o estándar. Para ello, entre en Gretl, abra el archivo `DSTAR.GDT` (File –> Open Data –> User File, y seleccione el directorio y el archivo correspondiente); después, vaya a Model –> Other linear models –> ANOVA y seleccione *pscore* como "response variable" y *small* como "treatment variable".

   b) Contraste que la calificación media es independiente del tamaño de la clase, medida de acuerdo con la variable *small*. (Pista: haga un contraste de la t de igualdad de medias).

   c) Realice una regresión mínimo cuadrática de la calificación *pscore* sobre el tamaño *size*. Para ello, una vez en Gretl y con el archivo `DSTAR.GDT` abierto, vaya a Model –> Ordinary Least Squares, seleccione *pscore* como "response variable" y añada *size* como "treatment variable".

   d) ¿Cuál es su conclusión acerca del efecto del tamaño de la clase en el rendimiento académico?

   e) ¿Cómo cambiaría su conclusión si a lo largo del experimento algunos alumnos hubieran cambiado a clases de diferente tamaño? (por ejemplo, a causa de presiones de los padres cuyos hijos fueron asignados inicialmente a clases estándar)

6. En la formulación alternativa del modelo de regresión clásico, ¿se podría reemplazar el supuesto $E(\varepsilon|X) = 0$ para todo X por el supuesto $E(\varepsilon) = 0$? ¿Son ambos supuestos equivalentes? ¿Es posible que $E(\varepsilon) = 0$ y que $E(\varepsilon|X) = 0$ para todo x? ¿Es posible que $E(\varepsilon|X) = 0$ para todo x sin que $E(\varepsilon) = 0$?

   (Pista: ¿es posible que la renta media en los Estados Unidos sea igual a 20000 dólares sin que la renta media de cada uno de los estados sea también de 20000 dólares? ¿Es posible que la renta media de cada estado sea 20000 dólares y que la renta media en los Estados Unidos no sea también de 20000 dólares?).

7. Suponga que para establecer la relación lineal entre Y = variación porcentual del salario real y X = tasa de desempleo (en tanto por ciento) se considera la siguiente especificación:

$$Y = 8{,}33 - 0{,}84X + u. \text{ donde } E(u|X) = 0$$

   a) Interprete los coeficientes.

b) Suponga ahora que se considera una función recíproca, de manera que consideramos $X' = 1/X$ (inversa de la tasa de desempleo):

$$Y = -0{,}12 + 0{,}983X' + \varepsilon. \text{ donde } E(\varepsilon|X') = 0$$

Interprete los coeficientes.

8. El archivo `ALI.GDT` contiene datos de sección cruzada con 965 observaciones de parejas con o sin hijos en las que la edad del hombre está comprendida entre 25 y 65 años, seleccionadas aleatoriamente de la *Encuesta de Presupuestos Familiares de 1990-91*, elaborada por el Instituto Nacional de Estadística. Contiene las siguientes variables: $V1 = AL =$ gasto anual en alimentación de la familia (en euros), $V2 = GT =$ gasto anual total de la familia (en euros), $V3 = RF =$ renta familiar anual (en euros), $V4 = NH =$ Número de hijos menores de 18 años, $V5 = NA =$ número de adultos (incluídos los cónyuges), $V6 = EH =$ edad del marido, $V7 = EM =$ edad de la mujer, $V8 = UH =$ estudios universitarios del marido (que toma el valor 1 si tiene estudios universitarios y 0 en caso contrario), $V9 = UM =$ estudios universitarios de la mujer (que toma el valor 1 si tiene estudios universitarios y 0 en caso contrario), $V10 = MT =$ situación laboral de la mujer (que toma el valor 1 si trabaja y 0 en caso contrario). Tomemos solamente las siguientes variables:

$$\begin{aligned} V1 =& \text{ Gasto familiar anual (en euros) en alimentación} \\ V2 =& \text{ Gasto familiar total (en euros)} \end{aligned}$$

Hemos estimado con los mismos datos tres modelos alternativos, cuyos coeficientes presentamos a continuación. Compruebe con la ayuda del programa Gretl los resultados obtenidos, realizando las respectivas estimaciones MCO, e interprete los coeficientes que se obtienen en cada uno de dichos modelos alternativos.

a) $\ln(V1) = 3{,}67 + 0{,}48\ln(V2)$;

b) $V1 = -164567 + 2163\ln(V2)$;

c) $(V1/V2) \times 100 = 156{,}89 - 13{,}32\ln(V2)$ (Especificación "Working-Lesser", que considera los determinantes del gasto en alimentación como porcentaje del gasto total).

9. Una empresa multinacional con 1120 establecimientos repartidos por todo el mundo desea analizar los determinantes de las ventas en dichos establecimientos. Para ello, se propone el siguiente modelo:

$$[\ln(V) - \ln(NH)] = \beta_0 + \beta_1 [\ln(R) - \ln(NH)] + \varepsilon,$$

donde el término de error ε cumple los supuestos del modelo de regresión clásico y

$V =$	Ventas anuales (en miles de dólares) del establecimiento,
$R =$	Renta disponible agregada (en miles de dólares) de la localidad en que se halla el establecimiento,
$NH =$	Número de habitantes de la localidad en que se halla el establecimiento.

Interprete β_1.

10. El consumo de energía eléctrica *per capita*, en wiles de kWh, y la renta *per capita*, en miles de euros, para los países de la Unión Europea en el año 2001 están relacionados por medio del modelo lineal
$$C = -0{,}154 + 0{,}571R + \varepsilon.$$
Calcule la elasticidad-renta del consumo *per capita* para una renta *per capita* de 6000 euros.

11. El CAPM (Capital Asset Pricing Model) es un modelo de equilibrio para explicar los rendimientos esperados de los activos. La regresión del exceso de rendimiento (rendimiento por encima del activo seguro) tiene un formulación econométrica muy simple:
$$(R_i - r_i^f) = \beta_1 + \beta_2(R_i^M - r_i^f) + \varepsilon_i,$$
donde, para el mes i-ésimo, R_i es el rendimiento de un activo, r_i^f es el rendimiento mensual de un activo seguro o libre de riesgo (p.ej., las Letras del Tesoro a 30 días), R_i^M es el rendimiento de la cartera de mercado (es decir, el rendimiento de una cartera **ponderada** de **todos** los activos negociados), y ε_i es el término de error, que recoge fluctuaciones aleatorias en el rendimiento del activo analizado que son independientes de la cartera de mercado.

    a) Interprete β_2.

    b) ¿Qué puede decir de un activo para el que $\beta_2 = 1$? ¿Y cuando $\beta_2 > 1$? ¿Y cuando $\beta_2 < 1$?