Estadística II
Ejercicios Tema 5

1. Considera los cuatro conjuntos de datos dados en las transparencias del Tema 5 (sección 5.1)
   (a) Comprueba que los cuatro conjuntos de datos dan lugar a la misma recta de regresión.
   (b) Aplica los métodos de diagnóstico comentados en clase al conjunto de datos # 1, y comenta los resultados.
   (c) Aplica los métodos de diagnóstico comentados en clase al conjunto de datos # 2, y comenta los resultados.
   (d) Aplica los métodos de diagnóstico comentados en clase al conjunto de datos # 3, y comenta los resultados.
   (e) En el conjunto de datos # 3, identifica el dato atípico. Obtén la recta de regresión tras eliminar este dato atípico, y comenta el resultado.

2. A partir de una muestra de 30 observaciones, se estimó el modelo de regresión lineal simple $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$, con $\hat{\beta}_0 = 10.1$ y $\hat{\beta}_1 = 8.4$. La variación cuadrática de la respuesta debida al modelo es $\sum_i (\hat{y}_i - \bar{y})^2 = 128$, mientras que la variación cuadrática residual de la respuesta es $\sum_i e_i^2 = 286$.
   (a) Calcula e interpreta el coeficiente de determinación.
   (b) ¿Qué puedes decir sobre el coeficiente de correlación entre las x_i y las y_i?
   (c) Construye la tabla ANOVA correspondiente a partir de estos datos.
   (d) Contrasta a un nivel de significación del 5% la hipótesis de que la respuesta y no depende de x. Repite el contraste a un nivel de significación del 1%.
   (e) Da un estimador insesgado de la varianza del error.

3. El gestor de un concesionario de automóviles está interesado en la relación entre el número de vendedores que trabajan en fin de semana y el número de coches vendidos. Se obtuvieron los siguientes datos para seis fines de semana consecutivos:

	x_i (**# de vendedores**)	y_i (**# de coches vendidos**)
1	5	22
2	7	20
3	4	15
4	2	9
5	4	17
6	8	25

   (a) Determina la recta de regresión de y (# de coches vendidos) sobre x (# de vendedores).
   (b) Construye la tabla ANOVA y comprueba la validez de la descomposición ANOVA SCT = SCM + SCR.
   (c) Calcula e interpreta el coeficiente de determinación.
   (d) Utiliza la tabla ANOVA para contrastar, al 1% y al 5% de significación, la hipótesis de que el número de vendedores no influye en las ventas.
   (e) Realiza los contrastes del apartado (d) mediante el método visto en el Tema 4. Comprueba que el estadístico T de tal contraste y el estadístico F del contraste del apartado (d) cumplen la relación $F = T^2$.

4. Linealiza las siguientes relaciones no lineales, aplicando las transformaciones vistas en clase:
   (a) $y = \ln(5\sqrt{x})$.
   (b) $y = \frac{2}{3}8^x$.

(c) $y = 1/(4-x)$.

(d) $y = \frac{5}{4}\sqrt{x}$.

5. Supongamos que se han obtenido las siguientes observaciones para una variable respuesta y en función de la variable explicativa x:

x_i	y_i
1	5.47
2	7.54
3	9.13
4	10.47
5	11.65
6	12.72

   (a) Dibuja el gráfico de puntos (x_i, y_i). ¿Parece adecuada una recta para describir la relación entre los datos?

   (b) Suponiendo que el modelo correcto sea de la forma $y = ax^b\, u$, lleva a cabo las transformaciones adecuadas de las variables x e y, y estima los parámetros a y b a partir de una regresión lineal en las variables transformadas.

   (c) Construye la tabla ANOVA para las variables transformadas, y calcula e interpreta el coeficiente de determinación.

6. Para el conjunto de datos # 1 de los considerados en el ejercicio 1, calcula los estimadores de mínimos cuadrados de los coeficientes de regresión lineal empleando la formulación matricial.

7. Un análisis de regresión lineal múltiple a partir de $n = 34$ observaciones proporciona el modelo estimado $\hat{y} = 2.50 + 6.8x_1 + 6.9x_2 - 7.2x_3$. Los errores estándar de los coeficientes estimados de las variables explicativas son $s(\hat{\beta}_1) = 3.1$, $s(\hat{\beta}_2) = 3.7$ y $s(\hat{\beta}_3) = 3.2$. El coeficiente de determinación obtenido es $R^2 = 0.85$.

   (a) Calcula intervalos de confianza al 95% para los coeficientes de las variables explicativas.

   (b) Para cada variable explicativa, contrasta al 5% de significación la hipótesis de que la respuesta no depende de dicha variable.

   (c) Para cada variable explicativa, ¿existe evidencia significativa al 1% de que el coeficiente correspondiente es positivo?

8. Supongamos que has estimado los coeficientes de un modelo de regresión lineal múltiple $y_i = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + u_i$. Contrasta al 5% de significación la hipótesis de que la respuesta no depende de las variables explicativas a partir de las siguientes tablas ANOVA parciales:

(a)

Fuente de variación	SC	G.L.	Media	Cociente F
Modelo	4500	3		
Residuos/Errores	500	26		
Total				

(b)

Fuente de variación	SC	G.L.	Media	Cociente F
Modelo	9780	6		
Residuos/Errores	2100	32		
Total				

(c)

Fuente de variación	SC	G.L.	Media	Cociente F
Modelo	46000	8		
Residuos/Errores	25000	27		
Total				

9. Tenemos los siguientes datos de 10 viviendas unifamiliares, para las que se ha registrado el precio (en M€), la superficie (en m^2), la superficie del terreno (en Has.), y el número de cuartos de baño:

precio (M€)	superficie (m^2)	superf. terreno (Has.)	# baños
170	120.90	0.10	1
177	134.85	0.12	1.5
191	148.80	0.12	2
194	172.05	0.18	2
202	195.30	0.16	2
210	186.00	0.16	2.5
214	195.30	0.20	2
228	223.20	0.20	2.5
240	251.10	0.20	2.5
252	241.80	0.28	3

Más abajo se dan los resultados (obtenidos con Statgraphics) de un análisis de regresión lineal múltiple de y (precio) sobre x_1 (superficie), x_2 (superficie del terreno), y x_3 (# de baños).

(a) Calcula intervalos de confianza al 95% para los coeficientes del modelo de regresión $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + u$.

(b) Contrasta al 5% y al 10% de significación la hipótesis de que la respuesta no depende de la variable x_j, para $j = 1, 2, 3$.

(c) Calcula e interpreta el coeficiente de determinación múltiple. Estima la desviación típica del error.

```
Multiple Regression Analysis
-------------------------------------------------------------------------------
Dependent variable: precio
-------------------------------------------------------------------------------
                                    Standard          T
Parameter               Estimate        Error       Statistic        P-Value
-------------------------------------------------------------------------------
CONSTANT                 100,985      7,86246         12,844          0,0000
superfi                 0,354243    0,0975193         3,63255         0,0109
superfTerreno            109,115      73,4594         1,48537         0,1880
WCs                      10,3945      6,86311         1,51454         0,1807
-------------------------------------------------------------------------------

                           Analysis of Variance
-------------------------------------------------------------------------------
Source              Sum of Squares     Df  Mean Square    F-Ratio      P-Value
-------------------------------------------------------------------------------
Model                     6158,96      3      2052,99      73,92       0,0000
Residual                  166,635      6      27,7726
-------------------------------------------------------------------------------
Total (Corr.)              6325,6      9
```

10. En el modelo de regresión lineal múltiple $y_i = b_0 + b_1 x_{1i} + b_2 x_{2i} + u_i$, donde se supone que se cumplen las hipótesis habituales, se tiene una muestra aleatoria simple de tamaño 4. Los datos, en la forma (y_i, x_{1i}, x_{2i}) son $(y_1, 1, 2)$, $(2, 2, 3)$, $(y_3, 3, 5)$ y $(y_4, 1, 3)$, es decir, no conocemos la variable y para los individuos 1, 3 y 4. Se sabe que los estimadores de mínimos cuadrados de los parámetros del modelo son $\hat{b}_0 = 1.7$, $\hat{b}_1 = -0.3$ y $\hat{b}_2 = 0.1$ y también se sabe que

$$(X'X)^{-1} = \begin{pmatrix} 2.9 & 0.9 & -1.3 \\ 0.9 & 1.9 & -1.3 \\ -1.3 & -1.3 & 1.1 \end{pmatrix}.$$

Se pide:

(a) Calcular los valores desconocidos y_1, y_3, y_4.

(b) Estimar las varianzas de los estimadores de los parámetros $\mathrm{Var}(\hat{b}_i)$, para $i \in \{1, 2, 3\}$.

(c) Construir la tabla de análisis de la varianza y contrastar al 95 por ciento la hipótesis de la validez del modelo, es decir, $H_0 : b_1 = b_2 = 0$.

(d) Dado un modelo de regresión lineal simple $y_i = b_0 + b_1 x_i + u_i$, donde se supone que se cumplen las hipótesis habituales, supongamos que tenemos una muestra aleatoria simple de tamaño n, dada por los pares $(y_1, x_1), \cdots, (y_n, x_n)$. Podemos expresar la varianza de $\hat{b}_1$ de dos formas, la primera es $\text{Var}(\hat{b}_1) = \frac{\sigma^2}{nS_x^2}$, donde $S_x^2 = (1/n)\sum_{i=1}^n (x_i - \bar{x})^2$, y la segunda es, utilizando el modelo en notación matricial, el elemento correspondiente de $\sigma^2(X'X)^{-1}$. Se pide demostrar que las dos formas dan el mismo resultado.

11. La Consejería de Turismo de la Comunidad de Madrid ha realizado un estudio entre poblaciones de menos de 10000 habitantes para estudiar los gastos anuales en promoción turística con respecto al gasto en educación y al gasto en infraestructuras. Se seleccionaron 20 poblaciones en las que se midieron las siguientes variables:

- y = gasto anual en promoción turística (en millones de euros).
- x_1 = gasto anual en educación (en millones de euros).
- x_2 = gasto anual en infraestructuras (en millones de euros).

De dicho estudio se conocen los siguientes datos:

$$(X^T X)^{-1} = \begin{pmatrix} 52.63 & -18.22 & -17.70 \\ & 6.49 & 6.01 \\ & & 6.04 \end{pmatrix}, \quad X^T Y = \begin{pmatrix} 11.34 \\ 13.97 \\ 19.32 \end{pmatrix}.$$

Sabemos también que la variabilidad no explicada toma un valor de

$$\text{SCR} = \sum_{i=1}^{20} (y_i - \hat{y}_i)^2 = 0.034$$

y la variabilidad total de

$$\text{SCT} = \sum_{i=1}^{20} (y_i - \bar{y})^2 = 0.1$$

Se considera el modelo de regresión lineal multiple:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + u_i \quad i = 1 \ldots, 20,$$

para el que se pide que:

(a) Estimes e interpretes los coeficientes del modelo de regresión.

(b) Construyas la tabla ANOVA de la regresión y realices el contraste de significación general para el modelo de regresión.

(c) Realices los contrastes de significación individual de los parámetros del modelo especificando en cada caso las hipótesis nula y altenativa. ¿Qué observas respecto a la significación de los parámetros en comparación al resultado obtenido en el apartado anterior?

(d) Calcules la predicción para el gasto anual en promoción turística de una población que tiene un gasto anual en educación de 1.3 (millones de euros) y un gasto en infraestructuras de 1.2 (millones de euros).

12. Te dan una muestra de 20 observaciones $\{x, z, y\}$ de valores de tres variables, X, Y y Z. Para esta muestra se cumple que

$$\sum_{i=1}^{20} y_i^2 = 10.08, \quad \bar{y} = 0.488$$

Has calculado las estimaciones de mínimos cuadrados de los coeficientes del modelo de regresión lineal múltiple $y = \beta_0 + \beta_1 x + \beta_2 z + u$. Los valores obtenidos son:

$$\hat{\beta}_0 = 0.065, \quad \hat{\beta}_1 = -0.358, \quad \hat{\beta}_2 = 0.104, \quad s(\hat{\beta}_1) = 0.152, \quad s(\hat{\beta}_2) = 0.028, \quad \sum_{i=1}^{20} e_i^2 = 2.878$$

Si aceptamos que se cumplen las hipótesis del modelo de regresión lineal, contesta a las preguntas siguientes:

(a) Completa la tabla ANOVA para este modelo de regresión.

(b) Calcula el coeficiente de determinación múltiple para este model y comenta el significado del mismo.

(c) Contrasta si el modelo de regresión lineal múltiple es globalmente significativo, para un nivel de significación del 1%.

(d) Contrasta si tienes suficiente evidencia para concluir que un incremento en los valores de la variable X implica un decrecimiento en los valores de la variable Y (si se mantiene constante Z), para un nivel de significación del 5%.