Universidad de la República **Modelos Lineales**
Facultad de Ingeniería - IMERL **Primer semestre 2016**

Práctico 1

1. Si $Z \sim \chi^2_d$ hallar $\mathbb{E}(Z)$ y $\mathrm{Var}(Z)$.

2. En un modelo de regresión lineal simple $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$, demostrar las siguientes propiedades:

   a) $\sum_{i=1}^{n} e_i = 0$ siendo $e_i = y_i - \widehat{y}_i$.

   b) $\sum_{i=1}^{n} y_i = \sum_{i=1}^{n} \widehat{y}_i$

   c) $\sum_{i=1}^{n} x_i e_i = 0$

   d) $\sum_{i=1}^{n} \widehat{y}_i e_i = 0$

   e) $\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(\widehat{y}_i - \bar{y})^2 + \sum_{i=1}^{n}(y_i - \widehat{y}_i)^2$ (SCT = SCE + SCR).

   f) Interpretar geométricamente los resultados anteriores en el espacio $\mathbb{R}^n$ en términos de los vectores $y = (y_1, \ldots, y_n)'$, $\mathbb{1} = (1, \ldots, 1)', x = (x_1, \ldots, x_n)', \widehat{y} = (\widehat{y}_1, \ldots, \widehat{y}_n)', e = y - \widehat{y}$.

3. Dado el modelo lineal

$$\begin{pmatrix} y_1 \\ y_2 \end{pmatrix} = \begin{pmatrix} 2 \\ 1 \end{pmatrix} \theta + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \end{pmatrix}$$

   hallar el estimador por el método de los mínimos cuadrados de θ y la suma de cuadrados residuales.

4. En el caso de la regresión lineal simple, hallar los estimadores $\widehat{\beta_1}$ y $\widehat{\beta_0}$ como en un problema de minimización (verificar que efectivamente la solución encontrada es un mínimo).

5. A partir de los datos observados de una variable de respuesta y_i y de una variable regresora x_i se definen nuevas variables estandarizadas:

$$u_i = \frac{x_i - \overline{x}}{s_x} \qquad v_i = \frac{y_i - \overline{y}}{s_y} \quad \forall\, i = 1, \ldots, n$$

   Se define el modelo de regresión estandarizado como

$$v_i = b_1 u_i + \epsilon_i \quad i = 1, \ldots, n$$

   a) Probar que $\bar{u} = \bar{v} = 0$, $s_u^2 = s_v^2 = 1$ y que $\widehat{\beta}_1 = \widehat{b}_1 \frac{s_y}{s_x}$, $\widehat{\beta}_0 = \overline{y} - \widehat{\beta}_1 \overline{x}$ (siendo $\widehat{\beta}_0, \widehat{\beta}_1$ los parámetros estimados del modelo original).

   b) Probar que $\widehat{b}_1 = r$, siendo r el coeficiente de correlación de las variables estandarizadas u, v.

6. Mostrar que la solución matricial $\widehat{\beta} = (\widehat{\beta}_0, \widehat{\beta}_1)' = (X'X)^{-1}X'y$, coincide con

$$\widehat{\beta_1} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \qquad \widehat{\beta}_0 = \bar{y} - \widehat{\beta}_1 \bar{x}.$$

7. Supongamos que las observaciones independientes $Y_1, \ldots, Y_n$ tienen distribución normal con la misma varianza σ^2 y con esperanza $\beta_0 + \beta_1 x_i$ donde β_0 y β_1 son parámetros desconocicos, y $x_1, \ldots, x_n$ son conocidos.

   Hallar por el método de máxima verosimilitud estimadores para β_0, β_1 y σ.

   Deducir que $\widehat{\beta_0}$ y $\widehat{\beta_1}$ tiene distribución normal y hallar sus esperanzas y varianzas.

8. *Test Lack of Fit*

   a) Probar que en el test Lack of Fit, si k es la totalidad de variables independientes $x_1, \ldots, x_k$, para todo $i = 1, \ldots, k$ si n_i es la cantidad de replicas $y_{i1}, \ldots, y_{in_i}$ de la observación x_i, de manera que $n_1 + \cdots + n_k = n$, $\overline{y}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} y_{ij}$, entonces la suma de los errores al cuadrado es:

$$SCR = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij} - \widehat{y}_i)^2 = \underbrace{\sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij} - \overline{y}_i)^2}_{\text{SCR(experimentales)}} + \underbrace{\sum_{i=1}^{k} n_i (\overline{y}_i - \widehat{y}_i)^2}_{\text{SCR(lack of fit)}}$$

   b) Un ingeniero químico desea odelar el rendimiento de un proceso químico (y) en función de la temperatura (x). Al experimentar a distintas temperaturas obtiene los siguientes resultados:

y	x
77.4	
76.7	150
78.2	
84.1	
84.5	200
83.7	
88.9	
89.2	250
89.7	
94.8	
94.7	300
95.9	

   Aplicar el test Lack of Fit a nivel $\alpha = 0{,}05$ para los datos anteriores. En caso que no rechace la hipótesis nula, ajustar un modelo de regresión lineal simple.

   c) Consideramos el *modelo saturado*, es decir el que tiene tanto parámetros como grupo de datos. Esto se hace redefiniendo el predictor como un factor (`factor(x)`). Ajustar un modelo lineal y realizar un test de *anova* con el modelo de la parte (b). Relacionar el estadístico F encontrado con el test Lack of Fit.

   d) Repetir las partes (b) y (c) con los datos de CORROSION del paquete FARAWAY. Observar que el R^2 es de 0.97 pero que el test de Lack of Fit rechaza la hipótesis nula a nivel $0,05$.

Ejercicios en R

1. En este ejercicio usaremos el conjunto de datos `cars` del paquete `base`.

   a) Dibuje el conjunto de puntos en el plano $x - y$ donde $x =$ `speed` e $y =$ `dist`. Calcular, sin usar la función `lm()` de **R**, los coeficientes estimados por mínimos cuadrados del modelo de regresión lineal simple mediante las fórmulas $\widehat{\beta_1} = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) / \sum_{i=1}^{n}(x_i - \bar{x})^2$, $\widehat{\beta}_0 = \bar{y}_n - \widehat{\beta}_1 \bar{x}$.

   b) Verificar el punto anterior mediante la ecuación matricial $\widehat{\beta} = (X'X)^{-1}X'y$.

   c) Explore la posibilidad de obtener un mejor ajuste lineal a los datos transformando las variables.

2. Se considera un vector aleatorio (X, Y) con distribución normal bivariada de media $\mu = (1, 2)'$, $\text{var}(X) = \text{var}(Y) = 1$ y $\text{cov}(X, Y) = 0.8$.

   a) Simular $n = 1000$ observaciones del par (X, Y) simulando primero de la distribución de X y luego de la de $Y|X = x$. Graficar las parejas (x_i, y_i) de valores simulados.

   b) A partir de los valores simulados estimar los parámetros β_0, β_1 del modelo de regresión lineal simple $y_i = \beta_0 + \beta_1 x + \epsilon_i$, por el método de mínimos cuadrados. Graficar conjuntamente las parejas $(x_1, y_1), \ldots, (x_n, y_n)$ y la recta de regresión estimada.

   c) Mostrar que $E(Y|X = x)$ se puede escribir en la forma $b_0 + b_1 x$ y comparar esta recta con la estimada en el punto anterior (agregarla al gráfico).

3. La siguiente tabla presenta cinco conjuntos de datos para cinco modelos de regresión simple diferentes: los datos bajo el encabezamiento x(a-d) son los valores de una variable regresora que es común en las cuatro regresiones con las variables respuesta respectivas y(a), y(b), y(c) y y(d). Las series de datos x(e) y y(e) definen otra regresión. Estimar los cinco modelos de regresión lineal simple, visualizar gráficamente el ajuste de la recta estimada al gráfico de dispersión respectivo y calcular en cada caso el R^2.

obs.	x(a-d)	y(a)	y(b)	y(c)	y(d)	x(e)	y(e)
1	7	5.535	0.103	7.399	3.864	13.715	5.654
2	8	9.942	3.77	8.546	4.942	13.715	7.072
3	9	4.249	7.426	8.468	7.504	13.715	8.496
4	10	8.656	8.792	9.616	8.581	13.715	9.909
5	12	10.737	12.688	10.685	12.221	13.715	9.909
6	13	15.144	12.889	10.607	8.842	13.715	9.909
7	14	13.939	14.253	10.529	9.919	13.715	11.327
8	14	9.45	16.545	11.754	15.86	13.715	11.327
9	15	7.124	15.62	11.676	13.967	13.715	12.746
10	17	13.693	17.206	12.745	19.092	13.715	12.746
11	18	18.1	16.281	13.893	17.198	13.715	12.746
12	19	11.285	17.647	12.59	12.334	13.715	14.164
13	19	21.385	14.211	15.04	19.761	13.715	15.582
14	20	15.692	15.577	13.737	16.382	13.715	15.582
15	21	18.977	14.652	14.884	18.945	13.715	17.001
16	23	17.69	13.947	29.431	12.187	33.281	27.435