Modelo Lineal
PRACTICA 5

1. Sea X una matriz de $\Re^{n\times p}$ tal que $X = [X_1, X_2]$, donde $X_1 \in \Re^{n\times k}$ y $X_2 \in R^{n\times(p-k)}$. Sean $P_1 = X_1(X_1'X_1)^{-1}X_1'$ la matriz de proyección generada por las columnas de X_1 y sea $W = (I - P_1)X_2$ la proyección de X_2 sobre el complemento ortogonal de X_1. Finalmente, sea $P_2 = W(W'W)^{-1}W'$ la matriz de proyección correspondiente a W. Probar que

$$P = P_1 + P_2 = P_1 + (I - P_1)X_2(X_2'(I - P_1)X_2)^{-1}X_2'(I - P_1).$$

2. Supongamos que $X \in \Re^{n\times p}$ contiene una columna de $1's.$, por ejemplo, sin pérdida de generalidad, la primera, es decir $X = [\mathbf{l}, X_2]$ y sea $P = X(X'X)^{-1}X'$. Pruebe que

a) $P = P_1 + P_2$ donde $P_1 = n^{-1}\mathbf{l}\,\mathbf{l}'$ $(\mathbf{l} \in \Re^n, \mathbf{l} = (1, 1, .., 1)')$ y $P_2 = \widetilde{X}(\widetilde{X}'\widetilde{X})^{-1}\widetilde{X}'$ siendo $\widetilde{X} = (I - n^{-1}\mathbf{l}\,\mathbf{l}')X_2$ la matriz con las columnas centradas.

b) $p_{ii} \geq \dfrac{1}{n}$.

c) $p_{ii} = \dfrac{1}{n} + \widetilde{p}_{ii}$ donde $\widetilde{p}_{ii} = (P_2)_{ii}$.

3. Pruebe que la distancia de Cook de la i-ésima observación puede calcularse como

$$D_i = \frac{1}{p}\frac{p_{ii}}{1 - p_{ii}}r_i^2$$

4. Pruebe que $\widehat{Y}_i = (1 - p_{ii})\,\mathbf{x}_i'\widehat{\beta}_{(i)} + p_{ii}Y_i$. Huber (1981) interpretó de esta igualdad que $\dfrac{p_{ii}}{1 - p_{ii}}$ es el cociente entre la parte de $\widehat{Y}_i$ que se explica por Y_i y la parte de $\widehat{Y}_i$ que puede predecirse a partir de $\mathbf{x}_i'\widehat{\beta}_{(i)}$.

5. Los datos que se muestran en la Tabla 1 (y en el archivo salud.txt) corresponden a 30 miembros de un club de salud. Las variables son

$X_1 =$ peso en libras

$X_2 =$ pulso en reposo

$X_3 =$ fuerza del brazo y pierna (número de libras que puede levantar)

$X_4 =$ tiempo en segundos en que corre 1/4 de milla

$Y =$ tiempo en segundos en que corre 1 milla.

a) Estime por cuadrados mínimos los parámetros del modelo

$$E(Y) = \beta_o + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4$$

Realice el gráfico de los residuos estandarizados versus $\hat{Y}$ y el QQ–plot . ¿Hay algún indicio de que haya groseras violaciones de supuestos ?

b) Realice un boxplot y un esquema de tallo y hoja para los residuos estandarizados. ¿A qué número de observación corresponde el residuo de mayor valor absoluto?

c) Calcule el leverage de cada observación y realice un boxplot. ¿Cuál es la observación con mayor leverage? Use los criterios dados para identificar aquellas observaciones cuyo leverage puede indicar problemas.

d) Calcule la distancia de Cook para cada observación y realice un boxplot. ¿Cuáles son los cuatro puntos que aparecen como outliers en este gráfico? ¿Cuál es la observación de mayor influencia? ¿Qué pasa con el leverage de esta observación?

e) Según la tabla de estimación de los coeficientes, cuáles serían los coeficientes significativos? Realice los plots de residuos parciales (component plus residual plot) para verificar estos resultados. ¿A quién corresponde el punto más alejado en el grafico correspondiente a X_3 ?

f) *Plot L–R.* El L–R plot permite graficar residuos vs. leverage de manera de identificar qué tipo de patología tiene un dato. Se calculan los residuos como:

$$a_i = \frac{e_i}{\sqrt{(e^t e)}}$$

y se grafica a_i^2 vs p_{ii} .

Realice el L–R plot correspondiente a este problema. ¿Cómo caracterizaría a los puntos correspondientes a las observaciones 23, 28 y 30?

g) Recalcule los estimadores de mínimos cuadrados omitiendo una a la vez las observaciones 23, 28 y 30. ¿Cuándo observa el mayor cambio en los estadísticos t?

6. La Tabla 2 (y el archivo salario.txt) corresponde a un estudio para relacionar el salario mensual de una muestra aleatoria de 31 empleados y un conjunto de factores que se piensa pueden determinar las diferencias en los salarios. Las variables observadas son:

$X_1 =$ evaluación de trabajo

$X_2 =$ sexo (1=hombre, 0=mujer)

$X_3 =$ número de años en la compañía

$X_4 =$ número de años en el mismo cargo

$X_5 =$ ranking de rendimiento (1=no satisfactorio, 5=muy bueno)

$Y =$ salario mensual.

a) Estime por cuadrados mínimos los parámetros del modelo

$$E(Y) = \beta_o + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5$$

Realice el gráfico de los residuos estandarizados versus $\hat{y}$ y el QQ–plot . ¿Hay algún residuo que le llame la atención ?

b) Realice un boxplot y un esquema de tallo y hoja para los residuos estandarizados. ¿A qué observaciones corresponden los residuos mayores?

c) Para cada variable independiente realice un gráfico de los residuos estandarizados versus X_i. Observe qué ocurre con la observación número 6 en cada uno de estos gráficos.

d) Calcule el leverage de cada observación y realice un boxplot. ¿Qué observación se destaca?

e) Realice el L–R plot para este conjunto de datos. ¿A qué conclusiones llega?

7. La Tabla 3 (gener2.txt) contiene un conjunto de 30 datos generados correspondientes a dos variables: X e Y.

a) Ajuste un modelo lineal utilizando la variable Y como respuesta y la variable X como explicativa. Realice el QQ–plot de los residuos. En base a este gráfico, ¿cree que es válida la hipótesis de normalidad de los residuos? Aplique el test de Kolmogorov–Smirnov a los residuos. ¿Cuál es su conclusión?

b) Seleccione una transformación adecuada en la familia de transformaciones de Box–Cox y repita el análisis realizado en a) con la variable transformada como respuesta.

8. Considere el modelo $Y = X\beta + \epsilon$, donde $\epsilon \sim N_n(0, \sigma^2 I_n)$ y X es una matriz $n \times p$ de rango p. Sean $\lambda_1, \lambda_2, \ldots, \lambda_p$ los autovalores de la matriz $X'X$ y llamemos $\widehat{\beta}$ al estimador de mínimos cuadrados de β.

a) Muestre que $\sum_{j=1}^{p} Var(\widehat{\beta}_j) = \sigma^2 \sum_{j=1}^{p} \lambda_j^{-1}$. ¿Cuál es el efecto de pequeños autovalores sobre la varianza de los coeficientes?

b) Si $tr(X'X) = c$, donde c es una constante dada, pruebe que $\sum_{j=1}^{p} Var(\widehat{\beta}_j)$ se minimiza si $X'X = c/pI_p$.

c) Muestre que $E(\widehat{\beta}'\widehat{\beta}) = \beta'\beta + \sigma^2 \sum_{j=1}^{p} \lambda_j^{-1}$. ¿Cuál es el efecto de pequeños autovalores sobre la tendencia de $\widehat{\beta}'\widehat{\beta}$ a sobreestimar $\beta'\beta$?

d) Suponga que podría incorporar m observaciones adicionales que provienen del modelo supuesto y que estas observaciones están contenidas en la matriz X^* de dimensión $m \times p$ con $m \geq p$ y además que $X^{*\prime}X^* = dI_p$. Pruebe que después de incorporar las m observaciones adicionales $\sum_{j=1}^{p} Var(\widehat{\beta}_j) = \sigma^2 \sum_{j=1}^{p} (d + \lambda_j)^{-1}$. ¿Cuándo sería conveniente incorporar esta información adicional para reducir $\sum_{j=1}^{p} Var(\widehat{\beta}_j)$?

9. Sean $R^2_{a,k}$ y $R^2_{a,p}$ los R^2 ajustados de los modelos (1) y (2) respectivamente

$$E(Y) = 1 + \beta_1 x_1 + \cdots + \beta_{k-1} x_{k-1} \qquad (1)$$

$$E(Y) = 1 + \beta_1 x_1 + \cdots + \beta_{p-1} x_{p-1} \qquad (2)$$

donde $p < k$.

a) Pruebe que $1 + \dfrac{C_p - p}{n - p} = \dfrac{1 - R^2_{a,p}}{1 - R^2_{a,k}}$, con lo cual $R^2_{a,p} > R^2_{a,p'}$ no es equivalente a $C_p < C_{p'}$.

b) Pruebe que $C_k = k$.

c) Sea F el estadístico del test F para testear la hipótesis de de que el modelo (2) es válido frente a la alternativa de que el modelo (1) es cierto. Pruebe que

$$F = \frac{1}{k - p}(C_p - 2p + k).$$

10. La Tabla 4 (archivo llamadas.txt) muestra los datos correspondientes al número total de llamadas telefónicas internacionales (en decenas de millón) entre 1950 y 1973 registradas en el Belgian Statistical Survey publicada por el Ministerio de Economía de Bélgica.

a) Ajuste por mínimos cuadrados un modelo lineal simple usando como variable explicativa a *year* y como respuesta a *calls*. Calcule los residuos estandarizados y studentizados, el leverage, la distancia de Cook y los Dfits para el modelo ajustado. Calcule los puntos de corte para cada medida y realice los gráficos que le parezcan apropiados con estas medidas para identificar posibles outliers y/o puntos influyentes.

b) Realice un scatterplot de *year* vs. *calls* en el que se grafique además la recta obtenida por mínimos cuadrados. ¿Qué le parece el ajuste obtenido? ¿Le parece que lo que se observa en el gráfico es bien reflejado por las medidas calculadas en a)? Si la respuesta es negativa, ¿cuál puede ser el motivo?

11. Los datos de la Tabla 5 (stars.txt) forman el diagrama de Hertzsprung–Russell del cluster de estrellas CYG OB1, que contiene 47 estrellas en la dirección de Cygnus. *Ltemp* es el logaritmo de la temperatura en la superficie de la estrella y L es el logaritmo de intensidad lumínica.

a) Ajuste por mínimos cuadrados un modelo lineal simple usando como variable explicativa a *Ltemp* y como respuesta a L. Calcule los residuos estandarizados y studentizados, el leverage, la distancia de Cook y los Dfits para el modelo ajustado. Calcule los puntos de corte para cada medida y realice los gráficos que le parezcan apropiados con estas medidas para identificar posibles outliers y/o puntos influyentes.

b) Realice un scatterplot de *Ltemp* vs. L en el que se grafique además la recta obtenida por mínimos cuadrados. ¿Qué conclusiones obtiene? ¿Le parece compatible este resultado con los del item a) ?

12. En la Tabla 6 (stack.txt) se presentan los datos conocidos como *Stackloss Data* ampliamente tratados en la literatura. Los 21 datos corresponden a una muestra real y describen

la operación de una planta de oxidación de amoníaco a ácido nítrico. La variable de respuesta Y es el stackloss, $X1$ es la tasa de operación, $X2$ es la temperatura del agua de enfriado y $X3$ es la concentración de ácido.

Resumiendo el análisis realizado por diversos autores, puede decirse que las observaciones 1, 3, 4 y 21 fueron clasificadas como outliers, mientras que la observación 2 también fue calificada como outlier moderado por algunos de estos autores.

Realice un ajuste por mínimos cuadrados, realice un gráfico de residuos estandarizados vs. número de observación y verifique si el valor absoluto de algún residuo estandarizado es mayor que la cota 2.5.

13. Utilizando los datos del ejemplo de Biomasa (Tabla 6 de la Práctica 3 y archivo biomasa.txt),

a) Calcule para todos 31 subconjuntos basados en las 5 variables independientes: K, $SODIO$, PH, SAL y ZN, el C_p de Mallows, el R^2 y el R^2 ajustado. Elija los subconjuntos que serían los mejores candidatos según cada criterio. Compare las elecciones realizadas por los distintos métodos. Realice un gráfico de p vs. C_p y observe cuáles son los subconjuntos con C_p pequeño y cercanos a p.

b) Use el procedimiento Stepwise con la opción "forward" para elegir el mejor modelo. Describa el test F que se realiza en cada paso (es decir cuál es la hipótesis nula y cuál es el modelo en cada paso). Verifique su respuesta realizando el test adecuado.

c) Use el procedimiento Stepwise con la opción "backward" para elegir el mejor modelo. Describa el test F que se realiza en cada paso (es decir cuál es la hipótesis nula y cuál es el modelo en cada paso).

d) Use el procedimiento Stepwise con la opción "Efroymson" para elegir el mejor modelo. Describa el test F que se realiza en cada paso (es decir cuál es la hipótesis nula y cuál es el modelo en cada paso).

Compare la elección elegida automáticamente con las resultantes del ítem a).

14. Para los datos del ejercicio 18 de la Práctica 2 (peak.txt) transformados con logaritmo,

a) calcule, con la función leaps, para los mejores subconjuntos que selecciona por default basados en las 9 variables independientes, el C_p de Mallows, el R^2 y el R^2 ajustado. ¿Cuáles son los mejores candidatos de acuerdo a cada criterio ? ¿Son muy diferentes sus elecciones?

b) Use el procedimiento Stepwise con la opción Efroymson para elegir el mejor modelo automáticamente. Compare este modelo con los obtenidos en el ítem a). ¿Cómo se pueden interpretar las diferencias observadas?

15. Los datos de la Tabla 7 (webster.txt) fueron generados por Webster, Gunst y Mason (1974). Los generaron de manera tal que $\sum_{j=1}^{4} x_{ij} = 10$ para las observaciones 2 a 12 y $\sum_{j=1}^{4} x_{ij} = 11$ para la observación 1. Las variables x_5 y x_6 fueron generadas con distribución normal, mientras que la respuesta Y satisface el modelo:

$$Y = 10 + 2x_1 + x_2 + 0.2x_3 - 2x_4 + 3x_5 + 10x_6 + \epsilon$$

donde $\epsilon \sim N(0,1)$.

a) Calcule la matriz de correlación de las variables independientes $x_i, i = 1,6$ ¿Le parece que están altamente correlacionadas?

b) Calcule el estimador de mínimos cuadrados y observe cuáles son los estimadores de los parámetros que son significativos con nivel 0.05.

c) Calcule los Factores de Inflación de la Varianza (VIF) para este ejemplo. ¿Qué le sugieren?

d) Calcule los índices de condición para X_s, la matriz de diseño escalada (es decir, luego de dividir a cada columna de la matriz de diseño por su norma). ¿Qué le sugieren ?

Tabla 1. Club de Salud

obs	X_1	X_2	X_3	X_4	Y
1	217	67	260	91	481
2	141	52	190	66	292
3	152	58	203	68	338
4	153	56	183	70	357
5	180	66	170	77	396
6	193	71	178	82	429
7	162	65	160	74	345
8	180	80	170	84	469
9	205	77	188	83	425
10	168	74	170	79	358
11	232	65	220	72	393
12	146	68	158	68	346
13	173	51	243	56	279
14	155	64	198	59	311
15	212	66	220	77	401
16	138	70	180	62	267
17	147	54	150	75	404
18	197	76	228	88	442
19	165	59	188	70	368
20	125	58	160	66	295
21	161	52	190	69	391
22	132	62	163	59	264
23	257	64	313	96	487
24	236	72	225	84	481
25	149	57	173	68	374
26	161	57	173	65	309
27	198	59	220	62	367
28	245	70	218	69	469
29	141	63	193	60	252
30	177	53	183	75	338

Tabla 2. Salarios

obs	X_1	X_2	X_3	X_4	X_5	Y
1	350	1	2	2	5	1000
2	350	1	5	5	5	1400
3	350	0	4	4	4	1200
4	350	1	20	20	1	1800
5	425	0	10	2	3	2800
6	425	1	15	10	3	4000
7	425	0	1	1	4	2500
8	425	1	5	5	4	3000
9	600	1	10	5	2	3500
10	600	0	8	8	3	2800
11	600	0	4	3	4	2900
12	600	1	20	10	2	3800
13	600	1	7	7	5	4200
14	700	1	8	8	1	4600
15	700	0	25	15	5	5000
16	700	1	19	16	4	4600
17	700	0	20	14	5	4700
18	400	0	6	4	3	1800
19	400	1	20	8	3	3400
20	400	0	5	3	5	2000
21	500	1	22	12	3	3200
22	500	1	25	10	3	3200
23	500	0	8	3	4	2800
24	500	0	2	1	5	2400
25	800	1	10	10	3	5200
26	475	1	10	4	3	2400
27	475	0	3	3	4	2400
28	475	1	8	8	2	3000
29	475	1	6	6	4	2800
30	475	0	12	4	3	2500
31	475	0	4	2	5	2100

Tabla 3. Datos generados

obs	X	Y	obs	X	Y
1	0.2775889	696.20174	16	−0.886568	6.8234601
2	−0.517632	72.295359	17	0.4935309	72.466261
3	−0.303792	113.55840	18	0.7128258	2621.0541
4	−2.078444	1.8890014	19	1.0458933	293.29129
5	1.4366185	650.52223	20	−1.559868	5.5906742
6	−0.235768	96.996475	21	0.9805923	339.38955
7	−0.868704	35.627923	22	−0.597517	39.637404
8	1.8890238	31566.051	23	−1.448257	36.087594
9	−1.743802	3.9250497	24	0.443873	446.68841
10	−0.305831	445.65564	25	0.1877068	520.12349
11	0.2756379	290.04645	26	−0.295656	62.706673
12	1.0474655	402.32246	27	−0.284534	116.83186
13	1.3234957	930.59459	28	1.140876	3917.2343
14	−0.237969	56.925312	29	0.1878306	179.25413
15	0.7173053	800.51180	30	−0.153660	551.13103

Tabla 4. Llamadas

obs	*year*	*calls*
1	50	4.4
2	51	4.7
3	52	4.7
4	53	5.9
5	54	6.6
6	55	7.3
7	56	8.1
8	57	8.8
9	58	10.6
10	59	12.0
11	60	13.5
12	61	14.9
13	62	16.1
14	63	21.2
15	64	119.0
16	65	124.0
17	66	142.0
18	67	159.0
19	68	182.0
20	69	212.0
21	70	43.0
22	71	24.0
23	72	27.0
24	73	29.0

Tabla 5. Stars

obs	*Ltemp*	*L*	obs	*Ltemp*	*L*
1	4.37	5.23	24	4.49	4.85
2	4.56	5.74	25	4.38	5.02
3	4.26	4.93	26	4.42	4.66
4	4.56	5.74	27	4.29	4.66
5	4.30	5.19	28	4.38	4.90
6	4.46	5.46	29	4.22	4.39
7	3.84	4.65	30	3.48	6.05
8	4.57	5.27	31	4.38	4.42
9	4.26	5.57	32	4.56	5.10
10	4.37	5.12	33	4.45	5.22
11	3.49	5.73	34	3.49	6.29
12	4.43	5.45	35	4.23	4.34
13	4.48	5.42	36	4.62	5.62
14	4.01	4.05	37	4.53	5.10
15	4.29	4.26	38	4.45	5.22
16	4.42	4.58	39	4.53	5.18
17	4.23	3.94	40	4.43	5.57
18	4.42	4.18	41	4.38	4.62
19	4.23	4.18	42	4.45	5.06
20	3.49	5.89	43	4.50	5.34
21	4.29	4.38	44	4.45	5.34
22	4.29	4.22	45	4.55	5.54
23	4.42	4.42	46	4.45	4.98
			47	4.42	4.50

Tabla 6. Datos de Stackloss

obs	$X1$	$X2$	$X3$	Y
1	80	27	89	42
2	80	27	88	37
3	75	25	90	37
4	62	24	87	28
5	62	22	87	18
6	62	23	87	18
7	62	24	93	19
8	62	24	93	20
9	58	23	87	15
10	58	18	80	14
11	58	18	89	14
12	58	17	88	13
13	58	18	82	11
14	58	19	93	12
15	50	18	89	8
16	50	18	86	7
17	50	19	72	8
18	50	19	79	8
19	50	20	80	9
20	56	20	82	15
21	70	20	91	15

Tabla 7. Datos de Webster et al.

obs	$X1$	$X2$	$X3$	$X4$	$X5$	$X6$	Y
1	8	1	1	1	0.541	−0.099	10.006
2	8	1	1	0	0.13	0.07	9.737
3	8	1	1	0	2.116	0.115	15.087
4	0	0	9	1	−2.397	0.252	8.422
5	0	0	9	1	−0.046	0.017	8.625
6	0	0	9	1	0.365	1.504	16.289
7	2	7	0	1	1.996	−0.865	5.958
8	2	7	0	1	0.228	−0.055	9.313
9	2	7	0	1	1.38	0.502	12.96
10	0	0	0	10	−0.798	−0.399	5.541
11	0	0	0	10	0.257	0.101	8.756
12	0	0	0	10	0.44	0.432	10.937