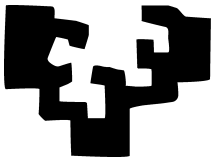

Universidad del País Vasco
Euskal Herriko Unibertsitatea

TAREA 7

EJERCICIOS

1. Los datos correspondientes a este ejercicio están en un fichero de nombre `mortal.dat`. Recogen la tasa de mortalidad en diversas comunidades de Estados Unidos, junto con variables recogiendo información sobre la polución atmosférica y otras presumiblemente relacionadas con la calidad de vida. Las variables —en orden de aparición en cada fila— se describen en el Cuadro 1. La mortalidad `MORT` se proporciona "ajustada por edad", sin que se indique cómo.

   a) Escoge un modelo que ajuste los datos. Puedes servirte de criterios de ajuste como los vistos en clase (C_p de Mallows, etc.).

   b) Interpreta los resultados. ¿Encuentras evidencia de que la polución está relacionada con la tasa de mortalidad? Resume en unas pocas líneas tus hallazgos.

   c) Calcula —y representa gráficamente— los residuos borrados. ¿Hay alguna observación cuyo comportamiento se separe notablemente del de las restantes?

   d) Obtén las curvas de influencia de cada observación sobre cada uno de los parámetros del modelo que hayas seleccionado.

2. Los datos correspondientes a este ejercicio están en un fichero llamado `longley.dat`, en siete columnas. Se reproducen en el Cuadro 2. La primera columna, `GNP.deflator`, es el regresando. Son datos de la economía U.S.A. entre 1947 y 1962, y se utilizan frecuentemente como banco de pruebas cuando se requiere un conjunto de regresores acusadamente multicolineal; casi cualquier conjunto de series macroeconómicas no despojadas de sus tendencias exhibiría análogo comportamiento.

   a) Ajusta una regresión lineal de la primera columna sobre las restantes, utilizando los procedimientos estudiados en clase para hacer frente a multicolinealidad fuerte (ridge regression y regresión en componentes principales).

   b) Dibuja la traza ridge de algunos de los parámetros. Compara los diferentes estimadores entre sí y con los MCO.

   c) (OPTATIVO si trabajas sobre S-PLUS) Haz estimación ridge seleccionando el parámetro k por validación cruzada. En R, la función `lm.ridge` (del paquete MASS; mira [9] o la documentación on-line) hace todo el trabajo por tí. En S-PLUS habrías de escribir una pequeña función.

Cuadro 1: Datos en el fichero `mortal.dat`

PREC	Precipitación anual (en pulgadas; multiplicado por 25.6 = litros/m^2).
JANT	Temperatura promedio en Enero (grados Farenheit).
JULT	Temperatura promedio en Julio (grados Farenheit).
OVR65	Porcentaje de población de 65 ó más años.
POPN	Tamaño medio de la familia.
EDUC	Mediana de años de escolarización de personas mayores de 22 años.
HOUS	Porcentaje de viviendas no en ruinas y con comodidades mínimas.
DENS	Población por milla cuadrada en áreas urbanas (en 1.960).
NONW	Porcentaje de población no blanca en áreas urbanas en 1960.
WWDRK	Porcentaje de empleados en ocupaciones de "cuello blanco".
POOR	Porcentaje de familias con renta de menos de $3000.
HC	Potencial relativo de polución por hidrocarburos.
NOX	Idem óxidos nítricos.
SO2	Idem SO_2.
HUMID	Promedio anual de humedad relativa a las 13 horas.
MORT	Mortalidad total ajustada por edad por 100,000 habitantes.

Cuadro 2: Datos en el fichero `longley.dat`

	GNP.deflator	GNP	Unemployed	Armed.Forces	Population	Year	Employed
1947	83.00	234.29	235.60	159.00	107.61	1947.00	60.32
1948	88.50	259.43	232.50	145.60	108.63	1948.00	61.12
1949	88.20	258.05	368.20	161.60	109.77	1949.00	60.17
1950	89.50	284.60	335.10	165.00	110.93	1950.00	61.19
1951	96.20	328.98	209.90	309.9	112.08	1951.00	63.22
1952	98.10	347.00	193.20	359.40	113.27	1952.00	63.64
1953	99.00	365.38	187.00	354.70	115.09	1953.00	64.99
1954	100.00	363.11	357.80	335.00	116.22	1954.00	63.76
1955	101.20	397.47	290.4	304.80	117.39	1955.00	66.02
1956	104.60	419.18	282.20	285.70	118.73	1956.00	67.86
1957	108.40	442.77	293.60	279.80	120.44	1957.00	68.17
1958	110.80	444.55	468.10	263.70	121.95	1958.00	66.51
1959	112.60	482.70	381.30	255.20	123.37	1959.00	68.66
1960	114.20	502.60	393.10	251.40	125.37	1960.00	69.56
1961	115.70	518.17	480.60	257.20	127.85	1961.00	69.33
1962	116.90	554.89	400.70	282.70	130.08	1962.00	70.55

Cuadro 3: Equivalencia de funciones en S-PLUS y R.

En S-PLUS	En R	Paquete:
`leaps`	`leaps`	`leaps`
-	`regsubsets`	`leaps`
`ls.diag`	`ls.diag`	`base`
`lsfit`	`lsfit`	`base`
`lm`	`lm`	`base`
`drop1`	`drop1`	`base`
`add1`	`add1`	`base`
`step`	`step`	`base`
-	`stepAIC`	`MASS`
`stepwise`	-	-
`lm.influence`	`lm.influence`	`base`
-	`lm.ridge`	`MASS`

AYUDAS, SUGERENCIAS, COMENTARIOS

1. En la Tabla 3 tienes una breve relación de funciones disponibles en S-PLUS y R. Funciones del mismo nombre hacen lo mismo con leves matices: consulta en cada caso la documentación del paquete que utilices. Observa que algunas funciones en R forman parte de paquetes añadidos que has de cargar con un `library(MASS)`, etc.

2. Puedes emplear `read.table` para leer los dos ficheros necesarios. En el caso de ambos, l a primera línea da los nombres de las variables, por lo que debes emplear la opción `header=T`.

3. Tanto las funciones `lm` como `lsfit` son utilizables. Si lees los datos como data-frames, puedes emplear `lm`. Cuando hayas seleccionado un modelo —para lo que la sintaxis y facilidades de `lm` resultan generalmente más cómodas—, puedes invocar dicha función con las opciones `x=T` y `y=T`. Esto tiene por efecto devolver un objeto con componentes `$x` e `$y` conteniendo respectivamente la matriz de regresores y el vector de observaciones del regresando. Con las matriz `x` y vector `y` así obtenidos, puedes invocar a continuación `lsfit` y cualquiera de las funciones (como `ls.diag`) que requieren como argumento un objeto como los que proporciona `lsfit`.

   Observa que si hay regresores cualitativos —que deben ser desdoblados en columnas de unos y ceros—, `lm` hace todo el trabajo por tí. Incluso elimina una columna redundante para evitar colinealidad con la columna de "unos", si la hay.

4. El argumento `keep=T` en la función `drop1` especifica que queremos guardar toda la información generada en el ajuste de cada uno de los modelos que resulta al dejar de lado un regresor por turno. Por ejemplo, en

```
> ajuste5 <- lm(Edad ~ Partido + poly(Eleccion,3))
> drop1(ajuste5,keep=T)
$anova:
Single term deletions

Model:
Edad ~ Partido + poly(Eleccion, 3)
                    Df  Sum of Sq        RSS         Cp
          <none> "  " "        " "1517.552" "1916.908"
         Partido " 1" " 13.8648" "1531.417" "1850.902"
poly(Eleccion, 3) " 3" "467.0518" "1984.604" "2144.347"
```

```
$keep:
                  coefficients      fitted   residuals       x.residuals
          Partido numeric, 4   numeric, 43 numeric, 43 numeric, 43
poly(Eleccion, 3) numeric, 2   numeric, 43 numeric, 43 model.matrix, 129
                      effects          R
          Partido numeric, 43 matrix, 16
poly(Eleccion, 3) numeric, 43 matrix, 4
```

se guardan los coeficientes estimados, residuos, valores ajustados, etc. En el ejemplo presentado, `ajuste5$keep` es una lista. Los distintos componentes están ordenados por "columnas". El primer componente en el ejemplo anterior serían los cuatro coeficientes al ajustar el modelo sin Partido entre los regresores; se obtendría como `ajuste5$keep[[1]]`. Para obtener los residuos ordinarios en la misma regresión, bastaría tomar `ajuste5$keep[[5]]`.

5. Además de `leaps` (empleable en conjunción con `lsfit`) y `add1` y `drop1` (junto con `lm`) dispones de la función `step` que hace regresión escalonada o *stepwise* al modo convencional en otros programas y `regsubsets` (ésta última en `leaps`). Son útiles, pero —al igual que en el caso de `leaps`— no debes aceptar acríticamente lo mejor que te presente. Puede que seas capaz de mejorarlo.

6. Tienes ayuda *on line* sobre todas las funciones. También dispones del manual y de [3]. Para una descripción del output de `drop1` puedes también querer hacer `help(lm.object)`.

7. Sobre regresión *ridge*: es habitual la pregunta de qué valores probar para k, parámetro de "engordado" de la diagonal principal de $(X'X)$. *Si las X están reescaladas de modo que $(X'X)$ es una matriz de correlación* (tiene unos a lo largo de la diagonal principal) valores de unas pocas centésimas —quizá hasta 0.10— suelen ser lo adecuado. Si se opera con las variables X sin reescalar, los k adecuados serán proporcionalmente mayores (o menores).

   Nota que cuando las escalas de los regresores son muy diferentes, hacer estimación *ridge* sin corregir este efecto es inadecuado.

   Si empleas la función `lm.ridge` de la biblioteca MASS (disponible sobre R) no te has de preocupar de las escalas de las variables. La función reescala los regresores hasta que $(X'X)$ es una matriz de correlación y luego deshace el cambio. Los k's que has de proporcionar son los que corresponderían a una matriz $(X'X)$ de correlación.

8. Si empleas la función `step` para hacer regresión escalonada, verás que se emplea como criterio para la selección de modelos el AIC (An Information Criterion, o Akaike's Information Criterion, por el nombre de su proponente; véase [1]). Se define como

$$AIC = -2\log_e(\text{Verosimilitud maximizada}) + 2p, \tag{1}$$

   siendo p el número de parámetros ajustados. Si consideras modelos de regresión lineal con perturbaciones normales, al tomar el logaritmo neperiano de la verosimilitud comprobarás que obtienes algo parecido a $-\frac{1}{2}SSE/\hat{\sigma}^2$. En la práctica, es casi igual emplear C_p o AIC (mira por ejemplo [9], p. 185, para ver cuál es exactamente la relación entre ambos estadísticos).

9. Errores comunes en el pasado que has de evitar:

   a) Los objetos devueltos por diferentes funciones no necesariamente lo son en formatos compatibles. Un error frecuente en el pasado ha sido calcular la curva de influencia empírica así:

$$\texttt{SIC}_i = (N-1)(\hat{\beta} - \hat{\beta}_i), \tag{2}$$

   en que $\hat{\beta}$ era el vector devuelto por `lsfit` y $\hat{\beta}_i$ el proporcionado por `lm.influence`. Comprueba que las dimensiones de las cosas que restas no son iguales. La forma de operar de S-PLUS —hacer conformables las cosas aunque sea a martillazos—es muy cómoda en muchas ocasiones, pero aquí resulta insidiosa: obtendrás sin ningún aviso un resultado incorrecto.

   Has de preocuparte de que las cosas que restas sean realmente "restables".

b) Cuando se hace estimación en componentes principales, hay que restaurar los factores de escala antes de comparar con MCO.

c) Es común hacer la estimación, observar el t-ratio de `SO2` y contrastar a continuación la hipótesis de que el parámetro correspondiente es nulo. ¿Ves en qué se viola aquí el principio de que las hipótesis *no* deben ser escogidas a la vista de los resultados experimentales (y, si se hace, hay que tener en cuenta este hecho empleando métodos de inferencia simultánea)?

Referencias

[1] H. Akaike. Use of an information theoretic quantity for statistical model identification. In *Proc. 5th. Hawai Int. Conf. on System Sciences*, pages 249–250, 1972.

[2] R.A. Becker, J.M. Chambers, and A.R. Wilks. *The New S Language. A Programming Environment for Data Analysis and Graphics*. Wadsworth & Brooks/Cole, Pacific Grove, California, 1988.

[3] J.M. Chambers and T.J. Hastie. *Statistical Models in S*. Wadsworth & Brooks/Cole, Pacific Grove, Ca., 1992.

[4] A.E. Hoerl and R.W. Kennard. Ridge regression: Biased estimation for non-orthogonal problems. *Technometrics*, 12:55–67, 1970.

[5] J.W. Longley. An appraisal of least squares programs for the electronic computer from the point of view of the user. *Journal of the American Statistical Association*, 62:819–841, 1967.

[6] R.H. Myers. *Classical and Modern Regression with Applications*. PWS-KENT Pub. Co., Boston, 1990.

[7] G.A.F. Seber. *Linear Regression Analysis*. Wiley, New York, 1977.

[8] A. Fdez. Trocóniz. *Modelos Lineales*. Serv. Editorial UPV/EHU, Bilbao, 1987.

[9] W.N. Venables and B.D. Ripley. *Modern Applied Statistics with* S-PLUS. Springer-Verlag, New York, third edition, 1999.