eman ta zabal zazu

Universidad del País Vasco

Euskal Herriko Unibertsitatea

TAREA 7

EJERCICIOS

1. El fichero `Rentabilidades.csv` contiene las rentabilidades sobre diferentes periodos de todos los fondos de pensiones que se comercializan en España[1] Son datos en formato CSV ("comma separated values") que puedes leer con una orden como:

```
Rentab <- read.csv("Rentabilidades.csv", header=TRUE,
   sep=";", dec=",")
```

   a) Regresa la rentabilidad en el periodo 2001-2010 sobre las variables `TipoInv` (tipo de activos en que invierte el Fondo), y `Gestora`[2].

   b) ¿Qué tipo de inversión ha sido mejor, en promedio, para el periodo reseñado? ¿Es significativa la diferencia?

   c) ¿Es significativa la diferencia de rentabilidad obtenida por las diferentes gestoras? ¿Qué gestoras, si es que alguna, parecen haberlo hecho mejor[3]?

   d) ¿Hay observaciones anómalas? ¿Influyentes?

2. Los datos correspondientes a este ejercicio están en un fichero llamado `longley.dat`, en siete columnas. Se reproducen en el Cuadro 1. 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.

[1] Los datos proceden del Ministerio de Economía y Hacienda, y están accesibles en `http://www.dgsfp.meh.es/rentabilidades/Rentabilidades.aspx`.

[2] Puedes emplear equivalentemente la variable `ClaveGest`, que proporciona etiquetas mucho más cortas; siempre puedes examinar en el fichero a qué corresponde cada clave si tienes interés en identificar a una gestora.

[3] O quizá simplemente han cobrado menos comisión de gestión por el mismo trabajo.

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

Año	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

a) Ajusta una regresión lineal de la primera columna sobre las restantes, utilizando regresión *ridge*. Dibuja la traza ridge de algunos de los parámetros. Compara los diferentes estimadores entre sí y con los MCO.

b) Haz ahora estimación ridge seleccionando el parámetro k por validación cruzada. En R, la función `lm.ridge` (del paquete MASS; mira [11], la documentación *on-line* o el ejemplo que tienes en los apuntes) hace todo el trabajo por tí.

3. Los datos en `clouds.dge` recogen los resultados de un experimento tendente a evaluar la posibilidad de provocar (o incrementar) la lluvia artificialmente. El procedimiento consiste en "inseminar" *(seed)* as nubes lanzando bengalas de yoduro de plata, sustancia que se supone provoca la condensación y aumento de tamaño de gotas de agua y favorece su precipitación.

   Los datos aparecen en el Cuadro 2. Una explicación somera de los significados de las variables es la siguiente: `seeding` es una variable cualitativa: indica si se inseminó o no la nube, decisión tomada al azar[4]. `time` es el número de días a contar desde el primer experimento (para ver si existe una pauta o tendencia temporal a lo largo de la temporada); `sne` es un "suitability criterion" o índice de idoneidad de la nube. `cloudcover`, el porcentaje de cubierta nubosa en el área experimental. `prewetness` el total de lluvia caída en el área experimental una hora antes de realizarse la inseminación. `echomotion` indica si el eco en el radar de la nube era estacionario o móvil. Por fin, `rainfall` es la variable respuesta, cantidad de lluvia caída.

[4]Para evitar, consciente o inconscientemente, inseminar aquéllas que parecen más "prometedoras"; a esto se llama "aleatorizar el experimento".

Tu trabajo consiste en contrastar con criterio estadístico si la inseminación parece haber producido un incremento en la cantidad de lluvia. Te convendrá hacer un análisis descriptivo previo, y examinar diferentes modelos hasta encontrar uno que parezca razonable. En el contexto del mejor modelo que encuentres, debes contrastar la hipótesis que se te propone.

Como parte del análisis, debes examinar los residuos y decidir: a) Si la hipótesis de normalidad de las perturbaciones parece razonable, b) Si hay alguna observación anómala y c) Si hay alguna observación influyente.

Cuadro 2: Datos sobre lluvia artificial mediante inseminación de nubes con yoduro de plata

seeding	time	sne	cloudcover	prewetness	echomotion	rainfall
no	0	1.75	13.4	0.274	stationary	12.85
yes	1	2.70	37.9	1.267	moving	5.52
yes	3	4.10	3.9	0.198	stationary	6.29
no	4	2.35	5.3	0.526	moving	6.11
yes	6	4.25	7.1	0.250	moving	2.45
no	9	1.60	6.9	0.018	stationary	3.61
no	18	1.30	4.6	0.307	moving	0.47
no	25	3.35	4.9	0.194	moving	4.56
no	27	2.85	12.1	0.751	moving	6.35
yes	28	2.20	5.2	0.084	moving	5.06
yes	29	4.40	4.1	0.236	moving	2.76
yes	32	3.10	2.8	0.214	moving	4.05
no	33	3.95	6.8	0.796	moving	5.74
yes	35	2.90	3.0	0.124	moving	4.84
yes	38	2.05	7.0	0.144	moving	11.86
no	39	4.00	11.3	0.398	moving	4.45
no	53	3.35	4.2	0.237	stationary	3.66
yes	55	3.70	3.3	0.960	moving	4.22
no	56	3.80	2.2	0.230	moving	1.16
yes	59	3.40	6.5	0.142	stationary	5.45
yes	65	3.15	3.1	0.073	moving	2.02
no	68	3.15	2.6	0.136	moving	0.82
yes	82	4.01	8.3	0.123	moving	1.09
no	83	4.65	7.4	0.168	moving	0.28

AYUDAS, SUGERENCIAS, COMENTARIOS

1. En el problema 1, considera parametrizaciones alternativas (¿te interesan contrastes `contr.sum` o `contr.treatment`?) y ten en cuenta que según cómo contrates la hipótesis de interés tendrás un problema de inferencia simultánea.

2. Puedes emplear `read.table` para leer el primer fichero y `dget("clouds.dge")` para el segundo, ambos en el lugar habitual. En el caso del primero, la 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 *dataframes*, 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`. En general, te bastará emplear la función `lm` y asociadas.

   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. Para probar diferentes modelos puede que te resulte de utilidad la función `drop1`, además de la ya vista en la tarea anterior `anova`.

5. Tienes ayuda *on line* sobre todas las funciones. También dispones del manual y de [1]. Para una descripción del output de `drop1` puedes también querer hacer `help(lm.object)`. Los manuales de [5] y [4] proporcionan ejemplos de manejo de R.

6. 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.

7. 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:

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

   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". Si empleas `lm` y absolutamente quieres hacer uso de alguna función que emplea el resultado de `lsfit`, puedes pedir a `lm` la matriz de diseño y el vector de observaciones de la variable respuesta, tal como se explica en el apartado 3 más arriba.

Referencias

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

[2] P. Dalgaard. *Introductory Statistics with R*. Statistics and Computing. Springer-Verlag, 2002. Signatura: 519.682 DAL.

[3] N. R. Draper and H. Smith. *Applied Regression Analysis*. Wiley, third edition, 1998. Signatura: 519.233.5 DRA.

[4] J. J. Faraway. *Linear Models with R*. Chapman & Hall/CRC, 2005. Signatura: 519.233 FAR.

[5] J. Fox. *An R and S-Plus Companion to Applied Regression*. Sage Pub., 2002.

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

[7] 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.

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

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

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

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