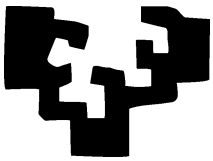
eman ta zabal zazu



Universidad del País Vasco

Euskal Herriko Unibertsitatea

TAREA 8

EJERCICIOS

1. Los datos correspondientes a este ejercicio están en un fichero de nombre `boston.dat`. Proceden de [5]. Recogen diversas variables, con posibilidad de ser influyentes en el precio y de las viviendas de una comunidad. Son datos referidos a distintos barrios del Gran Boston, y han sido profusamente utilizados como ejemplo y en un artículo seminal sobre modelos hedónicos. Son también un conjunto de datos de tamaño y características bien habituales en el quehacer de un economista.

   El objetivo es comprobar si la polución atmosférica tiene efecto discernible sobre el precio de las viviendas (representado por la variable MEDV).

   a) Ajusta un modelo a los datos. Puedes servirte como aproximación inicial de algún método de búsqueda automática junto con alguno de los criterios de ajuste (C_p de Mallows, etc.) que conoces.

   b) Representa gráficamente los residuos studentizados (en cualquier versión). ¿Hay evidencia de observaciones extrañas?

   c) Representa gráficamente los residuos (brutos o studentizados: ¿qué trascendencia tiene emplear unos u otros?) frente a $\hat{y}$. ¿Ves alguna cosa llamativa (debieras ver al menos una)?

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

   e) Dibuja los residuos contra cada una de las variables incluidas. ¿Hay algún indicio que sugiera un cambio de especificación? ¿Hay observaciones con gran influencia sobre algún o algunos parámetros?

   f) Examina la influencia de las observaciones sobre cada una de las variables subsistentes en tu modelo.

Cuadro 1: Variables en el fichero `boston.dat`

VARIABLE	SIGNIFICADO
CRIM	Tasa de criminalidad en la vecindad.
ZN	Porcentaje de terreno residencial en parcelas de más de 25000 pies cuadrados.
INDUS	Proporción de empresas no minoristas en la vecindad.
CHAS	Fachada al Charles River (= 1 si limita con el río; 0 en otro caso).
NOX	Oxidos de nitrógeno (partes en 10 millones).
RM	Número mediod e habitaciones por vivienda.
AGE	Proporción de viviendas ocupadas antes de 1940.
DIS	Distancia ponderada a cinco centros de negocios de Boston.
RAD	Indice de accesibilidad a autopistas radiales.
TAX	Impuesto sobre bienes inmuebles por \$10.000
PTRATIO	Ratio alumnos/profesor en la vecindad.
B	$1000(B - 0{,}63)^2$ siendo B la proporción de negros en el barrio.
LSTAT	% status bajo de la población.
MEDV	Mediana del valor de las viviendas ocupadas por sus propietarios (en \$1000's).

g) Considera la pertinencia y oportunidad de transformar alguna de las variables, tanto regresores como respuesta.

h) Interpreta los resultados. En el contexto del mejor modelo que hayas podido construir, ¿Encuentras evidencia de que la polución está relacionada con el precio de las viviendas? (ésta era la hipótesis inicial de los investigadores que compilaron los datos). Resume en unas pocas líneas tus hallazgos.

AYUDAS, SUGERENCIAS, COMENTARIOS

1. Para leer los datos, puedes servirte de la instrucción

```
boston <- read.csv(file="boston.dat",header=TRUE)
```

2. Lectura recomendada: Cualquiera de los manuales empleados a lo largo del curso de ayudará. Por ejemplo, [8], [9] o [7]. Los libros [2] y [3] tratan específicamente transformaciones, residuos, e influencia. [6] tiene también sendos capítulos (el 5 y 6) muy legibles sobre las mismas cuestiones. [1]. Para R tienes documentación on-line. Te vendrán bien [10] y [4].

3. Piensa en *el problema* ¿Qué sugiere el sentido común? Puedes utilizar las variables como te las dan o transformarlas si ello da un modelo más interpretable o de mejor ajuste.

4. Puedes servirte de cualquiera de los tipos de gráficos de residuos que se comentaron en clase para ayudarte en la especificación.

5. Tienes excelentes herramientas en R para visualizar datos. En particular pueden interesarte las funciones `pairs` y `coplot`, entre otras. Tienes también en R un interface a `ggobi` (lo puedes invocar cargando la librería `rggobi` e invocando a continuación `ggobi()` con un argumento de tipo dataframe).

Referencias

[1] 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.

[2] D. A. Belsley, E. Kuh, and R. E. Welsch. *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*. Wiley, New York, 1980.

[3] R. D. Cook and S. Weisberg. *Residuals and Influence in Regression*. Chapman and Hall, New York, 1982.

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

[5] D. Harrison and D. L. Rubinfeld. Hedonic prices and the demand for clean air. *J. Environ. Economics and Management*, 5(3):81–102, 1978.

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

[7] D. Peña. *Estadística Modelos y Métodos. 2. Modelos Lineales y Series Temporales*. Alianza Editorial, Madrid, 1987.

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

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

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