eman ta zabal zazu

Universidad del País Vasco
Euskal Herriko Unibertsitatea

TAREA 3

EJERCICIOS

Los ejercicios que siguen tienen por objeto proporcionar práctica adicional en el manejo de R y afianzar algunos conceptos sobre los rudimentos del modelo de regresión lineal. Puedes presentar los resultados en forma de un fichero de salida de R, debidamente comentado, y entregado por correo electrónico o, mejor, vía Moodle.

1. Esta pregunta ilustra algunos cosas aprendidas sobre proyecciones, problemas de rango incompleto y estimabilidad.

   a) Crea una matriz X de orden 5×3 con los siguientes datos:

   $$\begin{pmatrix} 5 & 6 & 11 \\ 2 & 8 & 10 \\ 3 & 6 & 9 \\ 2 & 4 & 6 \\ 1 & -2 & -1 \end{pmatrix}$$

   Observa que la tercera columna es combinación lineal de las dos primeras.

   b) ¿Existe $(X'X)^{-1}$¿Por qué?

   c) ¿Cuáles son los valores propios de $(X'X)$?

   d) Sea $X_{(12)}$ la matriz formada por las dos primeras columnas de X y $X_{(13)}$ la formada por las columnas primera y tercera. ¿Generan las columnas de $X_{(12)}$ y de $X_{(13)}$ el mismo subespacio de R^5? ¿Generan las columnas de $X_{(12)}$ el mismo subespacio de R^5 que las columnas de X?

   e) Computa:

   $$\begin{aligned} M &= (X'_{(12)}X_{(12)})^{-1} \\ N &= (X'_{(13)}X_{(13)})^{-1} \\ P &= X_{(13)}(X'_{(13)}X_{(13)})^{-1}X'_{(13)} \\ Q &= X_{(12)}(X'_{(12)}X_{(12)})^{-1}X'_{(12)} \end{aligned}$$

   ¿Qué proyecciones representan estas matrices? Compara P y Q. ¿Qué observas? ¿Es lógico?

f) ¿Son M y N iguales? Toma M y forma una matriz M_0 de dimensión 3×3 orlándola con una fila y columna de ceros:

$$M_0 = \begin{pmatrix} M & 0 \\ 0 & 0 \end{pmatrix}$$

Realiza el producto:

$$(X'X)M_0(X'X)$$

y compara el resultado con $X'X$. ¿Qué observas? ¿Qué es M_0 respecto a $(X'X)$? Guarda M_0.

g) Toma un vector arbitrario $\vec{y}$ de R^5 y computa:

$$\hat{\beta}_{(12)} = (X'_{(12)}X_{(12)})^{-1}X'_{(12)}\vec{y} \quad (1)$$

$$\hat{\beta} = (M_0)X'\vec{y} \quad (2)$$

h) Calcula ahora:

$$\vec{u} = X_{(12)}\hat{\beta}_{(12)} \quad (3)$$

$$\vec{v} = X\hat{\beta} \quad (4)$$

Compara $\vec{u}$ y $\vec{v}$. ¿Qué observas?

i) ¿Es $\hat{\beta}_{(12)}$ único siempre que ajustamos mínimo-cuadráticamente $\vec{y}$ sobre las dos primeras columnas de X?

j) ¿Es $\hat{\beta}$ único al ajustar mínimo-cuadráticamente $\vec{y}$ sobre las tres columnas de X?

k) Proporciona una restricción sobre los valores de los parámetros que convierta $\vec{\beta}$ en estimable. Comprueba, añadiendo dicha restricción a las ecuaciones normales, que el sistema resultante tiene solución única.

2. Se te proponen a continuación diversas situaciones para que expliques qué modelo lineal ajustarías, pronunciándote en particular sobre: i) La pertinencia o no de incluir una ordenada en el origen, β_0, y ii) La plausibilidad de una especificación lineal.

   a) Las observaciones Y_i son datos de consumo de carburante de un mismo vehículo, medido en N ensayos diferentes. En cada ensayo se han recorrido X_i kilómetros.

   b) Las observaciones Y_i son datos de consumo de energía eléctrica total en una población, N días consecutivos. En cada día se ha registrado la temperatura media X_i.

   c) Las observaciones Y_i son datos de consumo total de la familia i-ésima en una población, durante un cierto periodo. Las observaciones X_i recogen la renta de la familia i-ésima en el periodo anterior.

   d) Las observaciones Y_i son datos de consumo calórico de un animal de tiro. Las X_i recogen el número de kilómetros recorridos y el número de kilogramos arrastrados (en condiciones cuidadosamente controladas: mismo trayecto y mismo carro).

3. El fichero de datos `choco.dat`, contiene precios de 2006 en euros por 100 gramos de distintas marcas de chocolate. Se recoge también su contenido porcentual en cacao. Con los datos disponibles, y prescindiendo de un posible efecto "marca", estima la influencia de las variables CacaoPC y Peso.

a) ¿Hay evidencia de que los parámetros asociados a las variables en el modelo

$$\text{Precio} = \beta_0 + \beta_1\ \text{CacaoPC} + \beta_2\ \text{Peso} + \epsilon$$

son distintos de cero? (Hasta tanto avancemos en teoría, puedes recurrir a declarar significativo un parámetro estimado si excede de dos desviaciones típicas.) ¿Qué probabilidad hay de que esto suceda si el verdadero parámetro es cero y la distribución del estimador es normal? ¿Y si la distribución del estimador fuera distinta de la normal? (Para responder a esta última pregunta, puedes desempolvar la desigualdad de Tchebycheff. ¿A que no imaginabas que acabaría sirviéndote para algo?)

b) ¿Cuál es el R^2 de la regresión? ¿Qué interpretación tiene?

c) Cabría la posibilidad de que la dependencia del precio sobre el contenido en cacao fuera no lineal. Por ejemplo, por cada punto porcentual adicional de cacao, el precio podría aumentar más en los tramos altos (chocolates "de lujo") que en los bajos. O viceversa. Especifica y estima un modelo que de cuenta de esta eventualidad.

d) Compara los modelos estimados en los apartados $3a$ y $3c$. ¿Qué ocurre con la R^2?

e) Imagina que, llevado de tu ardor de recién llegado al mundo de la regresión lineal, ajustas un modelo en que Precio depende del contenido en cacao de una forma muy elaborada: nada menos que un polinomio de grado 17. Así:

$$\begin{aligned}\text{Precio} &= \beta_0 + \beta_1\ \text{CacaoPC} + \beta_2\ (\text{CacaoPC})^2 + \\ &\quad + \ldots + \beta_{17}\ (\text{CacaoPC})^{17} + +\beta_{18}\ \text{Peso} + \epsilon\end{aligned}$$

Tras estimar el modelo, observas que ¡los residuos son todos cero! ¿Has dado con el Santo Grial de los modelos que explican el precio del chocolate, o lo que te ocurre es algo esperable (y sin mucho interés)? Explica.

AYUDAS, SUGERENCIAS Y COMPLEMENTOS

1. Cuando se te pregunta si una matriz es singular (o si existe su inversa) has de dar un argumento que justifique tu respuesta. Decir que una matriz es singular "porque el determinante es cero" es tautológico. No obstante, puedes querer asegurarte de algo calculando un determinante. En R tienes una función `det` que lo hace directamente.

2. En R puedes calcular una inversa generalizada mediante la función `ginv` de la biblioteca `MASS` (has de hacer un `library(MASS)` previo). No estaría de más que al menos la primera vez la calcularas por el procedimiento descrito en clase.

3. Los vectores y valores propios pueden ser calculados en R mediante la función `eigen`. Observa que una matriz simétrica puede no parecerlo en coma flotante, y la función te puede dar valores propios tales como `1.2345234 +2.39485-16i` que a todos los efectos prácticos son reales.

4. Sobre R tienes, además de [14], las notas [11] o su traducción española [12]. Libros recientes que también puedes querer mirar son [3], [6], y [5]. Documentación *on line* incluye [8] y todo lo que puedes encontrar en CRAN (tienes una réplica en el Laboratorio: `http://b012526.bs.ehu.es/cran`, sólo accesible desde dentro de la UPV/EHU).

5. Para leer un fichero de datos con estructura tabular (como `choco.dat`) puedes hacer uso en R de la instrucción `read.table`. Así:

```
> choco <- read.table(file = "../datos/choco.dat", header = TRUE)
> choco[1:5, ]

  Peso CacaoPC Precio    Marca
1   50      99   4.39    Lindt
2  100      74   0.77 Consumer
3  200      70   1.15    Valor
4   80      70   1.94   Mascao
5  100      70   1.47    Lindt
```

   Debes sustituir `"../datos/choco.dat"` por el camino completo hasta tus datos. Si trabajas en Windows, será seguramente algo como `C:/datos/choco.dat`. Si trabajas en el Laboratorio y descargas los datos al directorio en que estás trabajando, bastará algo como `çhoco.dat"`. El `header=TRUE` indica que la primera línea del fichero nombra las variables.

6. Si quieres hacer regresión sobre distintas potencias de una misma variable con la función `lsfit`, la parte más complicada consiste en construir la matriz de regresores. Hasta tanto nos familiarizamos con funciones más potentes y elaboradas que nos dispensarán de este trabajo, puedes recurrir a la siguiente receta:

```
> CacaoPC <- choco[, 2]
> potencias <- outer(CacaoPC, 1:5, FUN = "^")
> potencias[1:5, ]

     [,1] [,2]   [,3]     [,4]       [,5]
[1,]   99 9801 970299 96059601 9509900499
[2,]   74 5476 405224 29986576 2219006624
[3,]   70 4900 343000 24010000 1680700000
[4,]   70 4900 343000 24010000 1680700000
[5,]   70 4900 343000 24010000 1680700000
```

   (la explicación, en clase, o mira la ayuda para la función `outer`).

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

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

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

[5] J. J. Faraway. *Linear Models with R*. Chapman & Hall/CRC, 2004. Signatura: 519.233 FAR.

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

[7] J. H. Grafe. *Matemáticas Universitarias*. MacGraw-Hill, Madrid, 1985.

[8] P. Kuhnert and W. Venables. *An Introduction to R: Software for Statistical Modelling and Computing*. CSIRO Mathematical and Information Sciences, Cleveland, Australia, 2005.

[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] B. Venables, D. Smith, R. Gentleman, and R. Ihaka. *Notes on R: A Programming Environment for Data Analysis and Graphics*. Dept. of Statistics, University of Adelaide and University of Auckland, 1997. Available at `http://cran.at.r-project.org/doc/R-intro.pdf`.

[12] B. Venables, D. Smith, R. Gentleman, R. Ihaka, and M. Mächler. *Notas sobre R: Un Entorno de Programación para Análisis de Datos y Gráficos*, 2000. Traducción española de A. González y S. González.

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