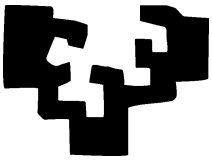
eman ta zabal zazu



Universidad del País Vasco
Euskal Herriko Unibertsitatea

TAREA 5

EJERCICIOS

1. Del modo que ya conoces, genera 100 muestras con 20 observaciones cada una de una variable aleatoria definida así:

$$Y_i = \beta_0 + \beta_1 X_1 + \ldots + \beta_4 X_4 + \varepsilon \tag{1}$$

   Los regresores los puedes escoger a tu antojo (pero cuidando de que no haya dependencias lineales entre las columnas de la matriz X). Los β's también puedes escogerlos como quieras. Emplea un nivel de significación $\alpha = 0{,}05$ en lo que sigue.

   a) Calcula para cada muestra el estadístico Q_h resultante de contrastar la hipótesis (cierta) de que los parámetros tienen por valores precisamente los que les has dado. Compara la distribución teórica con la empírica.

   b) Repite ahora el experimento, pero calculando el estadístico Q_h asociado al contraste de una nueva hipótesis, esta vez falsa. Observa su distribución empírica (puedes hacer un histograma de sus valores).

   c) En el apartado anterior, puesto que h no es cierta, sería deseable que se produjera su rechazo. Recordarás que la probabilidad de dicho rechazo se llama *potencia* del contraste. Estíma dicha potencia a la vista de los resultados de la simulación en (1*b*).

2. El fichero `presis.dat` está en el lugar habitual. Contiene datos sobre la edad en el momento de la elección y filiación política de los presidentes norteamericanos, junto con la fecha de la elección, desde 1853. La primera variable toma valor REP ó DEM, según el presidente correspondiente fuera republicano o demócrata. La segunda, recoge los años en que fueron elegidos. La tercera, las edades respectivas en el momento de la elección. Las observaciones están rotuladas con los nombres de los respectivos presidentes[1].

   a) Haz un gráfico representando la edad en el momento de la elección (Y) frente a la fecha de la elección (X). Emplea un caracter o color diferentes para representar a demócratas y republicanos. Puedes emplear las funciones `plot` y `points`.

[1] En `http://www.whitehouse.gov/history/presidents/chronological.html` puedes encontrar si lo deseas información biográfica adicional.

El gráfico sugiere que la edad media de los presidentes en el momento de alcanzar la presidencia ha ido aumentando. Sugiere también muchas otras interesantes preguntas. Para contestarlas puedes utilizar modelos de regresión, con cuya ayuda debes contestar lo siguiente.

b) ¿Que modelo te parece que describe mejor los datos? ¿Por qué?

c) En conexión con la pregunta anterior: justifica la inclusión o no de un parámetro β_0 multiplicando la columna de "unos", y explica, en su caso, cuál sería su interpretación. Si *no* lo incluyeras, explica cual sería la interpretación de los demás parámetros que sí sean incluídos..

d) ¿Ves evidencia en favor de la hipótesis de que la edad media de los presidentes en el momento de la elección ha ido creciendo?

e) ¿Ves evidencia en favor de la hipótesis de que los candidatos republicanos acostumbran a ser elegidos (quiza por su carácter más conservador) a edad más avanzada?

f) Supón que el modelo adecuado fuera: $Edad = \beta_0 + \beta_1(Fecha) + \beta_2(Fecha)^2 + \varepsilon$. Contrasta la hipótesis de que los modelos generando las dos submuestras (la de republicanos y demócratas) tienen sus parámetros correspondientes iguales.

g) Supón que quisieras contrastar la hipótesis de que la edad de los presidentes republicanos ha ido creciendo linealmente con el tiempo, pero la de los demócratas no. ¿Qué modelo ajustarías? ¿Qué contraste realizarías?

3. Un *geyser* es una fuente de agua caliente que, de tanto en tanto, experimenta variaciones de régimen y produce una erupción de vapor o agua sobrecalentada. El mecanismo físico subyacente es complejo y no perfectamente predecibl. La experiencia muestra, sin embargo, que aunque las erupciones no se producen a intervalos regulares, hay cierta relación entre la duración de una erupción y el tiempo que transcurre hasta la siguiente.

   Uno de los *geyser* más famosos está situado Yellowstone National Park, y se conoce como Old Faithfule[2]. Entra en erupción a intervalos de entre 40 y 100 minutos, y cada erupción dura entre 1 y 6 minutos. Los guardas han observado que, cuanto más larga es una erupción, más tarda en presentarse la siguiente, y emplean la fórmula empírica $T = 30 + 10d$, en que T es el tiempo aproximado hasta la siguiente erupción, y d es la duración de la erupción previa.

   a) ¿Puedes tú hacerlo mejor? En el fichero `geyser.dat` tienes datos correspondientes a 272 erupciones consecutivas[3], acaecidas en Octubre de 1.980. Las variables recogidas son duración e intervalo, ambas en minutos; una representación gráfica de una frente a otra aparece en los apuntes.

   b) Supón que quieres llevar a los turistas a ver el *geyser* en el momento oportuno: tienen que no esperar mucho. ¿Cómo lo harías? (Ayuda: si les llevas en el momento "justo" desde un punto de vista mínimo cuadrático, tienes garantizado llegar un poco antes o un poco después, con lo que aproximadamente la mitad de las veces tus turistas perderán por los pelos la primera erupción y habrán de esperar a la siguiente. Tu estrategia ha de ser otra; por ejemplo, podrías llevarles en el último momento tal que con probabilidad 0.975 alcancen a ver la erupción inmediata).

[2] Puede interesarte buscar "Old Faithfull geyser" en Google; hay toneladas de vídeos y explicaciones.

[3] Los datos proceden de Cook & Weisberg (1982).

AYUDAS, SUGERENCIAS, COMPLEMENTOS

1. Los datos en el fichero `presis.dat` están encolumnados con formato libre. Una manera cómoda de leerlos es mediante la función `read.table`. Por ejemplo, así:

   ```
   > presis <- read.table(file="presis.dat",header=TRUE)
   ```

   lo que producirá (sólo se muestran las primeras líneas:

   ```
   > presis[1:5,]
   ```

   ```
                      Partido FechaElec Edad
   Franklin Pierce        DEM      1853   48
   James Buchanan         DEM      1857   65
   Abraham Lincoln-I      REP      1861   52
   Abraham Lincoln-II     REP      1865   56
   Andrew Johnson         REP      1865   56
   ```

2. El ejemplo con las edades de los presidentes tiene sólo finalidad didáctica. Pero ilustra algunas cuestiones de interés. Entre otras, muestra que un modelo proporciona ajuste, en el mejor de los casos, en una cierta región. Claramente, si se encuentra una tendencia creciente en las edades, no tiene sentido hacer inferencias del tipo: "En el siglo XXIII los presidentes demócratas de Estados Unidos serán elegidos a una edad media de 230 años".

3. Ambos ejemplos —el de los presidentes y el del *geyser*— presentan una peculiaridad: la matriz X no puede ser escogida por el analista. Es sólo observada. No podemos decidir la edad que tendrá el próximo presidente a añadir a la muestra. Tampoco la duración de las erupciones del *geyser*. Esta situación nos coloca fuera de la teoría estudiada (uno de nuestros supuestos era que la matriz X era fija, no estocástica: aquí la matriz X podría verse como aleatoria).

   Ello no obstante, toda la teoría "pasa". Basta que consideremos nuestra inferencia como condicional en los valores observados de las X. En Econometría estudiarás lo que ocurre cuando los regresores son aleatorios, problema que nosotros soslayaremos.

4. Ambos ejemplos muestran también que está fuera de lugar, en general, una interpretación causal del tipo "la duración de la erupción *causa* intervalos más largos entre erupciones" o "el hecho de que los presidentes sean Republicanos *causa* que sean elegidos a edades más tardías". Lo que los modelos recogen es sólo una asociación lineal entre diferentes variables, cualquier interpretación causal que añadamos es de nuestra exclusiva responsabilidad.

5. Hay una característica adicional en ambos ejemplos que puede ser explotada, y de la que tampoco nos ocuparemos: los datos presentan una ordenación natural (en ambos casos, a lo largo del tiempo). Cuando esto ocurre, suele existir dependencia temporal entre las observaciones, y esta dependencia puede explotarse para mejorar la estimación. De nuevo, éste es un problema típicamente econométrico, y lo veréis tratado en la asignatura de Econometría. Cuando estudiéis lo que es autocorrelación, podéis desear volver sobre los datos del *geyser* para comprobar que se puede hacer mejor de lo que lo habéis hecho.

La ordenación natural no tiene porqué limitarse al tiempo (aunque éste sea con mucho el caso más frecuente). Puede haber también una ordenación espacial. Como en el caso temporal, cabe esperar dependencia entre observaciones contiguas.

Una sucesión de variables aleatorias en que hay un orden natural se denomina un *proceso estocástico* (esto dista de ser una definición digna de tal nombre). Una realización (= una muestra) procedente de un proceso estocástico es una *serie temporal*. El estudio de series temporales, a caballo entre Econometría y Estadística, es una materia que no abordaremos en esta asignatura.

6. Además de Seber & Lee (1998) —el libro al que más nos aproximamos— tenéis manuales como Draper & Smith (1998), Trocóniz (1987), Stapleton (1995) o Myers (1990), cuya consulta os ayudará. Para cosas relacionadas con R (y por muchos atinados comentarios en relación con cuestiones estadísticas de fondo) es útil Faraway (2005).

7. Sobre las funciones gráficas `plot` y `points` que se te sugiere utilizar, puedes ver la ayuda *on line* o libros como Venables & Ripley (1999*a*) (completado con Venables & Ripley (1999*b*)), Fox (2002), Dalgaard (2002), Kuhnert & Venables (2005), Ugarte et al. (2008), o los ejemplos intercalados entre los apuntes)

8. Para poderte referir directamente a las variables en una *data frame*, puedes emplear `with`. Por ejemplo, puedes teclear:

   ```
   > with(presis, plot(FechaElec,Edad))
   ```

   Sin el `which`, deberías hacer uso de la notación, más farragosa,

   ```
   > plot(presis[,"FechaElec"],presis[,"Edad"])
   ```

   Una alternativa (menos recomendable) es usar `attach`, cuyo efecto dura toda la sesión, y se anula mediante un `detach`.

9. R ofrece ayuda en el análisis exploratorio. Si dibujas puntos mediante, por ejemplo,

   ```
   > plot(FechaElec,Edad)
   ```

   (lo que requiere un attach previo) puedes rotularlos interactivamente. Teclea algo como

   ```
   > identify(FechaElec,Edad,labels=rownames(presis))
   ```

   y podrás identificar cada punto haciendo "click" sobre él con el botón izquierdo del ratón. *Ten cuidado de acabar apretando el botón derecho de tu ratón, de otro modo colgarás Emacs.* La función `identify` devuelve un vector con los números de las observaciones que has señalado.

Referencias

Cook, R. D. & Weisberg, S. (1982), *Residuals and Influence in Regression*, Chapman and Hall, New York.

Dalgaard, P. (2002), *Introductory Statistics with R*, Statistics and Computing, Springer-Verlag. Signatura: 519.682 DAL.

Draper, N. R. & Smith, H. (1998), *Applied Regression Analysis*, third edn, Wiley. Signatura: 519.233.5 DRA.

Faraway, J. J. (2005), *Linear Models with R*, Chapman & Hall/CRC. Signatura: 519.233 FAR.

Fox, J. (2002), *An R and S-Plus Companion to Applied Regression*, Sage Pub.

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

Myers, R. H. (1990), *Classical and Modern Regression with Applications*, PWS-KENT Pub. Co., Boston.

Seber, G. A. F. & Lee, A. J. (1998), *Linear Regression Analysis*, Wiley.

Stapleton, J. H. (1995), *Linear Statistical Models*, Wiley, New York.

Trocóniz, A. F. (1987), *Modelos Lineales*, Serv. Editorial UPV/EHU, Bilbao.

Ugarte, M., Militino, A. & Arnholt, A. (2008), *Probability and Statistics with R*, CRC Press.

Venables, W. & Ripley, B. (1999*a*), *Modern Applied Statistics with S-Plus*, third edn, Springer-Verlag, New York.

Venables, W. & Ripley, B. D. (1999*b*), 'R complements to *Modern Applied Statistics with S-Plus.*', En `http://www.stats.ox.ac.uk/pub/MASS3`.