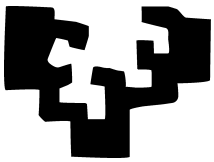

Universidad del País Vasco
Euskal Herriko Unibertsitatea

TAREA 5

EJERCICIOS

1. Del modo que ya conoces, genera 50 muestras con 20 observaciones cada una de una variable aleatoria definida así:

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

   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. Deduce analíticamente el valor medio de la distribución de Q_h. Observa su distribución empírica (puedes hacer un histograma de sus valores), y examina si hay acuerdo aproximado entre el valor medio teórico que has calculado y el muestral.

   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*).

   d) (OPTATIVO) Sea $h : A\vec{\beta} = \vec{c}$ y $\vec{\delta} = A\vec{\beta} - \vec{c}$. Una medida razonable de la "incorrección" de h vendría dada por la norma de $\vec{\delta}$. Cabría esperar que, cuanto mayor es dicha norma (= más "incorrecta" es h), mayor sería la probabilidad de rechazar h (= potencia). Comprueba empíricamente que éste es efectivamente el caso, repitiendo la simulación en (1*b*) para hipótesis h progresivamente "más falsas" (con mayor $\vec{\delta}$).

   e) (OPTATIVO) ¿Cuál sería el parámetro de no centralidad de la F de Snedecor cuando, como en el apartado anterior, h no se verifica? Calcúlalo para la hipótesis h incorrecta que hayas escogido. ¿Depende dicho parámetro del tamaño muestral (del número de filas de X)? ¿Es razonable que sea así?

2. El fichero `presis.dat` está en el lugar habitual. La primera variable toma valor 1 ó 0, 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.

   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`.

   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 incluyas.

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

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 no es bien conocido. 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 Faithful. 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[1], acaecidas en Octubre de 1.980. Las variables recogidas son intervalo y duración; 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).

AYUDAS, SUGERENCIAS, COMPLEMENTOS

1. Hay que hacer bastantes simulaciones. Escribe las instrucciones y depúralas haciendo unas pocas. Sólo al final realiza todas las que te piden en cada apartado.

   Para ello puedes definir una variable `iter` al comienzo, y escribir todo lo que sigue en función de `iter`. Mientras hagas pruebas, `iter` puede tener asignado el valor 2. Cuando estés seguro de que todo es correcto, basta que asignes a `iter` el valor que sea (cincuenta, en esta tarea, para el ejercicio 1) y envíes tu trabajo a la cola de batch.

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

[1] Los datos proceden de [3].

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 se estudia lo que ocurre cuando los regresores son aleatorios, problema que nosotros soslayaremos.

4. 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.

   Quizá no esté de más añadir que 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.

5. Además de [6] —el libro al que más nos aproximamos— tenéis manuales como [8], [7] o [4], cuya consulta os ayudará.

6. Sobre las funciones gráficas `plot` y `points` que se te sugiere utilizar, puedes ver la ayuda on-line o libros como [9] (completado con [10]). También los ejemplos intercalados entre los apuntes [5].

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

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

[5] V. Núñez and F. Tusell. Regresión lineal y análisis de varianza. Notas de clase, Octubre 2002.

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

[7] J.H. Stapleton. *Linear Statistical Models*. Wiley, New York, 1995.

[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.

[10] W.N. Venables and B.D. Ripley. 'R' complements to Modern Applied Statistics with S-PLUS. En `http://www.stats.ox.ac.uk/pub/MASS3`, 1999.