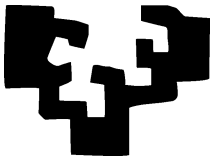
eman ta zabal zazu



Universidad del País Vasco

Euskal Herriko Unibertsitatea

TAREA 6

EJERCICIOS

Los ejercicios que siguen tiene por objeto proporcionar alguna práctica en el uso de R para especificar y estimar un modelo de regresión lineal

1. Los datos para este ejercicio están en la *dataframe* `vida.dge`, en el lugar habitual. Leelos mediante la función `dget`. La interpretación de las diferentes columnas en dicha dataframe es la siguiente:

Nombre	**Tipo**	**Significado**
`Natal`	Numérica	Nacimientos por 1000 habitantes.
`Mortal`	Numérica	Defunciones por 1000 habitantes.
`MortInf`	Numérica	Mortalidad infantil por 1000 habitantes.
`EspVida`	Numérica	Esperanza de vida en años.
`Sexo`	Cualitativa	Hombre o mujer.
`PNBpc`	Numérica	Producto nacional bruto en US$ *per capita.*
`Grupo`	Cualitativa	Grupo de países.
`País`	Cualitativa	País.

Cuadro 1: Listado de variables en la *dataframe* `vida`.

Los datos proceden de Day, A. (ed.) (1992), *The Annual Register 1992*, London: Longmans y U.N.E.S.C.O. *1990 Demographic Year Book (1990),* New York: United Nations. [1]

   a) Haz un análisis descriptivo, previo a cualquier otra cosa.

   b) Parece evidente que la esperanza de vida (`EspVida`) ha de estar relacionada con algunas o todas las restantes variables. ¿Influye el sexo en la esperanza de vida? ¿Influye el grupo de países? ¿El producto nacional bruto *per capita*?

[1] Obtenidos de `http://www.amstat.org/publications/jse/jse_data_archive.html` y ligeramente editados para adaptarlos al uso que se les da en esta tarea.

c) ¿Influye *de modo diferenciado* el grupo de países sobre la esperanza de vida de hombres y mujeres?

2. Los datos correspondientes a este problema están en un fichero llamado `camionero.dat`. Se trata del cuaderno de ruta de un camionero metódico que antes de iniciar un porte llena su depósito de gasoil y reposta en ruta las veces que necesite, apuntando siempre los kilómetros recorridos desde el último llenado. Los kilometrajes están recogidos en la primera columna, y las columnas 2-15 recogen los litros repostados en diferentes gasolineras. Los trayectos son todos homogéneos en cuanto a cargas y orografía.

   a) Un rumor se extiende por la carretera: algunas estaciones de servicio han manipulado los contadores del surtidor de combustible, de forma que sirven menos litros de los que cobran. Nuestro camionero os aporta su cuaderno de ruta, y os pregunta si encontráis evidencia de que los litros repostados en distintas gasolineras "cunden" desigualmente. ¿Cómo lo harías? (Ayuda: Al final tienes algunas orientaciones).

   b) ¿Variaría tu modo de operar si te dijera: "Me sospecho que la gasolinera correspondiente a la columna 13 me está estafando"?

   c) En el apartado (2*a*) puedes haber hecho dos cosas, según tu interés fuera sólamente detectar el posible fraude en *alguna* gasolinera o en detectar *qué gasolinera(s) parece(n) estar cometiéndolo*. ¿Cuál sería la probabilidad de que culpabilizaras indebidamente *al gremio* de gasolineras? ¿Cuál sería una cota superior de la probabilidad de que culpabilizaras a una gasolinera inocente si te limitaras a comparar cada t-ratio con valores críticos en una distribución t de Student?

3. Ha habido una ingente cantidad de investigación dedicada a analizar si de forma predecible la Bolsa sube o baja en épocas determinadas. Se ha hablado así de que Octubre es un mal mes para la Bolsa (dos *cracks* históricos acontecieron en Octubre, en 1929 y 1987, al que hay que agregar el reciente de 2008; pero el mes de Octubre de 2006 y otros muchos han sido excelentes). Se ha buscado un "efecto Enero", "efecto Lunes", "efecto Martes", y todo tipo de cosas imaginables: ¡frecuentemente, encontrándolas!

   Los datos en la *dataframe* `Bolsa` (lee con un `dget("Bolsa.dge")` contienen datos de la Bolsa de las Batuecas, correspondientes a 2000 sesiones. La variable RENDIMIENTO contiene los rendimientos diarios. La variable SANTO es una variable cualitativa con 200 niveles, correspondiendo a otras tantas festividades del santoral.

   a) Haz una regresión para ver si la variable SANTO influye en los rendimientos.

   b) Observa los resultados: el nivel 98 de la variable SANTO corresponde a San Pancracio. El nivel 163 corresponde a Santa Filomena. ¿Se deduce —nota el signo de los β's— que ambos santos son funestos para las cotizaciones, y que hay que vender lo que se tenga la víspera? Explica.

AYUDAS, SUGERENCIAS, COMPLEMENTOS

1. **Datos observacionales versus diseño experimental.** Observa: uno de los supuestos que hicimos al desarrollar la teoría era que la matrix X es no aleatoria: fijamos los valores de X que deseamos, miramos los de las y y ajustamos nuestros modelos.

   Una situación en que el analista puede fijar los valores de los regresores permite un *experimento diseñado;* podemos hacerlo de modo que no haya problemas de multicolinealidad y, en general, de modo que rentabilicemos al máximo cada observación.

   En Ciencias Sociales esto será la excepción más que la regla: en general, tenemos *una* muestra, que es todo lo que hay y no ha sido escogida por nosotros. Es difícil exagerar la importancia de distinguir entre observación y experimentación, entre muestra aleatoria y lo que se ha dado en llamar *grab set* —un conjunto de datos que cae en nuestro poder sin que hayamos podido controlar como se generan—. Muchas cosas completamente injustificadas se hacen por ignorar esta distinción,

   Es preciso tener particular cuidado para no extrapolar conclusiones fuera del ámbito que la muestra cubre. Si en el Ejercicio 1 tuviéramos datos sólo de países pobres o de un continente, sería temerario ajustar un modelo a dichos países y dar por sentado que es de aplicación a todos los demás.

   Finalmente, cuando se tienen datos observacionales hay que estar alerta ante la posibilidad de que se produzca falta de datos *relacionada con el fenómeno que tratamos de estudiar*. Si la falta de datos se produce completamente al azar (MCAR = "missing completely at random"), el único efecto es disminuir el tamaño de la muestra. Cuando se produce de otro modo, podemos tener sesgos de selección: los datos entrar en nuestra muestra *de forma no independiente del fenómeno que estudiamos*, con lo que ciertos segmentos pueden estar sobre- o infrarrepresentados.

2. **Inferencia simultánea y *data mining*.** Las ideas que el tema de inferencia simultánea pretende transmitir son particularmente importantes en el contexto actual, en que, sobre todo en Marketing, se practica intensivamente el *data mining*. Se trata de técnicas tendentes a encontrar rasgos interesantes en ficheros masivos de información, al objeto de segmentar clientelas, hallar nichos de mercado, diseñar nuevos productos, etc.

   Cuando se procesa una información masiva sin una hipótesis previa, *dejando que los datos sugieran las hipótesis*, se ha de ser consciente de que esas hipótesis sugeridas por los datos no pueden ser contrastadas como si fueran hipótesis previas, preexistentes. Regresando al ejemplo del camionero, no podemos examinar los t-ratios de todas las gasolineras para a continuación hacer un contraste ordinario sobre el t-ratio mayor: si lo hiciéramos sin corregir consecuentemente el nivel de significación, estaríamos dando un nivel de significación incorrecto.

   Incluso a un nivel introductorio como el del presente curso, es importante que te esfuerces en entender lo que los ejercicios 2 y 3 tratan de comunicar.

   Una referencia (bastante avanzada) sobre esta cuestión es [11]. Una discusión, no técnica, divertida y demoledora, del uso inadecuado que se hace de la Estadística, es [9].

3. **Riqueza, salud, esperanza de vida.** En el ejercicio 1, junto al ajuste de modelos, seguramente te convendrá hacer unos pocos gráficos que ilustren lo que los datos muestran. La función `plot(x,y)` representa los valores de dos variables en un plano. Un poco más sofisticada es la función `coplot` que hace lo propio condicionando sobre valores de una tercera variable. Prueba, por ejemplo,

```
attach(vida)
coplot(Natal ~ PNBpc | Grupo)
```

La idea de realizar gráficas de una variable frente a otra condicionando sobre los valores de una tercera (o más de una) puede ser llevada mucho más lejos, como en los gráficos llamados Trellis (puedes ver, por ejemplo, el paquete de R llamado `lattice` o [6] y [7]).

Sobre el mismo tema de utilización de gráficos para investigar la relación entre riqueza y salud, puedes ver también:

`http://www.ted.com/index.php/talks/hans_rosling_shows_the_best_stats_you_ve_ever_seen.html`

(todo en una línea).

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, 2004. Signatura: 519.233 FAR.

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

[6] Deepayan Sarkar. Lattice. *R News*, 2(2):19–23, June 2002.

[7] Deepayan Sarkar. *Lattice. Multivariate Data Visualization with R*. Springer, 2008.

[8] G. A. F. Seber and A. J. Lee. *Linear Regression Analysis*. Wiley, 1998.

[9] N.N. Taleb. *The Black Swan.* Random House, 2007.

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

[11] C. Wang. *Sense and Nonsense of Statistical Inference*. Marcel Dekker, New York, 1993.