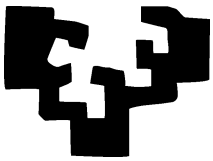
eman ta zabal zazu



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*?

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

[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.

2. 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 $\beta$'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. **Interacción.** La búsqueda de una respuesta al apartado 1*c* te llevará de modo muy natural al concepto de interacción. Si queremos estudiar el efecto sobre la esperanza de vida de dos variables cualitativas (sexo y región geográfica), no podemos emplear una modelización aditiva de los dos efectos, tal como `EspVida ~ Sexo + Grupo`. En un modelo así, los coeficientes de cada una de las variables (que `lm` desdobla automáticamente en el número de niveles preciso), recogen su influencia sobre la respuesta *sea cual fuere el valor que toma la otra.*

   Lo que querríamos es ver el efecto de cada combinación posible `Sexo x Grupo`, codificando cada combinación con una columna de ceros y unos. Si alguno o algunos de los parámetros multiplicando a estas columnas fuera significativamente distinto de cero, tendríamos evidencia de un efecto diferenciado.

   La función `lm` te simplifica bastante la vida, construyendo automáticamente las columnas de ceros y unos precisas, y eliminado las redundantes. Basta que especifiques algo como

   `EspVida ~ Sexo +Grupo + Sexo*Grupo`.

   De hecho, bastaría que especificaras `EspVida ~ Sexo*Grupo`; cuando incluyes un término de interacción, `lm` incluye automáticamente todos los términos de orden inferior.

2. **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 cómo 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. Por ejemplo, si en un estudio en que preguntamos sobre la renta disponible aquéllas familias con una renta más alta rehusan responder en mayor proporción, el efecto no será sólo que se reduce el tamaño de nuestra muestra: hay un *sesgo de selección:* la muestra se ve enriquecida en familias de rentas medias y bajas. Si estimáramos con dicha muestra la renta media sin tener presente este efecto, tendríamos un sesgo a la baja.

3. **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 problema de los datos bursátiles no podemos examinar los $t$-ratios de todos los días 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 el segundo ejercicio trata de comunicar.

   Una referencia (bastante avanzada) sobre esta cuestión es [12]. Una discusión, no técnica, iconoclasta y demoledora, del uso inadecuado que se hace de la Estadística en Finanzas es [10].

4. **Riqueza, salud, esperanza de vida.** En el ejercicio 1 se te pide un análisis descriptivo previo. Recuerda las cosas que aprendiste en EPE. En particular, 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. Otro gráfico de gran utilidad exploratoria que visteis en EPE es el boxplot; mira la función del mismo nombre en R.

   A partir de aquí, puedes invertir el tiempo y esfuerzo que quieras en representar y examinar tus datos sin que R se convierta nunca en un factor limitativo: estás limitado sólo por tu imaginación. Lo que sigue son sólo algunas sugerencias, totalmente optativas.

   Un poco más sofisticada que la función `plot` es la función `coplot` que hace lo mismo que `plot` pero 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` ó [7] y [8]).

Puedes visualizar tus datos desde todas las perspectivas imaginables, tres dimensiones a la vez, e incluso rotarlos en tiempo real[2]. Mira el paquete `ggplot2` (sobre el que, aparte la documentación *on-line*, puedes leer en [13]).

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). Con algún trabajo, puedes crear gráficos animados como los de dicha presentación en R; mira la función `Rosling.bubles` en el paquete `animation`.

Finalmente, aunque no es el objeto de esta tarea, puedes presentar información espacial sobre un mapa, también desde R; tienes un ejemplo en

`http://ryouready.wordpress.com/2009/11/16/infomaps-using-r-visualizing-german-unemployment-rates-by-color-on-a-map/`

(todo en una línea) y referencias que contiene (también puedes ver [1]).

# Referencias


[1] Roger S. Bivand, Edzer J. Pebesma, and Virgilio Gómez-Rubio. *Applied Spatial Data Analysis with R*. Springer Verlag, 2008.

[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, 2005. Signatura: 519.233 FAR.

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

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

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

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

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

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

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

[13] H. Wickham. *ggplot2 : elegant graphics for data analysis*. Springer-Verlag, 2009.


[2] Esto último, mejor en el LEC: si lo haces desde una ubicación remota sin una conexión muy rápida, es probable que no funcione bien.