eman ta zabal zazu

Universidad del País Vasco
Euskal Herriko Unibertsitatea

TAREA 11

EJERCICIOS

1. Tienes un conjunto de datos en R sobre cáncer de esófago y diversos factores de riesgo, que puedes recuperar así:

   ```
   library(datasets)
   data(esoph)
   ```

   Los datos son casi autoexplicativos, pero un `help(esoph)` aclarará cualquier duda que puedas tener.

   a) Haz un análisis de los datos tal cual te los dan. Observa que están agrupados, y la forma de invocar `glm` es un poco diferente de la vista en clase. La tienes detallada en la página de ayuda de `esoph`.

   b) Los regresores son *factores ordenados*. A efectos puramente ilustrativos, conviértelos en factores no ordenados así:

   ```
   age <- factor(unclass(agegp))
   levels(age) <- levels(agegp)
   ```

   (análogamente para `tobgp` y `alcgp`) y repite el análisis en el número anterior. Lee en el Capítulo 7 de las notas de clase las diferencias entre ambos tipos de regresores y compara los ajustes que obtienes en este apartado y el anterior.

   c) Resume en pocas líneas tus hallazgos. ¿Influye la ingesta de alcohol en el desarrollo de cáncer de esófago? ¿El fumar? ¿La edad? ¿Influye alguna de estas cosas en combinación con otra *más de lo que cabría esperar de la consideración aditiva de los efectos*?

2. Como consecuencia de sus actividades con una becaria, consideradas inadecuadas por parte de la opinión pública de los EE.UU., se abrió al anterior presidente de los Estados Unidos, Bill Clinton, un proceso tendente a lograr su remoción del cargo. Se le hicieron dos cargos (perjurio y obstrucción a la justicia). Los datos en `impeach.frame` (léelos mediante un `dget`) recogen el voto de cien senadores sobre ambos cargos y algunas variables como Estado de procedencia, porcentaje de votos

de Clinton en las anteriores elecciones en dicho Estado, si se trataba de un senador bisoño o experimentado, y el año en que debía presentarse a la reelección o retirarse. Hay también un "índice de conservadurismo", medido en una escala de 0 a 100.

Tu trabajo consiste en examinar si alguna de las variables explicativas anteriores permite predecir el voto de los senadores.

AYUDAS, SUGERENCIAS Y COMPLEMENTOS

1. Puedes servirte de textos como [6], que tiene una introducción a los modelos de variable binaria, o más especializados como [4] y [5].

2. Te será de utilidad la función `glm` (modelo lineal generalizado). Funciona con una sintaxis muy similar a `lm`, que te resultará familiar. Pertenece al grupo de funciones "nuevas", y la encontrarás documentada en [2] y en la ayuda *on line* (pero no en [1]).

   `glm` es una función sumamente general, que en particular permite hacer regresión logística si especificamos

   ```
   family=binomial(link=logit).
   ```

   Si consultas la documentación de `family` verás que hay otras muchas posibilidades, que se obtienen especificando una combinación de la función de enlace *(link)* y de la distribución de las perturbaciones. El "lado derecho" de la fórmula de regresión, es siempre lineal. Si se quiere relajar también este supuesto, hay una familia aún más amplia de modelos (aditivos generalizados) que permite hacerlo. No lo haremos en este curso, pero puedes mirar la documentación de `gam` y [3] si tienes interés.

3. En las notas de clase tienes completamente desarrollado un ejemplo que estudia la influencia de diversas variables sobre la propensión a contraer cáncer de esófago. Tanto edad como ingesta de alcohol y consumo de tabaco son variables que en teoría podrían registrarse de modo continuo. No obstante, han sido categorizadas. Hay diferencia, sin embargo, con variables categóricas nominales: aquí hay un orden natural.

   S-PLUS y R soportan el concepto de *factores* (variables categóricas, que toman uno de varios estados sin orden natural entre ellos: sexo, raza, nacionalidad, profesión, . . .) y *factores ordenados*, en los que si hay un orden natural. En este caso, todas tus variables lo tienen, porque "descienden" de variables continuas.

   Hay otras situaciones en que también hay un orden natural: en las respuestas "Muy de acuerdo", "De acuerdo", "Indiferente", "En desacuerdo" y "Muy en desacuerdo" hay un orden, aunque sea completamente arbitrario asignar valores numéricos[1]. S-PLUS trata de modo diferente los factores y los factores ordenados. Mira la documentación de `glm`, `factor`, `ordered` y `contrasts`, y asegúrate de entenderla. Hay métodos para manejar variables cualitativas ordinales mediante extensiones del modelo de regresión logística que no hemos estudiado (mira por ejemplo la documentación de la función `polr` en la librería MASS de R y en [7], pág. 231).

4. Observa que si una de las variables predictoras separase perfectamente los 0's y 1's de la respuesta (por ejemplo, si la edad o el sexo separase a la perfección los supervivientes y fallecidos en el ejemplo del *Titanic*, desarrollado en clase) la verosimilitud no tendría máximo (¿ves por qué?). Lo mismo ocurre si una combinación lineal cualquiera de los predictores separa a la perfección los valores de la respuesta.

[1] ¿"Muy de acuerdo" es el "doble" de acuerdo que simplemente "De acuerdo"? ¿O quizá tres veces más?.

5. Al analizar los datos de supervivencia en el naufragio del *Titanic*, simples tabulaciones de los datos (mira como emplear la función `table`) son ya suficientemente elocuentes. Observa, no obstante, que no puedes en general tabular Supervivientes frente a Clase y extraer conclusiones como que hubo mucha mayor proporción de ahogados entre los tripulantes: podría ser que entre los tripulantes hubiera también muchas menos mujeres, y que su menor supervivencia reflejara este hecho. Aunque en el ejemplo desarrollado en clase las conclusiones puedan resultar bastante obvias, en general *necesitas un modelo* que deslinde los efectos de los diferentes factores.

   Si empleas R puedes también encontrar de interés la función `mosaicplot`, una herramienta gráfica para examinar tablas de contingencia. Ejecuta,

   ```
   par(ask=TRUE)
   example(mosaicplot)
   ```

   para ver un ejemplo de utilización.

A. Datos `impeach` (detalle votos senatoriales para destituir a Bill Clinton)

Variable	Tipo	Niveles	Descripción
Nombre	Alfanumérica		Nombre abreviado del senador
Estado	Factor	50	Estado (código postal) por el cual es senador
Voto1	Factor	2	Voto sobre acusación perjurio
Voto2	Factor	2	Voto sobre acusación obstrucción justicia
NVotos	Numérica		Número de votos “Culpable” (0, 1 ó 2)
Partido	Factor	2	Partido del senador
Conserv	Numérica		Conservadurismo del senador (de 0 a 100)
ClintonPC	Numérica		% votos demócratas en el estado del senador
Final	Numérica		Año en que finaliza el mandato del senador
Primerizo	Factor	2	¿Es el primer mandato como senador?

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] T.J. Hastie and R.J. Tibshirani. *Generalized Additive Models*. Chapman & Hall, London, 2nd. edition, 1991.

[4] D.W. Hosmer and S. Lemeshow. *Applied Logistic Regression*. Wiley, 1989.

[5] D.G. Kleinbaum. *Logistic Regression. A Self-Learning Test*. Springer Verlag, 1994.

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

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