eman ta zabal zazu

Universidad del País Vasco
Euskal Herriko Unibertsitatea

TAREA 6

EJERCICIOS

1. 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 (1*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?

   d) Genera una vector de 30 observaciones aleatorias $N(0,1)$, y regrésalo sobre las columnas de una matriz arbitraria X de dimensiones 30×15. Repite el experimento 100 veces (¡en batch!), y mira en cuantas ocasiones el mayor $\hat{\beta}_i/\hat{\sigma}_{\hat{\beta}_1}$ supera en valor absoluto $t_{15}^{0,025}$. Explica el resultado.

2. En el fichero `agua.dat` tienes con encabezamientos autoexplicativos datos[1] de consumo de gasóleo (en Kg.), consumo de agua caliente (en m^3), temperatura interior y exterior en grados centígrados, y una variable cualitativa, indicando si en el momento de hacerse la observación existía instalado un aislamiento térmico mediante inyección de una espuma sintética en la cámara de aire de los muros exteriores.

   El depósito de gasóleo es único para atender calefacción y agua caliente central. Hay un contador a la salida del mismo que permite saber el consumo realizado, pero no su distribución entre el quemador de la calefacción y el del agua caliente.

   a) ¿Te parece que el aislamiento ha disminuido, *ceteris paribus*, el consumo de gasóleo, o podría decirse que es ineficaz?

   b) ¿Cuál es el coste en Kg. de gasóleo del metro cúbico de agua caliente?

[1] Ficticios.

AYUDAS, SUGERENCIAS, COMENTARIOS

1. Te resultará de utilidad la función `read.table` para leer los datos. Puedes referirte a las notas repartidas en clase, a [2], y a las indicaciones a continuación; `read.table` es una función moderna y no aparece descrita en [1], aunque sí en [2] y [5] y en la `help` on-line.

2. Al leer `agua.dat` puedes hacerlo así:

   ```
   agua <- read.table(file="/users/practicas/p4ges/agua.dat",header=T)
   ```

   Al hacerlo, estás indicando a S-PLUS (o R) que `agua` es una data-frame, y que deseas que la primera línea del fichero (la cabecera o header) sea utilizada para nombrar las variables, lo que hará tus salidas mucho más auto-documentadas y fáciles de interpretar. Si tecleas,

   ```
   agua
   ```

   veras tus observaciones igual que si `agua` fuera una matriz. Pero observa la diferencia: hay una columna que, en lugar de valores numéricos, tiene datos cualitativos (`SI/NO`).

3. Si quieres puedes emplear la función `lsfit`, pero para ello deberas recodificar los valores `SI/NO` en numéricos (por ejemplo, 0 y 1). Mejor emplear la función `lm`; en este caso, la recodificación de variables (y eventual desdoble en tantas columnas de ceros y unos como sea preciso) es algo automático. En general, `lsfit` es más rápida y permite un mejor acceso a "las tripas" de los cálculos, en tanto `lm` es una función mucho más flexible y cómoda de utilizar; en especial, con datos cualitativos o cuando se quiere mezclar mezclar regresión lineal y no lineal o no paramétrica.

4. Claramente, en el ejemplo del agua caliente el signo del coeficiente estimado asociado a la variable aislamiento tiene importancia: no es creible que el aislamiento empeore las cosas. Si haces una de tus variables dicotómica (SI/NO), S-PLUS la convierte en 0/1 para hacer la regresión: nada garantiza que los SI sean 1 y los NO sean 0 o viceversa, sin embargo, y para interpretar el signo del coeficiente te interesa saber qué codificación se ha empleado. Puedes forzar a `lm` a devolverte la matriz X de diseño que ha construido especificando el argumento opcional `x` (mira la documentación anexa). Esto es muy cómodo: en ocasiones, puede interesarte hacer uso de `lm` sólo para construir una matriz de diseño a partir de variables categóricas, y a continuación emplear sobre dicha matriz `lsfit` o hacer un tratamiento *ad hoc* de la misma.

5. Recuerda: toda inferencia que hagas es relativa a un modelo, válida en la medida en que el modelo sea una adecuada representación de la realidad (qué significa esto, es arduo de dilucidar). Que estás interesado en el efecto del aislamiento no significa que hayas de tomar en tu modelo sólo esta variable en consideración. Tu modelo debe ser "el adecuado", tan simple como sea posible (Ockham) pero no más simple (Einstein). Puede que introduciendo variables adicionales al aislamiento detectes un efecto que de otro modo no hubiera sido significativo (¿ves por qué?).

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

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

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