UNIVERSITAT DE BARCELONA

Universitat Oberta de Catalunya

Regresión, modelos y métodos
Prueba de evaluación continua 1

Francesc Carmona y Mireia Besalú

Fecha publicación del enunciado: 11-04-2020
Fecha límite de entrega de la solución: 26-04-2020

Presentación Esta PEC consta de ejercicios similares a los discutidos en los debates con los que podréis contrastar vuestra asimilación de los conceptos y métodos presentados en las dos primeras unidades.

Objetivos El objetivo de esta PEC es trabajar los conceptos básicos de regresión lineal simple y múltiple trabajados en la primera parte de la asignatura.

Descripción de la PEC Debéis responder cada problema por separado. Recordad que tan importante como el resultado es el razonamiento y el proceso que os leva a ello, es decir el consultor debe poder ver no tan sólo donde habéis llegado sino también como y porque habéis llegado hasta allí. Incluid el código de R en la solución.

Criterios de valoración Cada PEC representa un 50 % de la nota de la asignatura. La presentación de los ejercicios aportará una puntuación que **se sumará** a los puntos obtenidos por las PECs.

Código de honor Cuando presentáis ejercicios individuales os adherís al código de honor de la UOC, con el que os comprometéis a no compartir vuestro trabajo con otros compañeros o a solicitar de su parte que ellos lo hagan. Asimismo aceptáis que, de proceder así, es decir, en caso de copia probada, la calificación total de la PEC será de cero, independientemente del papel (copiado o copiador) o la cantidad (un ejercicio o todos) de copia detectada.

Formato Para hacer la entrega se tiene que enviar un mensaje al buzón de entregas del aula. En este mensaje debéis adjuntar un fichero PDF (obtenido a partir de vuestra solución en Word, Open Office, Latex, Lyx o RMarkdown). El nombre del fichero debe ser la composición de vuestro apellido y vuestro nombre seguido de `_Reg_PEC1.pdf` (por ejemplo: si vuestro nombre es "Jordi Pujol", el fichero debe llamarse `pujol_jordi_Reg_PEC1.pdf`).
Es importante que el examen sea legible, por lo que valoraremos que separéis el código **R** de los resultados y la discusión. Podéis hacerlo por ejemplo dejando el código completo en un apéndice -o en un archivo `.R` adjunto-. En medio de las explicaciones podéis poner vuestro código pero controlad la longitud de las resultados (evitad por ejemplo páginas enteras que únicamente contienen números).

Problema 1 (50 pt.)

Los datos sobre los que trabajaremos provienen de un estudio sobre 18 mujeres y 14 hombres que investigó por qué las mujeres muestran una menor tolerancia al alcohol y desarrollan enfermedades hepáticas relacionadas con el alcohol más fácilmente que los hombres.
El conjunto de datos que tenemos en el archivo `alcohol.txt` contiene las siguientes variables:

- Metabol: El metabolismo de primer paso del alcohol en el estómago (mmol/litro-hora).
- Gastric: La actividad gástrica de la alcohol deshidrogenasa en el estómago (μmol/min/g de tejido)
- Sex: El sexo del sujeto.
- Alcohol: Si el sujeto es alcoholico o no.

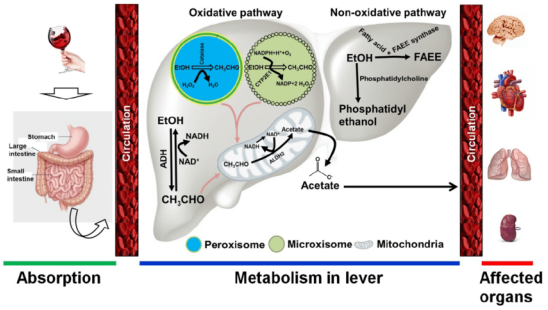

Figura 1: Alcohol metabolism in the body

Como vamos a trabajar con modelos de regresión, es mejor considerar a los factores como variables **numéricas dicotómicas** con valores 0 y 1.
Así crearemos la variable Male que vale 1 para los hombres y 0 en otro caso y la variable Female que vale 1 para las mujeres y 0 en otro caso.
También crearemos la variable Alcoholic tendrá los valores 0 para `Non-alcoholic` y 1 para `Alcoholic`.
Consideramos 4 modelos distintos:

$$\begin{aligned}
\text{Metabol} &= \beta_0^{(1)} + \beta_1^{(1)}\text{Gastric} + \beta_2^{(1)}\text{Female} + \text{error} && (M_1)\\
\text{Metabol} &= \beta_0^{(2)} + \beta_1^{(2)}\text{Gastric} + \beta_2^{(2)}\text{Male} + \text{error} && (M_2)\\
\text{Metabol} &= \beta_0^{(3)} + \beta_1^{(3)}\text{Gastric} + \beta_2^{(3)}\text{Male} + \beta_3^{(3)}\text{Female} + \text{error} && (M_3)\\
\text{Metabol} &= \beta_1^{(4)}\text{Gastric} + \beta_2^{(4)}\text{Male} + \beta_3^{(4)}\text{Female} + \text{error} && (M_4)
\end{aligned}$$

(a) Entre el modelo M_1 y el modelo M_2, está claro que $\beta_2^{(2)} = -\beta_2^{(1)}$. ¿Cual es la relación entre $\beta_0^{(2)}$ y los parámetros del modelo M_1?

(b) ¿Cual es la diferencia en término medio de Metabol entre hombres y mujeres que tienen un mismo nivel alcohol deshidrogenasa?

(c) De los cuatro modelos M_i, $i = 1, \ldots, 4$, ¿cual es el mejor según el coeficiente de determinación? ¿y según el RMSE?

(d) ¿Cual es el rango de la matriz de diseño del modelo M_3? Resolver las ecuaciones normales para este modelo y hallar una estimación alternativa de los parámetros con la ayuda de la g-inversa de Moore-Penrose.

Comprobar que los residuos son los mismos que proporciona R con la función `lm()`.

(e) Hallar el intervalo de confianza de la función paramétrica $\beta_0^{(2)} + \beta_1^{(2)}$ en el modelo M_2.

Nota: Utilizar la fórmula 3.5 del apartado 3.5 del libro de Carmona donde los elementos son los mismos que se necesitan en el test t de Student como en el ejercicio 5.6.

(f) Comparar las rectas de regresión que relacionan el metabolismo Metabol con la actividad gástrica Gastric para hombres y para mujeres. ¿Son paralelas? ¿Son iguales?

Nota: Podemos construir la matriz de diseño conjunta como se explica en el apartado 6.7.1 del libro de Carmona, teniendo en cuenta que el número de mujeres es distinto del número de hombres y el orden de los datos.

(g) Si consideramos el modelo completo con interacciones:

$$\begin{aligned}\mathsf{Metabol} = \beta_0 &+ \beta_1 \mathsf{Gastric} + \beta_2 \mathsf{Male} + \beta_3 \mathsf{Alcoholic} \\ &+ \beta_4 \mathsf{Gastric} \cdot \mathsf{Male} + \beta_5 \mathsf{Gastric} \cdot \mathsf{Alcoholic} + \beta_6 \mathsf{Male} \cdot \mathsf{Alcoholic} \\ &+ \beta_7 \mathsf{Gastric} \cdot \mathsf{Male} \cdot \mathsf{Alcoholic} + \mathsf{error}\end{aligned}$$

¿Podemos prescindir de todas las interacciones y también de la variable `Alcoholic` y quedarnos con el modelo M_2?

Problema 2 (50 pt.)

Para este problema vamos a trabajar con la base de datos `senic.txt` en la que encontraremos datos de 113 hospitales en referencia al *Study on Efficacy of Nosocomial Infection Control* (SENIC). El propósito de este estudio es buscar características de los hospitales asociadas a altos (o bajos) ratios de infecciones adquiridas en hospitales, que se llaman técnicamente infecciones nosocomiales. Las variables incluidas en el archivo son las siguientes:

Variable Number	Variable Name	Description
1	Identification number	1–113
2	Length of stay	Average length of stay of all patients in hospital (in days)
3	Age	Average age of patients (in years)
4	Infection risk	Average estimated probability of acquiring infection in hospital (in percent)
5	Routine culturing ratio	Ratio of number of cultures performed to number of patients without signs or symptoms of hospital-acquired infection, times 100
6	Routine chest X-ray ratio	Ratio of number of X-rays performed to number of patients without signs or symptoms of pneumonia, times 100
7	Number of beds	Average number of beds in hospital during study period
8	Medical school affiliation	1 = Yes, 2 = No
9	Region	Geographic region, where: 1 = NE, 2 = NC, 3 = S, 4 = W
10	Average daily census	Average number of patients in hospital per day during study period
11	Number of nurses	Average number of full-time equivalent registered and licensed practical nurses during study period (number full time plus one half the number part time)
12	Available facilities and services	Percent of 35 potential facilities and services that are provided by the hospital

El estudio SENIC está descrito en una secuencia de artículos[1], pero nosotros utilizaremos el conjunto de datos del Apéndice C1 del libro de Kutner[2] que constan en el archivo `senic.txt`.

(a) Calcular la matriz de correlaciones entre las variables que sea posible. ¿Qué variables son las más correlacionadas con `infrisk`? Añadir algún gráfico adecuado.

Dibujar los gráficos de caja (*boxplot*) para la variable `infrisk` primero separados según la variable `medschl` y después según `region`. ¿El riesgo de infección es igual en todas las regiones? ¿Y según la variable `medschl`?

Nota: Plantear las dos últimas preguntas como un contraste de modelos con la función `anova()`.

(b) Calcular el modelo de regresión que tiene como variable respuesta `infrisk`. Escribir el modelo obtenido. ¿Es significativa la variable `region`? ¿Como interpretas el coeficiente de la variable `medschl`? ¿Y el de `stay`?

(c) Utilizar un test F para determinar la significación de la regresión del modelo. Escribe las hipotesis de este test e interpreta el resultado obtenido. ¿Qué predictoras son significativas al 5 %? ¿Concuerdan estas variables con las del apartado (a)? ¿Cuales son las variables más correlacionadas? ¿Y las que menos? ¿Concuerdan con los resultados del modelo de regresión?

(d) Si simplificamos el modelo tomando únicamente las variables significativas al 5 %, contrastar si se puede aceptar ese modelo simplificado frente al completo.

(e) Consideramos ahora el modelo con las variables: `stay`, `culratio` y `region`. Estudia la normalidad y la heterocedasticidad del error. ¿Hay alguna observación con un alto leverage? ¿Y con una gran influencia? Dibujar los gráficos oportunos para explicar los resultados.

(f) Un amigo americano está a punto de entrar en un hospital. Quiere saber el intervalo de confianza al 90 % para la predicción del riesgo de infección utilizando el modelo del apartado anterior. Sabe que el hospital tiene los valores de `stay` = 9.6 días, `cultratio` = 15.5 y `region` = NE.

(g) En la estimación del modelo del apartado (e), el coeficiente de `region==NE` no aparece. ¿Puedes dar alguna explicación?

Observar las tres últimas columnas de la matriz de diseño del modelo. ¿Como se codifican los 4 valores del factor `region`?

¿Como se calcula *a mano* una predicción para un hospital con `region=="NE"` con la ecuación de este modelo?

Si ejecutamos la siguiente instrucción

```
options(contrasts=c('contr.sum','contr.poly'))
```

y recalculamos la estimación del modelo, ¿como se codifican ahora los 4 valores del factor `region`?

Ahora, ¿como se calcula *a mano* una predicción para un hospital con `region=="NE"` con la ecuación de este otro modelo?

Nota: Si queremos devolver R a su estado por defecto, ejecutaremos:

```
options(contrasts=c('contr.treatment','contr.poly'))
```

(h) Consideremos ahora el modelo con 4 variables regresoras: `stay`, `age`, `xratio` y `medschl`. Alguien sugiere que el efecto de `medschl` sobre el riesgo de infección puede interactuar con `age` y con `xratio`. Añadir los términos de interacción apropiados al modelo de regresión, ajustar el modelo ampliado y contrastar si los términos de interacción ayudan. Usar $\alpha = 0.1$.

Indicar la hipótesis nula, la alternativa, la regla de decisión y la conclusión.

Hospital universitario Johns Hopkins en Baltimore, Maryland, USA.

Referencias

[1] *The American Journal of Epidemiology*, Volume 111, Issue 5, 1980, Pages 465-653

[2] Kutner, Nachtscheim, Neter and Li (2005) *Applied Linear Statistical Models*, 5th Edition.