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: 9-11-2019
Fecha límite de entrega de la solución: 24-11-2019

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.

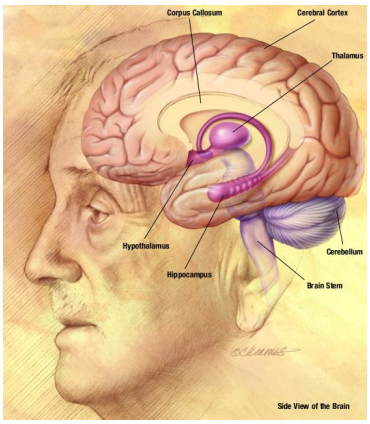
Descripción de la PEC Debéis responder cada problema por separado. Recordad que tan importante como el resultado es el razonamiento y el proceso que os leva a ello, es decir el consultor debe poder ver no tan sólo donde habéis llegado sino también como y porque habéis llegado hasta allí. Incluid el código 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 (30 pt.)



Gladstone (1905)[2] estudió la relación entre el peso del cerebro humano y algunas medidas de la cabeza en personas de ambos sexos y diferentes edades, fallecidas por diversas causas. Los datos se hallan agrupados por sexos y grupos de edad en el documento original y en el archivo `cbrain.dat`, los dos adjuntos.

Las medidas recogidas por Gladstone fueron:

`cause` Causa de la muerte con tres valores 1. "acute" (A), 2. "otherwise" (-) y 3. "chronic" (C).

`age` (A) Edad en años.

`height` (S) Estatura en pulgadas.

`headht` (H) Altura auricular en mm.

`length` (L) Longitud máxima de la cabeza en mm.

`breadth` (B) Anchura máxima de la cabeza en mm.

`size` (P) El producto $H \times L \times B$ en cm^3.

`brnweight` (w) Peso del cerebro en gr.

`circum` (U) Circunferencia horizontal de la cabeza en mm.

`cephalic` Índice cefálico o relación entre la anchura máxima de la cabeza respecto a su longitud máxima multiplicada por 100.

`sex` El sexo con valores 1. F y 2. M.

`ageClass` Tres categorías de edad: 1. "child", 2. "age 20-45" y 3. "over 45".

Estos datos fueron estudiados por Blakeman et al. (1905)[1] desde el punto de vista descriptivo y con diversos modelos de regresión.

(a) Incorporar a **R** esta base de datos, teniendo en cuenta la codificación de los valores faltantes (*missing values*).

Identificar con su número correspondiente las observaciones con algún dato faltante. ¿Cuantas observaciones son?

Eliminar de la base de datos todas las observaciones para las que falta algún dato. Fijaros que conservamos la columna `obs` que permite identificar las observaciones en la base de datos original.

(b) Calcular la variable `new.cephalic` calculando la relación $(B/L) \times 100$ y estudiar la discrepancia con `cephalic` con una recta de regresión.

¿Cual es la correlación entre ambas?

¿Podemos aceptar que la pendiente es 1?

¿Podemos aceptar con un contraste que el coeficiente de intercepción es cero y la pendiente es 1 (las dos cosas a la vez)?

En esta regresión, identificar posibles residuos atípicos o *outliers* (sin hacer ningún contraste) y valorar si el dato de `cephalic` anotado en la base de datos original es una errata.

(c) Estudiar la regresión del peso del cerebro `brnweight` con la variable producto `size`. ¿Veis alguna dificultad?

Eliminar del estudio los individuos de menos de 20 años y repetir la regresión. ¿Mejora?

Estudiar la normalidad del error.

Dado que la variable `size` es un producto de tres variables, proponer una transformación potencia[1] $h(x) = x^\lambda$ que mejore su simetría o "normalidad".

Repetir la regresión con la variable `size` transformada y valorar el modelo.

Hallar el intervalo de confianza para la pendiente al 97 %.

(d) Comparar las rectas de regresión que relacionan el peso del cerebro `brnweight` con la variable transformada $h($`size`$)$ para hombres y para mujeres sin los menores de 20 años. ¿Son paralelas? ¿Son iguales?

Problema 2 (45 pt.)

Con la base de datos `cbrain` del problema anterior sin los menores de 20 años, calcular el modelo de regresión que tiene como respuesta el peso del cerebro `brnweight` y como predictoras las variables `age`, `sex`, `height`, `headht`, `lenght`, `breadth`, $h($`size`$)$ i `circum`. La transformación $h()$ es la que hemos decidido en el apartado (c) del problema anterior.

(a) Escribir el modelo de regresión estimado e interpretar el coeficiente de la variable $h($`size`$)$.

(b) Utilizar un test F para determinar la significación de la regresión del modelo. ¿Qué significa esto último?

¿Qué predictoras son significativas al 5 %?

¿Quiere esto decir que podemos eliminar del model las no significativas al 5 %?

Calcular la matriz de correlaciones entre las variables continuas del modelo de regresión.

¿Cuales son las variables más correlacionadas? ¿Y las que menos? ¿Concuerdan con los resultados del modelo de regresión?

(c) Contrastar si podemos aceptar un modelo más simple sin las variables `headht`, `lenght`, `breadth`.

(d) En el modelo más simple del apartado anterior, ¿hay alguna observación con un alto *leverage*? ¿Y con una gran influencia?

Dibujar los gráficos oportunos para explicar el resultado.

(e) Con el modelo simple del apartado (c), dibujar el gráfico de regresión parcial con la variable predictora $h($`size`$)$.

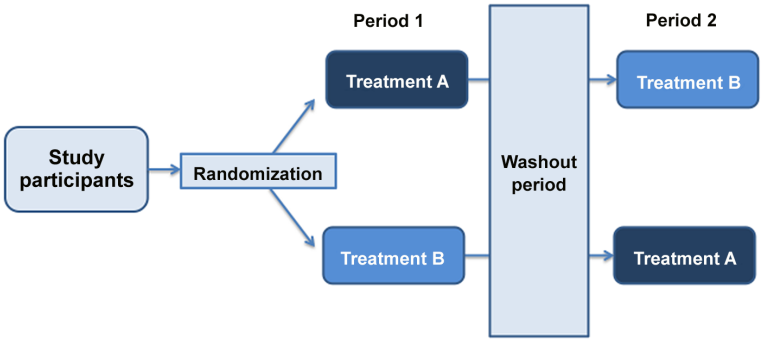
[1] La función `transformTuckey()` del paquete `rcompanion` puede ayudar. Otra posibilidad es aplicar la función `boxcox()` del paquete `MASS` al modelo, pero entonces la transformación se aplica sobre la variable respuesta.

(f) Hallar el intervalo de confianza para la predicción del peso del cerebro, por el modelo simple, cuando un individuo toma los valores de la observación 5 de la base de datos.

(g) Estimar la varianza del error y calcular su intervalo de confianza al 99 % en el modelo de regresión más simple del apartado (c).

Problema 3 (25 pt.)

En los ejemplos 5.3.2 y 5.6.3 del libro de Carmona y con los datos de la tabla 5.2 se ha estudiado el diseño crossover simplificado. En este problema vamos a considerar un diseño un poco más sofisticado y con más parámetros.



En la página `https://newonlinecourses.science.psu.edu/stat509/node/123/` se explican los diseños crossover con todo detalle. Nosotros nos vamos a centrar en el caso 2×2 donde a dos grupos de pacientes se subministran dos fármacos o se realizan dos tratamientos de forma consecutiva en dos períodos de tiempo. En esta situación, los parámetros a considerar son: μ media general, α efecto del fármaco A, β efecto del fármaco B, γ_1 efecto del grupo 1 o secuencia AB, γ_2 efecto del grupo 2 o secuencia BA, π_1 efecto del primer período, π_2 efecto del segundo período, λ_A o efecto de arrastre de A (*first-order carryover effect*) y λ_B o efecto de arrastre de B. En total 9 parámetros.

Como trabajar con 9 parámetros es complicado, en la tabla Design 11 de `https://newonlinecourses.science.psu.edu/stat509/node/127/` nos proponen un conjunto más reducido de la siguiente forma:

[Design 11]	**Period 1**	**Period 2**
Sequence AB	$\mu_A + \nu + \rho$	$\mu_B + \nu - \rho + \lambda_A$
Sequence BA	$\mu_B - \nu + \rho$	$\mu_A - \nu - \rho + \lambda_B$

donde

$\mu_A = \mu + \alpha$ representa el efecto del fármaco A.

$\mu_B = \mu + \beta$ representa el efecto del fármaco B.

$\nu = \gamma_1$ y $-\nu = \gamma_2$ de modo que $\gamma_1 + \gamma_2 = 0$. ν representa el efecto de la secuencia o grupo.

$\rho = \pi_1$ y $-\rho = \pi_2$ de modo que $\pi_1 + \pi_2 = 0$. ρ representa el efecto del período.

λ_A y λ_B son los efectos de arrastre.

(a) Escribir la matriz de diseño reducida del modelo propuesto en la tabla anterior y calcular su rango.

La matriz reducida tendrá 4 filas, una por cada situación experimental, y 6 columnas, una por cada parámetro. El rango representa el número efectivo de parámetros.

(b) Como el objetivo es contrastar si el efecto de los fármacos A y B se puede considerar similar, la hipótesis paramétrica a contrastar es $H_0 : \alpha = \beta$ que es equivalente $H_0 : \mu_A = \mu_B$ con los nuevos parámetros.

Probar que esta hipótesis NO es contrastable con la matriz de diseño del apartado anterior.

(c) Reducir el número de parámetros de forma que $\lambda_A = \lambda_B = \lambda$, es decir, el efecto de arrastre de A es igual al efecto de arrastre de B.

Probar que la hipótesis principal de igualdad de efectos de los fármacos para este nuevo diseño SI es contrastable.

(d) Contrastar la hipótesis $H_0 : \mu_A = \mu_B$ con el modelo del apartado anterior y los datos[2] de la tabla 5.2 del libro de Carmona.

En un cierto orden, los datos de la variable respuesta en **R** son:

```
y <- c(17,34,26,10,19,17,8,16,13,11,
       17,41,26,3,-6,-4,11,16,16,4,
       21,20,11,26,42,28,3,3,16,-10,
       10,24,32,26,52,28,27,28,21,42)
```

Referencias

[1] J. Blakeman, Alice Lee and Karl Pearson (1905), A Study of the Biometric Constants of English Brain-weights, and their Relationships to External Physical Measurements, *Biometrika*, Vol. 4, Issue 1-2, pp 124–160,

[2] R.J. Gladstone (1905). A Study of the Relations of the Brain to to the Size of the Head, *Biometrika*, Vol. 4, pp 105-123

[2] Hay que tener cuidado con el orden de los datos que se debe corresponder con el orden de las filas de la matriz de diseño.