UNIVERSITAT DE BARCELONA

Universitat Oberta de Catalunya

Regresión, modelos y métodos
Prueba de evaluación continua 1

Susana Barcelo, Geòrgia Escaramís, Santiago Ríos y Francesc Carmona

Fecha publicación del enunciado: 20-04-2024
Fecha límite de entrega de la solución: 05-05-2024

Presentación Esta PEC consta de ejercicios similares a los planteados en los ejercicios con los que podréis contrastar vuestra asimilación de los conceptos y métodos presentados en las tres últimas unidades.

Objetivos El objetivo de esta PEC es trabajar los conceptos de regresión 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 lleva a ello, es decir el consultor debe poder ver no tan sólo donde habéis llegado sino también como y porqué 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.
Se valorará positivamente la contención en las respuestas del software y negativamente los volcados de datos innecesarios.

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 "Josep Benet", el fichero debe llamarse `benet_josep_Reg_PEC1.pdf`). También puede ser en formato HTML.

*Es **importante** que el examen sea legible y, a ser posible, elegante. Como si fuera un informe a vuestro jefe. Por ello valoraremos que separéis el código **R** (no necesario para la comprensión de la resolución) de los resultados y la discusión. Podéis hacerlo por ejemplo dejando el código completo en un apéndice. En medio de las explicaciones podéis poner vuestro código pero controlad la longitud de los resultados (evitad por ejemplo páginas enteras que únicamente contienen números).*

Ejercicio 1 (45 pt.)

The Framingham Heart Study (Levy, 1999) recogió datos sobre los factores de riesgo cardiovascular y el seguimiento a largo plazo de casi 5.000 residentes de la ciudad de Framingham (Massachusetts).

La muestra se compone de 4240 individuos, 1944 hombres y 2490 mujeres y 16 variables. Aunque el objetivo del estudio era predecir si el paciente tenía un riesgo a 10 años de sufrir en el futuro una cardiopatía coronaria, en este ejercicio se va a intentar predecir la presión sanguínea sistólica, factor de riesgo cardiovasvular, a partir de algunas variables registradas.

Las variables que se utilizarán en los ejercicios son las siguientes:

- `sysBP` (Systolic blood pressure): Presión arterial sistólica variable Y que se pretende predecir.
- `BMI` (Body mass index): Índice de masa corporal
- `age`: edad
- `male`: sexo (1: varón; 0: mujer)
- `totChol`: colesterol total
- `heartRate`: pulsaciones por minuto
- `currentSmoker`: Fumadxr (1: fumadxr; 0: no fumadxr)
- `cigsPerDay`: cigarrillos por día
- `diabetes`: (0:no; 1:sí)
- `glucose`: glucosa en sangre

(a) Estimar un modelo de regresión `lmod_inicial` que permita obtener la influencia, si existe, únicamente del índice de masa corporal (`BMI`) sobre la presión arterial sistólica (`sysBP`) y si esta relación varía con el sexo (`male`). Comentar el resultado de la regresión, en cuanto a la relación y si ésta varía con el sexo. Comentar también si la bondad de ajuste lineal es suficiente o si es necesario incluir más variables para explicar la varianza de `sysBP`.

Nota: Se recomienda eliminar las observaciones con valores faltantes (*missings*).

(b) Dibujar un gráfico de dispersión con las rectas de regresión de hombres y mujeres según el modelo del apartado anterior.

(c) Hallar los intervalos de confianza al 99 % para los coeficientes del modelo del apartado (a).

Calcular una estimación de la varianza del error en el mismo modelo.

(d) Además de las variables contempladas anteriormente, se cree que hay otras variables clínicas y demográficas que predicen linealmente la presión arterial sistólica como son la edad, el colesterol total, las pulsaciones por minuto, si se es o no fumador y el número de cigarrillos por día, si se es diábetico y la glucosa en sangre. Como se ha visto en los apartados anteriores la relación entre `sysBP` y `BMI` puede depender del sexo.

Estimar un modelo de regresión lineal múltiple `lmod_ampliado` para predecir `sysBP` que tenga como variables predictoras: `BMI`, `age`, sexo, `totChol`, `heartRate`, `currentSmoker`, `cigsPerDay`, `diabetes`, `glucose` y la interacción `BMI` con sexo.

¿Es significativo el modelo obtenido? Plantear la hipótesis nula y la alternativa del test. ¿Qué test estadístico se emplea para contestar a esta pregunta?

Explicar el resultado del coeficiente de la regresora `cigsPerDay` y su significación en el contexto de este modelo.

(e) Contrastar si nos podemos quedar con el modelo más reducido que no tiene en cuenta las variables regresoras: `currentSmoker`, `cigsPerDay` y `diabetes`. Escribir en forma paramétrica las hipótesis del test H_0 y H_1 de este contraste. Estimar un nuevo modelo `lmod_reducido` en el que no intervengan estas variables.

(f) Con el modelo elegido, calcular un intervalo de predicción al 95 % de un individuo con los siguientes valores de la variables predictoras: `BMI`=29, `age`=64, `male`=0, `totChol`=200, `heartRate`=70, `glucose`=96. Comprobar previamente que los valores observados no suponen una extrapolación. Para ello utilizaremos el elipsoide que se forma con el leverage máximo.

Ejercicio 2 (35 pt.)

En este ejercicio se pretende mejorar el modelo del ejercicio anterior haciendo algunas pruebas diagnósticas.

(a) Obtener con la función `plot()` los gráficos básicos de diagnóstico del modelo `lmod_reducido` del apartado (e) del ejercicio anterior. Explicar los cuatro gráficos.

Comprobar con algún test las hipótesis de homocedasticidad y normalidad de los errores. Utilizar dos tests distintos para cada una de las dos hipótesis.

(b) Realizar un estudio descriptivo de la variable `sysBP`, especialmente sobre su distribución. Buscar una transformación de esta variable que mejore sus propiedades.

Sugerencia: Utilizar la transformación $\log(x + a)$ basándose en el modelo reducido obtenido en el ejercicio anterior. Para ello, la función `logtrans()` del paquete `MASS` nos puede ayudar.

(c) Estimar el modelo reducido con la transformación obtenida en el apartado anterior y comprobar si se ha mejorado en la homocedasticidad y en la normalidad.

(d) Estudiar la presencia de valores atípicos, de alto leverage y/o puntos influyentes en el último modelo con la variable respuesta transformada.

Dibujar un gráfico resumen.

(e) Hallar los tres puntos más influyentes del apartado anterior. ¿Son también atípicos (*outliers*)?

Estimar un nuevo modelo sin esos 3 puntos y comprobar otra vez la homocedasticidad y la normalidad.

Ejercicio 3 (20 pt.)

Con el modelo reducido del apartado (e) del ejercicio 2 contestar las siguientes cuestiones:

(a) Hallar la matriz del diseño del modelo. ¿Cual es su rango? ¿Coincide con el número de parámetros β_i? ¿Tendremos problemas para estimar combinaciones lineales de los β_i?

(b) Discutir qué soluciones existen en el sistema de ecuaciones normales cuando el número de parámetros es igual al rango de la matriz de diseño y cuando son diferentes.

(c) Comprobar que los coeficientes obtenidos con las ecuaciones normales son iguales que los obtenidos en el ejercicio 2(e).

(d) Obtener la estimación de la varianza del error σ^2 y un intervalo de confianza al 95 % suponiendo normalidad.

Referencias

[1] The Framingham Heart Study. Publisher/Source: The NHLBI Biologic Specimen and Data Repository Information Coordinating Center (BioLINCC).

[2] https://en.wikipedia.org/wiki/Framingham_Heart_Study