Regresión, modelos y métodos
Prueba de evaluación continua 2

Geòrgia Escaramís, Susana Barcelo, Víctor Gallardo, Santiago Ríos y Francesc Carmona

Fecha publicación del enunciado: 17-06-2023
Fecha límite de entrega de la solución: 2-07-2023

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 segunda 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.

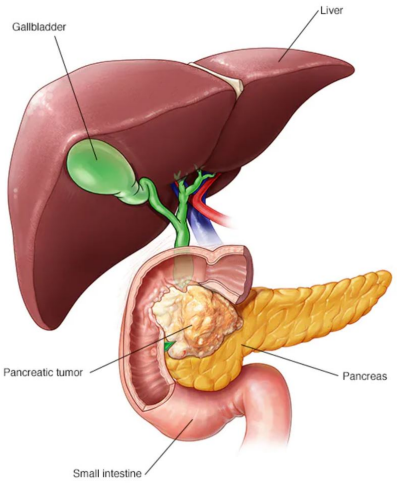
Formato Para hacer la entrega se tiene que enviar un mensaje al buzón de entregas del aula. En este mensaje debéis adjuntar un fichero PDF (obtenido a partir de vuestra solución en Word, Open Office, LaTeX, LyX o RMarkdown). El nombre del fichero debe ser la composición de vuestro apellido y vuestro nombre seguido de `_Reg_PEC2.pdf` (por ejemplo: si vuestro nombre es "Mireia García", el fichero debe llamarse `garcia_mireia_Reg_PEC2.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 (35 pt.)

El archivo `pancreas_biomarkers.txt` contiene el conjunto de datos del trabajo de Silvana Debernardi et al (2020), donde los investigadores estudian el uso de un panel de biomarcadores urinarios para el diagnóstico precoz del tipo más común de cáncer de páncreas, llamado *adenocarcinoma ductal pancreático* o PDAC. Para ello, recolectaron tres grupos de pacientes (variable `diagnosis` en el conjunto de datos):

- Controles saludables (valor 1)
- Pacientes con afecciones pancreáticas no cancerosas, como pancreatitis crónica (valor 2)
- Pacientes con adenocarcinoma ductal pancreático (valor 3)



En la medida de lo posible, estos pacientes fueron emparejados (*matched*) por edad y sexo. Para cada paciente se dispone de cuatro biomarcadores urinarios: creatinina, LYVE1, REG1B y TFF1.

creatinina una proteína que a menudo se usa como indicador de la función renal.

LYVE1 el receptor 1 de hialuronano endotelial de los vasos linfáticos, una proteína que puede desempeñar un papel en la metástasis tumoral.

REG1B una proteína que puede estar asociada con la regeneración del páncreas.

TFF1 el factor de trébol 1, que puede estar relacionado con la regeneración y reparación del tracto urinario.

Los datos también incluyen la edad, el sexo e información sobre el estadio del cáncer de páncreas y el diagnóstico para pacientes no cancerosos.

(a) Ajustar un modelo de regresión logística que incluya la edad categorizada y los cuatro potenciales biomarcadores urinarios, con la finalidad de diagnosticar o discriminar pacientes con cáncer frente a pacientes con afecciones pancreáticas no cancerosas. ¿Qué variables nos permiten predecir el riesgo de adenocarcinoma ductal pancreático?

(b) Interpretar los coeficientes que acompañan a edad categorizada y LYVE1.

(c) Contrastar si nos podemos quedar con un modelo más reducido que no incluya los biomarcadores creatinina y TFF1. Escribir las hipótesis H_0 y H_1 de este contraste.

(d) Verificar la suposición de aumento lineal en el *log odds* del modelo seleccionado en el apartado anterior para los biomarcadores LYVE1 y REG1B. Usar un modelo que agrega el cuadrado LYVE1^2, y un modelo que agrega el cuadrado REG1B^2, uno a la vez.

¿Deberíamos incluir cualquiera de las variables como una función cuadrática? Discutir vuestra conclusión aportando los estadísticos que la soportan.

(e) Suponemos que tenemos un paciente con afección pancreática, aunque no sabemos si es cancerosa o no, con las siguientes características: edad=68, creatinine=0.5, LYVE1=6, REG1B=140, TFF1=400. Según el modelo seleccionado en el apartado (c) para el diagnóstico de cáncer de páncreas, ¿lo clasificaríamos como afección cancerosa o no cancerosa? Argumentar la respuesta evaluando la posible extrapolación de la observación.

Ejercicio 2 (40 pt.)

Diversos estudios han demostrado que los casos de enfermedad de Parkinson han aumentado en los últimos 30 años. Una medida mayoritariamente aceptada en la comunidad científica para valorar la gravedad de la enfermedad de Parkinson en una persona es su puntuación en la Escala Unificada de Enfermedad de Parkinson (UPDRS). Así, un método que pueda predecir la puntuación UPDRS de un paciente de Parkinson será eficaz para determinar la gravedad de la afección del paciente. En este ejercicio trataremos de pronosticar con técnicas de regresión la puntuación UPDRS de un paciente a partir de otros valores de parámetros independientes que se sabe que afectan a dicha puntuación UPDRS.
El conjunto de datos que vamos a estudiar está compuesto por un rango de mediciones de voz biomédicas de 42 personas con enfermedad de Parkinson en etapa temprana reclutadas para un ensayo de seis meses de un dispositivo de telemonitorización para el seguimiento remoto de la progresión de los síntomas. Las grabaciones fueron capturadas automáticamente en los hogares del paciente.
Estos datos se hallan en el Machine Learning Repository:

`http://archive.ics.uci.edu/ml/datasets/Parkinsons+Telemonitoring`

Las columnas del archivo contienen el número del sujeto, la edad, el sexo, el tiempo desde la fecha de reclutamiento, el UPDRS motor, el UPDRS total y 16 medidas de voz biomédicas. Cada fila corresponde a una de las 5875 gravaciones de voz de estos individuos.

El principal objetivo es predecir el UPDRS total a partir de las 16 medidas de voz.

Los datos se pueden cargar en una sesión de **R** con las siguientes instrucciones[1]:

```
import.data <-
"http://archive.ics.uci.edu/ml/machine-learning-databases/parkinsons
                              /telemonitoring/parkinsons_updrs.data"
parkinson <- read.table(url(import.data), sep=",", skip=1)
names(parkinson) <- c("subject#","age","sex","test_time","motor_UPDRS","total_UPDRS",
  "Jitter(%)","Jitter(Abs)","Jitter:RAP","Jitter:PPQ5","Jitter:DDP",
  "Shimmer","Shimmer(dB)","Shimmer:APQ3","Shimmer:APQ5","Shimmer:APQ11",
  "Shimmer:DDA","NHR","HNR","RPDE","DFA","PPE")
set.seed(123)
```

Observemos que hemos fijado la semilla para que los cálculos pseudo-aleatorios sean fijos.
A continuación separamos la base de datos en dos grupos: el grupo `data.train` con el 80 % de las observaciones y el grupo `data.test` con el resto. La base de datos `data.train` servirá para calcular los diferentes modelos y la base de datos `data.test` para comprobar el ajuste a un grupo externo de datos.

Para evitar problemas es mejor que suprimamos de entrada las observaciones con valores perdidos, si los hay.

Utilizar la puntuación `total_UPDRS` como variable respuesta y las 16 medidas de voz como potenciales regresoras. Se trata de estudiar el mejor modelo por diferentes métodos. En cada caso se informará del número de variables (o componentes) que se utilitzan, el coeficiente R^2 ajustado (cuando se pueda) y el *root mean squared error* (RMSE) para el grupo de ajuste (*train*) y para el grupo de prueba (*test*).

Ajustar los siguientes modelos:

(a) Regresión lineal con todas las 16 variables predictoras.

Indicar los posibles problemas, que no consideramos, al prescindir del factor "sujeto".

(b) Regresión lineal con las variables seleccionadas paso a paso por AIC.

Nota: Para no mostrar todos los pasos, utilizar el parámetro `trace = F`.

[1] La dirección web debe estar en una única linea

(c) Regresión por componentes principales.

Nota: Utilizar el mínimo de componentes razonable a la vista del gráfico de RMSE.

(d) *Ridge regression.*

(e) ¿Cree necesario repetir estos métodos tomando la variable `motor_UPDRS` como respuesta?

(f) Con la base de datos `data.train` y el modelo OLS hacer un rápido análisis de los residuos para detectar incumplimientos de las condiciones de un modelo lineal. Estudiar especialmente si hay problemas de multicolinealidad.

Nota: Para calcular el coeficiente de determinación R^2 en algunos métodos que no lo dan explícitamente, hay que hacerlo a partir de su definición:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

Cuando el número de variables es grande, también podemos calcular el coeficiente R^2_{adj}

$$R^2_{adj} = 1 - \frac{(1 - R^2)(n-1)}{n - p - 1}$$

donde p es el número de variables explicativas.

Ejercicio 3 (25 puntos)

Seguimos con la base de datos del ejercicio anterior. En ese ejercicio hemos visto que el modelo OLS presenta algunas dificultades. Para superarlas podemos estudiar otros métodos.

(a) Eliminar de la base de datos los 3 puntos más influyentes y volver a calcular los modelos del ejercicio 2. Mostrar en una tabla los RMSE del grupo de prueba (*test*) para cada uno de los modelos con y sin los puntos influyentes.

Nota: En este apartado sólo hay que mostrar la tabla.

(b) Dado que el grupo de prueba (*test*) puede contener outliers, calcular un RMSE robusto para cada modelo utilizando la media recortada (*trimmed*) al 10 %. Añadir esta información a la tabla del apartado anterior.

(c) Dados los problemas observados con los residuos del modelo OLS, podemos probar un método robusto como el de Huber o el *Least trimmed squares* (LTS).

Referencias

[1] Silvana Debernardi, Harrison O'Brien, Asma S. Algahmdi, Nuria Malats, Grant D. Stewart, Marija Pljesa-Ercegovac, Eithne Costello, William Greenhalf, Amina Saad, Rhiannon Roberts, Alexander Ney, Stephen P. Pereira, Hemant M. Kocher, Stephen Duffy, Oleg Blyuss, Tatjana Crnogorac-Jurcevic (2020), *A combination of urinary biomarker panel and PancRISK score for earlier detection of pancreatic cancer: A case-control study*, **Plos Med**, Dec 10;17(12):e1003489.