Actividad 3: Modelización predictiva

2019.1

Noviembre de 2019

Índice

1. Modelo de regresión lineal **1**
- 1.1. Modelo de regresión lineal simple . . . 2
- 1.2. Modelo de regresión lineal múltiple (regresores cuantitativos) . . . 2
- 1.3. Modelo de regresión lineal múltiple (regresores cuantitativos y cualitativos) . . . 2
- 1.4. Efectuar una predicción de la concentración de hematocritos en los dos modelos . . . 2

2. Modelo de regresión logística **2**
- 2.1. Análisis crudo. Estimación de OR (Odds Ratio) . . . 2
- 2.2. Modelo de regresión logística . . . 3
- 2.3. Mejora del modelo . . . 3
- 2.4. Predicción . . . 3
- 2.5. Conclusiones . . . 3

En esta actividad se usará el fichero de datos: **datosA3.csv**, que contiene los datos sobre 2353 operaciones efectuadas en el Hospital Universitario de Santiago. Nuestro motivo es estudiar primero la relación entre diferentes variables de la base de datos y posteriormente la identificación de los factores de riesgo asociados a la infección post operatoria.

El conjunto de datos contine 2353 registros y 15 variables:

- EDAD (años)
- SEXO
- PATOL (Patología) 1=inflamatoria; 2=neoplasia;3=trauma; 4=otras.
- TIP_OPER (tipo operación): 1=limpia; 2=potencialente contaminada; 3=contaminada; 4=sucia
- ALB (albúmina)
- HB (Hemoglobina)
- HCTO (Hematocrito)
- LEUCOS (Leucocitos)
- LINFOPCT (Linfocitos (%))
- HEMAT (Hematíes)
- GLUC (Glucosa)
- OBES (Obesidad)
- DESNUTR (Desnutrición)
- DIABETES
- INFEC(Infección)
- GLUC_4 (categorización de glucosa)

Se estudiarán las posibles relaciones entre las variables que se indican a continuación.

Nota: importante a tener en cuenta para entregar la actividad: Es necesario entregar el fichero Rmd y el fichero de salida (PDF o html). El fichero de salida debe incluir: el código y el resultado de la ejecución del mismo (paso a paso). Se debe respetar la misma numeración de los apartados que el enunciado.

---

1. Modelo de regresión lineal

1.1. Modelo de regresión lineal simple

a) Estimar por mínimos cuadrados ordinarios un modelo lineal que explique la variable hematocrito en función de la hemoglobina. Evaluar la bondad de ajuste a través del coeficiente de determinación (R2). Podéis usar la instrucción de R lm.

b) Algunos estudios afirman que la relación calculada anteriormente varía según la persona esté en condiciones óptimas de salud o no. Para contestar a esta pregunta, se dividirá la muestra en dos, según si la persona presenta desnutrición o no. Posteriormente se repetirá el estudio para cada muestra por separado. A partir de los resultados del modelo lineal en cada una de las muestras, ¿se puede tomar como cierta dicha conclusión? Justificar la respuesta.

1.2. Modelo de regresión lineal múltiple (regresores cuantitativos)

Estimar por mínimos cuadrados ordinarios un modelo lineal que explique la variable hematocrito en función de la hemoglobina y la edad.

Evaluar la bondad del ajuste y comparar el resultado con el obtenido en el apartado 1.1.a). Podéis usar la instrucción de R lm y usar el coeficiente R-cuadrado ajustado en la comparación. Interpretar también el significado de los coeficientes obtenidos y su significación estadística.

1.3. Modelo de regresión lineal múltiple (regresores cuantitativos y cualitativos)

a) Queremos conocer en qué medida se relacionan los hematocritos, con la hemoglobina y la edad, dependiendo de si los pacientes tienen o no infección postquirúrgica. Aplicar un modelo de regresión lineal múltiple y explicar el resultado.

b) Se hará el mismo estudio, pero tomando sólo aquéllos pacientes, cuya cantidad de hematocritos sea $<$ 37. Comparar con el modelo anterior y extraer conclusiones.

1.4. Efectuar una predicción de la concentración de hematocritos en los dos modelos

Suponer un paciente de 60 años, con infección postquirúrgica y con un valor de hemoglobina de 10. Realizar la predicción del valor de hematocritos, con los dos modelos del apartado 1.3. Interpretar los resultados.

2. Modelo de regresión logística

2.1. Análisis crudo. Estimación de OR (Odds Ratio)

Se desea identificar cuáles son los factores de riesgo en la infección postquirúrgica. Por tanto, se evaluará la probabilidad de que un paciente pueda o no tener una infección, dependiendo si presenta o no unas determinadas características.

Para evaluar esta probabilidad, primero se realizará un análisis crudo de los posibles factores (características). Es decir, un análisis univariante de posibles factores de riesgo asociados a la infección postquirúrgica.

a) Estudiar la relación entre la infección postquirúrgica y cada una de las variables siguientes: diabetes, desnutrición, obesidad, edad y hematocrito. Estimar e interpretar las OR en cada caso. Dicha estimación será efectuada a partir de las tablas de contingencia. Antes de calcular los valores de las odds ratio, se recomienda aplicar el test chi-cuadrado, para valorar la relación entre las variables. Para el test Chi-cuadrado, podéis consultar: https://psicologiaymente.com/miscelanea/prueba-chi-cuadrado. Para calcular odds ratio, podéis consultar: https://www.r-bloggers.com/computing-odds-ratios-in-r/

b) Edad y hematocrito son variables continuas: ¿podríamos seguir el procedimiento anterior para el cálculo de la OR?

c) Si queremos ver la relación entre INFEC (Infección) y TIP_OPER (tipo de operación), ¿podríamos seguir el procedimiento anterior, para el cáculo de la OR? En el caso que la respuesta fuese negativa, ¿cuál sería una solución?

2.2. Modelo de regresión logística

a) Estimar el modelo de regresión logística donde la variable dependiente es "INFEC" y la explicativa es tener diabetes o no. ¿Podemos considerar que el hecho de tener diabetes es un factor de riesgo de infección? Justifica tu respuesta. Tiene relación con lo obtenido en el apartado anterior? Se recodificará la variable DIABETES en 0=NO y 1=SI.

b) Añadimos al modelo anterior las variables explicativas edad y hematocrito. Evaluar si alguno de los regresores tiene influencia significativa (p-valor del contraste individual inferior al 5 %).

2.3. Mejora del modelo

a) Entrenamos el mismo modelo anterior, pero categorizando ambas variables continuas: Edad: (edad>=65 y edad<65) y Hematocrito: (hb <37 y hb>=37). Explicar los resultados. ¿De qué forma influye la edad y los niveles de hematocritos en este modelo? Explicar como se interpretan los resultados del modelo.

b) Posteriormente se añadirá al modelo las variable explicativa desnutrición. ¿Se observa una mejora del modelo? Explicar.

2.4. Predicción

Según el modelo del apartado anterior, ¿cuál será la probabilidad de infección postquirúrgica de un paciente de 50 años, con diabetes, concentración de hematocritos de 34, y que no presente desnutrición?

2.5. Conclusiones

En este apartado deberéis exponer, de las variables explicativas estudiadas, cuáles pueden considerarse factores de riesgo en la infección postquirúrgica. Razonar en base a los resultados obtenidos.