Aprendizaje Estadístico - FIUBA
Guía 4 - Clasificación

1.

La cifosis es una deformación en la columna vertebral que puede presentarse en niños con cirugía correctiva de la espina dorsal. Se cree que la incidencia de deformaciones (**kyp**=1 si la deformación está presente, **kyp**=0 caso contrario) después de la cirugía de la columna vertebral está asociada a la edad en la que ésta se realiza (**Age**, en meses), a la primera vértebra operada (**Start**) y al número de vértebras involucradas en la operación (**Number**).

El conjunto de datos kyphosis.csv contiene la información correspondiente a 81 niños.

1. Graficar la respuesta binaria de la incidencia de la cifosis versus la edad del niño. Ajustar un modeo logístico simple usando la variable edad como regresora. Examinar el ajuste y la significación de **Age**.

2. Ajustar un modelo de regresión logística usando como covariables **Age**, **Start** y **Number**. Examinar el ajuste y la significación de todos las componentes del modelo.

3. Ajustar un modelo de regresión logística cómo en el ítem a) pero agregando un término cuadrático en **Age**. Examinar el ajuste y la significación de todos las componentes del modelo.

4. Ajustar un modelo de regresión logística usando como covariables **Age**, **Start**, **Number** y una componente cuadrática en **Age**. Examinar el ajuste y la significación de todos las componentes del modelo. Comprarar con el modelo del ítem anterior.

5. ¿Es más conveniente considerar un modelo que contenga una componente cuadática en Number y una interacción entre **Age** y **Number**? Justificar.

---

2. Un investigador está interesado en saber como las variables GRE (Graduate Record Exam scores), GPA (grade point average) y prestigio de la institutción de pregrado, influyen en la admisión a una escuela de grado. La variable de respuesta es la admisión o no del candidato.

La variable de respuesta de este conjunto de datos es la respuesta binaria **admit**, mientras las variables regresoras son: **gre**, **gpa** y **rank**. Trataremos a las variables gre y gpa como continua. La variable rank toma los valores de 1 a 4, donde 1 corresponde a las instituciones de mayor prestigio y 4 a las de menor.

`https://stats.idre.ucla.edu/stat/data/binary.csv`

1. Realizar un análisis exploratorio de las variables presentes.

2. Ajustar un modelo logístico usando todas las variables explicativas. (Tener presente la naturaleza de la variable **rank**. Transformar: rank = factor(rank)).

   De acuerdo, a este ajuste ¿ en cuánto aumenta el log odds de la admisión cuando la variable **gre** aumenta en una unidad? ¿ Cómo se interpretan las estimaciones de los coeficientes relacionados con la variable **rank**?

3. ¿Cuánto vale la probabilidad estimada de la admisión para los distintos niveles de la variable **rank** cuando los otros dos predictores toman como valor la media muestral? En consecuencia, ¿ Cuál es el valor predicho de la admisión para cada caso?

4. ¿Cuáles son los valores estimados del cociente de los odds asociados cuando cada variable aumenta en una unidad y el resto permanece constante? Hallar intervalos de confianza de nivel aproximado 0.95 para cada uno de ellos, ¿ alguno contiene al 1?

5. Realizar un análisis secuencial de salida mediante el comando anova de R con la opción test="Chisq".¿Cómo se obtendría cada uno de estos resultados? ¿Cómo se interpretan cada uno de los resultados que arroja esta función?

6. Comparar el resultado anterior con los que obtendría al realizar

```
# modelo solo intercept
mod1 <- glm(admit ~ 1,  data = datos, family = "binomial")
# modelo intercept + gre
mod2 <- glm(admit ~ gre, data = datos, family = "binomial")
# modelo intercept + gre + gpa
mod3 <- glm(admit ~ gre + gpa,data = datos, family = "binomial")
# modelo completo
mod4 <- glm(admit ~ gre + gpa + rank, data = datos, family = "binomial")

anova(mod1, mod2, test="LRT")
anova(mod2, mod3, test="LRT")
anova(mod3, mod4, test="LRT")
```

   ¿ Cambia algo en estos resultados de anova si se usa test="Chisq"?

7. Mediante la instrucción anova compare los tres modelos posibles con solo dos variables (gre+gpa, gre+rank, gpa+rank) con el modelo completo.

---

3. Consideremos los datos **iris**. Es un conjunto de datos analizados por Fisher que consisten en 4 mediciones realizadas en 50 flores iris de cada una de 3 especies distintas (Setosa, Versicolor y Virginica). Las 4 variables, medidas en centímetros, son

X1 = Longitud de los sépalos (sepal length)
X2 = Ancho de los sépalos (sepal width)
X3 = Longitud de los pétalos (petal length)
X4 = Ancho de los pétalos (petal width)

1. Utilizando las funciones del paquete `MASS` de R:

   a) Suponiendo normalidad en los datos, hacer una clasificación lineal y calcular el error de clasificación por los 2 métodos:

      1) Calcular la proporción de datos mal clasificados (o error aparente).
      2) Validación cruzada.

   b) Idem a) pero con la clasificación cuadrática y comparar los resultados.

2. Usando lo aprendido en teoría

   a) Graficar los pares de datos (X2,X4) en el plano. Para cada especie, estos datos ¿tienen aspecto de provenir de una distribución normal bivariada?

   b) Suponiendo que la distribución es normal bivariada para cada población, construir la regla de clasificación cuadrática, asumiendo probabilidad a priori de pertenecer a cada grupo iguales. Usando esta regla de clasificación recién construida clasificar la nueva observación $x_0 = (3.5, 1.75)$ como perteneciente a alguno de los 3 grupos.

   c) Supongamos que las matrices de covarianza Σ_i son las mismas para las 3 poblaciones normales bivariadas. Construir la regla de clasificación lineal, asumiendo probabilidad a priori de pertenecer a cada grupo iguales, y usarla para clasificar la nueva observación $x_0 = (3.5, 1.75)$ como perteneciente a alguno de los 3 grupos. Comparar los resultados obtenidos en los ítems anteriores.

d) Graficar en el scatterplot realizado en a) las regiones halladas en c).

e) Usando la clasificación lineal realizada en c), clasificar las observaciones de la muestra. Calcular la proporción de datos mal clasificados y la estimación insesgada del error que se obtiene por validación cruzada.

---

4. Aproximadamente 2 años antes de la bancarrota de algunas empresas se recolectan datos financieros de las mismas, y también se recolectan datos de empresas sanas financieramente alrededor del mismo momento. A continuación figuran las 4 variables correspondientes a los datos que se encuentran en el archivo finanzas:

X1 = (flujo de caja)/(deuda total)
X2 = (ingreso neto)/(total de activos)
X3 = (activos corrientes)/(pasivos corrientes)
X4 = (activos corrientes)/(ventas netas)

Grupo 1: Empresas en bancarrota
Grupo 2: Empresas sanas financieramente

a Graficar los datos para los pares de observaciones (X1, X2), (X1, X3) y (X1, X4). Para alguno de estos pares de variables, ¿tienen aspecto de provenir de una distribución normal bivariada?

b Usando los $n_1 = 21$ pares de observaciones de empresas en bancarrota y los $n_2 = 25$ pares de observaciones de empresas sanas financieramente, calcular los vectores de medias muestrales $\bar{x_1}$ y $\bar{x_2}$ y las matrices de covarianza muestrales S_1 y S_2.

c Usando los resultados de b) y asumiendo que las dos muestras aleatorias provienen de dos poblaciones normales, construir la regla de clasificación cuadrática asumiendo $\pi_1 = \pi_2$.

d Evaluar la performance de la regla de clasificación desarrollada en c) calculando el error aparente total y la estimación del error actual esperado que se obtiene por validación cruzada.

e Repetir los items c) y d) tomando $\pi_1 = 0.05$ y $\pi_2 = 0.95$. ¿Es razonable esta elección de probabilidades a priori?

f Usando los resultados de b), construir la matriz de covarianza ponderada y realizar el análisis de coordenadas discriminantes. Usar esta función para clasificar las observaciones muestrales y evaluar el error aparente.

g Repetir los items b) a e) usando ahora las variables (X1, X3) y luego las variables (X1,X4). ¿Parecen ser algunas variables mejores clasificadoras que otras?

h Repetir los items b) a e) usando las 4 variables.