Ejercicios Tema 3. Diagnóstico y validación en Regresión lineal múltiple.

Máster en Ciencia de Datos. Módulo: Análisis exploratorio de datos

Ana Navarro Quiles

Curso 2022/2023

Objetivos:

Se pretende practicar los conceptos relacionados con la regresión lineal multiple: diágnostico del modelo, validación/validación cruzada y técnicas Bootstrap.

Nuevos comandos útiles

Diagnostico y validación en regresión lineal múltiple

Como ejemplo aplicación nos basamos en el banco de datos **advertising**, y en el modelo predictivo para la variable *sales*, que es la variable respuesta.

```
## DIAGNÓSTICO

# Grafico general de diagnóstico
par(mfrow=c(2,2))
plot(ml1)

# Calculo de residuos estandarizados y studentizados
rstandard(ml1)
rstudent(ml1)

# Grafico de residuos vs fitted
plot(fitted(ml1),rstudent(ml1))
abline(h=0, col="gray")

# Loess de residuos vs fitted
 library(ggplot2)
 library(gridExtra)
  ggplot(data = advertising, aes(x =fitted(ml1), y = rstudent(ml1))) +
  geom_point() + geom_smooth(color = "coral",span=0.4) + geom_hline(yintercept = 0) +
  labs(y = "residuos studentizados",
       x = "valores ajustados") +
  theme_bw()

# Test de Breusche-Pagan

    library(lmtest)
```

```
    bptest(ml1)

# Transformaciones de box-cox

    library(MASS)
    boxcox(ml1, lambda = seq(0, 1, length = 10))

# Normalidad y outliers: qqplot

  qqnorm(rstudent(ml1))
  qqline(rstudent(ml1),col="red")

# Test de normalidad

  shapiro.test(rstudent(ml1))

# Resumen numérico del vector de residuos

  summary(rstudent(ml1))
  advertising[abs(rstudent(ml1)) > 5,]

# Medidas de influencia: Leverages  y distancia de Cook

  library(car)
  influencePlot(ml1)
  summary(influence.measures(ml1))

# Diagnostico leverages

  n<-nrow(advertising)
  plot(fitted(ml1),hatvalues(ml1),main="leverages vs fitted")
  abline(h=2*length(coef(ml1))/n,col="red",lwd=1);
  abline(h=3*length(coef(ml1))/n,col="red",lwd=3);

  boxplot(hatvalues(ml1))
  summary(hatvalues(ml1))

# Diagnostico distancia de Cooks

  plot(fitted(ml1),cooks.distance(ml1),main="Distancia de Cook vs fitted")
  abline(h=1,col="red",lwd=1);
  #identify(fitted(ml1),cooks.distance(ml1))

  boxplot(cooks.distance(ml1))
  summary(cooks.distance(ml1))

# Colinealidad

library(ISLR)
vif(ml1)  #valores menores que 10 no hay problema

library(GGally)
ggpairs(advertising[,c('sales', 'TV','radio','newspaper')], lower = list(continuous = "smooth"),
```

```
        diag = list(continuous = "barDiag"), axisLabels = "none")

library(corrplot)
corrplot(cor(advertising[c('TV','radio','newspaper')]), method = "number", tl.col = "black")

## VALIDACIÓN

# Reserva de 1/4  datos para validar.

set.seed(12345)
seleccion <- sample(nrow(advertising),round(nrow(advertising)*3/4))
entrenamiento <- advertising[seleccion,]
prueba <- advertising[-seleccion,]

ajuste <- lm(sales ~ TV+newspaper+radio,data=entrenamiento)
summary(ajuste)
prediccion <- predict(ajuste,prueba)
(ecm <- mean((prueba$sales-prediccion)^2)) # referencia: mean(residuals(ml1)^2)

# Validación cruzada: leave-one-out

library(boot)
datos <- subset(advertising,select = TV:sales)
glm1 <- glm(sales ~., data=datos)
ecm <- cv.glm(advertising, glm1)
ecm$delta # referencia: mean(residuals(ml1)^2)

# Validación cruzada: leave-one-out  k-grupos

library(boot)
datos <- subset(advertising,select = TV:sales)
glm1 <- glm(sales ~., data=datos)
ecm <- cv.glm(advertising, glm1, K=10)
ecm$delta # referencia: mean(residuals(ml1)^2)

# SE bootstrap (no cumplimiento de hipótesis)

library(boot)

         # Error estándar en un punto dado (Utilizando el modelo lineal)
x <- c(150,23,30)
predict(ajuste, newdata = data.frame(TV=x[1],newspaper=x[2],radio=x[3]),
        se.fit=TRUE, interval = 'confidence')

         # Error estándar en un punto dado (Utilizando bootstrap)
B <- 1000
boot.fun <- function(datos,indice,x) {
  coef <- coef(lm(sales ~ TV+newspaper+radio,data=datos,subset=indice))
  return(coef[1]+coef[2]*x[1]+coef[3]*x[2]+coef[4]*x[3])
}
set.seed(12345)
boot(advertising,boot.fun,B,x=x)
```

Ejercicios propuestos

Ejercicio 1

Los sistemas de entrega de productos son de vital importancia para las empresas. En particular, les suele interesar predecir el *tiempo* necesario para realizar los pedidos. Supongamos que la persona responsable de analizar los datos a cargo de una empresa sólo tiene acceso rápido a información sobre la distancia y el número de cajas que ha de distribuirse en cada pedido. En el fichero **cervezas** tenemos unos datos que representan las tres variables nombradas.

a) ¿Cuál es el porcentaje de varianza explicada por tu modelo?¿Qué variables son relevantes?

b) Diagnostica el modelo ¿Qué observas?¿ Puedes mejorar tu modelo solucionando el o los problemas observados?

c) Puedes realizar una predicción para un nuevo reparto que consiste en llevar 20 cajas a 40 km de distancia y dar su error de predicción. ¿Y si hay que llevarlas a 70km?

Ejercicio 2

En los procesos de producción, hay bastante confusión sobre cuáles son las partes del proceso que hacen que ocurran desviaciones del resultado final que se está buscando. Existen diferentes factores que pueden influir: la temperatura del proceso de producción, la densidad del producto, o la propia tasa de producción. El fichero **defectos** contiene información sobre el número medio de defectos (en cada lote analizado) encontrados en 30 pruebas, junto con el valor de las covariables antes comentadas.

a) Ajusta un modelo para predecir el número medio de defectos con la información disponible, diagnostica el modelo y evalúa la efectividad de posibles soluciones.

b) Ajusta un modelo con el que se pretende explicar la relación entre la temperatura y el número medio de defectos, con la información disponible, diagnostica el modelo y evalúa la efectividad de las posibles soluciones.

Ejercicio 3

En el fichero **puentes** se ajustó un modelo para predecir el tiempo que se tarda en diseñar un puente (variable *Time*) en base al número de planos estructurales, variable *Dwgs*, y el número de tramos, variable *Spans*, categorizada en dos niveles (hasta tres tramos y superior). Diagnostica el modelo y si hubiera algún problema toma las acciones oportunas para intentar solucionarlo.

Ejercicio 4

En el fichero **deportistas** ajusta un modelo multivariante con toda la información disponible para predecir el porcentaje de grasa *PrctGRASA*, utilizando un procedimiento automático de selección de variables. Diagnostica el modelo y si hubiera algún problema toma las acciones oportunas para intentar solucionarlo. La función `recode` de la librería `car` puede serte útil.

Ejercicio 5

El fichero *magazines.csv* contiene los datos de cuatro variables de interés:

- *AdRevenue*: los ingresos por publicidad (en miles de dólares EE.UU.)
- *AdPages*: el número de páginas con publicidad de pago
- *SubRevenue*: ingresos por suscripciones de pago (en miles de dólares EE.UU.)
- *NewsRevenue*: ingresos por ventas en quiosco (en miles de dólares EE.UU.)

El interés del estudio se centra en construir un modelo de regresión múltiple que explique los ingresos por publicidad en función de las otras tres variables. Realiza una selección del modelo que contenga el número de

covariables necesarias y que intente garantizar sus condiciones de aplicabilidad. Observa la matriz de gráficas de dispersión de las variables. ¿Qué te sugieren? Observa que la variable respuesta y las tres variables de predicción son bastante asimétricas.

Ejercicio 6

En el fichero **silicio.sav** están los datos del contenido de silicio en muestras de agua de mar recogida a ciertas distancias prefijadas de la costa. Se trata de estudiar la relación lineal entre *silicio* con las otras variables, y predecir el contenido de silicio en el agua en función de la distancia a la costa. Para ello:

- Representa el diagrama de dispersión.
- Ajusta una recta de regresión a los datos.
- ¿Qué variación se obtiene en el contenido de silicio por cada Km de alejamiento de la costa?
- ¿Qué porcentaje de variación en el contenido de silicio es explicado por la regresión?
- Después de ver la gráfica de los residuos tipificados sobre los valores pronosticados tipificados, ¿qué comentarías?
- Ajusta un nuevo modelo que resuelva los problemas observados.
- Obtén un valor pronosticado para el contenido de silicio a 12 y a 40 Km de la costa.
- ¿Qué contenido de silicio se estimaría a 70 Km de la costa?