Ejercicios Tema 1. Regresión lineal simple

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

Ana Navarro Quiles

Curso 2022/2023

Objetivos:

Con estos ejercicios se pretende practicar los conceptos vistos en el Tema 1: Regresión Lineal Simple

- Cálculo e interpretración de la recta de regresión lineal.
- Coeficiente de determinación y coeficiente de correlación de Pearson.
- Modelo de regresió lineal simple (intervalos de confianza, contrastes de hipótesis, predicciones).
- Condiciones de validez del modelo.
- Transformaciones y alternativas no lineales.

Comandos útiles

A continuación indicamos algunos comandos útiles que complementan aquellos que ya hemos visto en la teoría. Para la realización de los ejercicios se recomienda revisar los comandos vistos durante el Tema 1.

Regresión lineal

Como ejemplo de aplicación vamos utilizar el banco de datos **deportistas** (que podemos encontrar en **datosTema1.RData**), empleando la variable suma de pliegues, *SumPliegues*, para explicar el porcentaje de grasa, *PrctGrasa*, que es la variable respuesta.

En particular, vamos a realizar la regresión lineal para el subconjunto de los hombres, *male*.

```
load('datosTema1.Rdata')
```

```
#Selección subconjunto:
hombres <- subset(deportistas, Genero=='male')

#Ajuste modelo
lm_hombres <- lm(PrctGrasa ~ SumPliegues, data=hombres)

#Resumen Modelo
summary(lm_hombres)

#Cálculo de residuos
residuos<-residuals(lm_hombres)

#Representación del ajuste
plot(hombres$SumPliegues, hombres$PrctGrasa, col='BLUE')
abline(coef=coef(lm_hombres), col='RED')
```

```
#Extracción de coeficientes y sus Intervalos de Confianza
coef(lm_hombres)
confint(lm_hombres)

#Obtención de bandas de estimación:
minx<-range(hombres$SumPliegues)[1]; maxx<-range(hombres$SumPliegues)[2]
nuevos <- data.frame(list(SumPliegues = seq(minx,maxx,length=100)))
bandas_est<-predict(lm_hombres, newdata = nuevos, interval = "confidence")

#Representación gráfica:
plot(hombres$SumPliegues, hombres$PrctGrasa, col='BLUE')
abline(coef=coef(lm_hombres), col='RED')
lines(nuevos$SumPliegues,bandas_est[,2],col='BLACK')
lines(nuevos$SumPliegues,bandas_est[,3],col='BLACK')

#predicción  x0=100:
predict100<-predict(lm_hombres, newdata = data.frame(SumPliegues=c(100)),
                    interval = "prediction")

#Obtención de bandas de predicción
minx<-range(hombres$SumPliegues)[1]; maxx<-range(hombres$SumPliegues)[2]
nuevos <- data.frame(list(SumPliegues = seq(minx,maxx,length=100)))
bandas_pred<-predict(lm_hombres, newdata = nuevos, interval = "prediction")

#Representación gráfica:
plot(hombres$SumPliegues, hombres$PrctGrasa, col='BLUE')
abline(coef=coef(lm_hombres), col='RED')
lines(nuevos$SumPliegues,bandas_pred[,2],col='BLACK')
lines(nuevos$SumPliegues,bandas_pred[,3],col='BLACK')

# diagnóstico linealidad y homocedasticidad
residuos <- residuals(lm_hombres)
predichos <- fitted.values(lm_hombres)
par(mfcol=c(1,2))
plot(predichos,residuos, col='BLUE',main = 'Gráfica de residuos')
abline(h=0,lty=2)

# diagnóstico normalidad residuos
qqnorm(residuos, col='BLUE')
qqline(residuos)
```

Alternativas no lineales. Modelos más flexibles

Como ejemplo de aplicación usamos los datos de la base **Boston** que está en el paquete de R **MASS**. Son datos sobre los suburbios de Boston, con variables como el precio medio de la vivienda *medv*, que vamos a utilizar como variable respuesta, y el estatus de la población *lstat* que vamos a utilizar como predictora.

```
library(MASS)
attach(Boston)
```

```
#ajuste knn vecinos
library(FNN)  #libreria
xx <- seq(min(lstat),max(lstat),0.25) # puntos (ordenados)
new_lstat <- data.frame(list(lstat = xx))
reg_knn_80 <- knn.reg(lstat, new_lstat, y=medv, k=5) #ajuste para k=5

#representación gráfica knn
plot(lstat, medv,cex=.4)
lines(xx,reg_knn_80$pred,col='GREEN',lwd=2)

#ajuste loess
xx <- seq(min(lstat),max(lstat),0.25) # puntos (ordenados)
new_lstat <- data.frame(list(lstat = xx))
reg_loess_20 <- loess(medv ~ lstat, span = 0.20)  #ajuste
pred_loess_20 <- predict(reg_loess_20, newdata = new_lstat, se = T) #prediccion

#representación gráfica loess
plot(lstat, medv,cex=.4)
lines(xx,pred_loess_20$fit,col='BLUE',lwd=2)
```

Ejercicios propuestos

Ejercicio 1

Utilizando el banco de datos **deportistas**, considerad la variable respuesta *Peso* relacionandola con el predictor *PrctGrasa*

a) ¿Cuánto vale la pendiente de la recta? ¿Podemos afirmar que es positiva?

b) Compara la varianza de la variable respuesta con la varianza de los residuos: ¿Qué porcentaje de la variabilidad inicial está explicado por la recta de mínimos cuadrados? ¿Qué porcentaje de la variabilidad inicial falta todavía por explicar?

c) Obtén los intervalos de confianza al 95% sobre los parámetros de la recta.

d) Dibuja el diagrama de dispersión, la recta de regresión y las bandas de confianza para la estimación al 95%.

e) Si te parece adecuado estima el peso correspondiente a nuevos individuos con los siguientes porcentajes de grasa: $25, 50, 75\%$. Calcula sus respectivos intervalos de confianza al 95%.

Ejercicio 2

Repite el ejercicio anterior considerando IMC en lugar de peso y compara los resultados con los del ejercicio anterior.

Ejercicio 3

Utilizando el banco de datos **deportistas.cs**, considerad la variable respuesta *PrctGrasa* relacionándola con el predictor *MCMagra*.

a) Obtén la recta mínimos cuadrados utilizando todos los datos, sin tener en cuenta el *sexo*.

b) Evalua el efecto del *sexo* sobre *PrctGrasa*.

c) Obtén ahora una recta para los *hombres* y otra para las *mujeres*.

d) Dibuja en la misma gráfica las tres rectas y comenta los resultados.

Ejercicio 4

En la base **deportistas**,

a) Evalúa mediante regresión lineal si el *PrctGrasa* explica los resultados análiticos: *Hematocrito*, *Ferritina* y *Hemoglobina*.

b) Evalúa mediante regresión lineal si *Peso* y *Altura* explican el *PrctGrasa*.

c) Evalúa la relación entre *IMC* y las variables *SumPliegues* y *PrctGrasa*.

Ejercicio 5

Utilizando el modelo $Y = 25 + 2X + \epsilon$, siendo ϵ Normal con media 0 varianza $\sigma^2 = 4$, simula $N = 10000$ muestras de tamaño $n = 50$. Para ello, utiliza valores de X simulados de una Uniforme definida en el intervalo $(0, 5)$. A continuación, para cada una de las muestras simuladas, obtén el intervalo de confianza al 95% sobre la pendiente de la recta. ¿Qué porcentaje de intervalos no contienen al verdadero valor de la pendiente?

Vuelve a calcular ese porcentaje, pero ahora simulando ϵ de forma que $\epsilon/8$ sea t-Student con 4 grados de libertad. ¿Cómo afecta la falta de normalidad a la fiabilidad de ese intervalo?

Ejercicio 6

Utilizando el banco de datos **Auto**, en el paquete de R **ISLR**, se desea explicar el consumo de carburante, variable *mpg*, a partir de la potencia del motor, variable *horsepower*.

a) Dibuja el diagrama de dispersión y la recta de mínimos cuadrados.

b) ¿Hay relación entre esas dos variables? ¿Cómo de fuerte es esa relación? ¿Podemos afirmar si es positiva o negativa?

c) ¿Qué consumo se espera si potencia del motor es 75? Proporciona el intervalo de confianza y el de predicción para esa potencia de motor.

d) Analiza gráficamente los residuos y comenta los resultados.

Ejercicio 7

El banco de datos **cerebros** es un banco de datos famoso. En él se recogen los pesos del cuerpo y del cerebro de diversos animales. Vamos a explicar el peso del cerebro (en g) *cerebro* a partir del peso del cuerpo (en Kg) *cuerpo*.

a) Ajusta el modelo y realiza el diagnóstico del modelo.

b) Retira los datos que no pertenecen a la misma población que el resto y re-analiza.

Ejercicio 8

Utilizando los datos de mamíferos, del banco **cerebros**, y las variables en escala logarítmica, dibuja el diagrama de puntos con la recta de mínimos cuadrados. A continuación, analiza gráficamente los residuos ¿crees que el modelo lineal sería adecuado?

Suponiendo adecuado el modelo lineal, contesta a las siguientes preguntas:

a) ¿Cuánto vale la pendiente de la recta? ¿Podemos afirmar que es positiva?

b) Compara la varianza de la variable respuesta con la varianza de los residuos: ¿Qué porcentaje de la variabilidad inicial está explicado por la recta de mínimos cuadrados? ¿Qué porcentaje de la variabilidad inicial falta todavía por explicar?

c) Obtén los intervalos de confianza al 90% sobre los parámetros de la recta.

d) Estima el valor de la recta de regresión en el punto *log(cuerpo) = 3* y calcula su intervalo de confianza al 95%. Dibuja el diagrama de dispersión, la recta de regresión y las bandas de confianza al 95% sobre la estimación de la recta.

e) Obtén la predicción puntual y por intervalos (al 95%) de un nuevo mamífero con *log(cuerpo) = 6*. Añade a la gráfica anterior las bandas de predicción.

Ejercicio 9

El banco de datos **Advertising.csv** relaciona las ventas de ciertos productos con la inversión en publicidad, considerando diversos medios: televisión, radio y periódicos. Aquí vamos a estudiar la variable respuesta *sales* relacionándola con el predictor TV.

a) Obtén un ajuste mediante el método *KNN*, decidiendo el valor de k que consideres adecuado.

b) Obtén un ajuste mediante el método *loess*, decidiendo el valor de *span* que consideres adecuado.

c) Dibuja, en la misma gráfica, los dos ajustes anteriores junto con la recta de mínimos cuadrados