Ejercicios Tema 2. 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:

- Ajuste e interpretración de la recta de regresión lineal múltiple.
- Utilidad del modelo y de los predictores.
- Interacción entre predictores y regresión polinómica.
- Criterios de comparación y selección de modelos.

Comandos útiles

Regresión lineal múltiple

Como ejemplo aplicación nos basamos en el banco de datos **Deportistas**, usando como respuesta el porcentaje de grasa, *PrctGrasa*.

```
load('datosTema2.Rdata')
```

```
#Ajuste modelo con todos los posibles predictores
lm_completo <- lm(PrctGrasa ~ ., data=deportistas, na.action=na.exclude)

#Actualización de modelos

lm_sinAltura<-update(lm_completo,~.-Altura)
lm_sinRec<-update(lm_completo,~.-RecGR-RecGB)

#Regresión polinómica (polinomio de grado dos)
lm_parábola <- lm(PrctGrasa ~ MCMagra+I(MCMagra^2),data=deportistas, na.action=na.exclude)

#Regrasión polinómica (haciendo uso de polinomios ortogonales)
lm_parábola2 <- lm(PrctGrasa ~ poly(MCMagra, 2),data=deportistas, na.action=na.exclude)

#Interacción
lm_intera<-lm(PrctGrasa ~ MCMagra*Genero,data=deportistas, na.action=na.exclude)

#Test F parcial
anova(lm_completo,lm_sinRec,test="F")
# Como el p-valor es 0.6723>0.05, no rechazamos la hipótesis nula. Por lo que
# los coeficientes eliminados del modelo completo en el modelo sinRec (RecGR y RecGB)
# no son significativos.
```

```r
#AIC
AIC(lm_sinAltura)
AIC(lm_sinRec)

#Selección por regsubsets (del mejor subconjunto)
library(leaps)
ajuste.sel <- regsubsets(PrctGrasa ~. , data=deportistas, nvmax=20)
(resumen <- summary(ajuste.sel))

resultado <- cbind(resumen$rsq,resumen$adjr2,resumen$cp,resumen$bic)
colnames(resultado) <- c('Rsq','RsqAdj','Cp','BIC')

which.max(resumen$adjr2)
which.min(resumen$cp)
which.min(resumen$bic)

#  Selección stepwise
ajuste.sel2 <- step(lm_completo, direction = 'both',trace=0) #El argumento
# trace indica la información que se muestra.


# la siguientes opciones en R, permite ver visualmente la relación entre las variables
# seleccionadas, por lo que es útil en la búsqueda de confusores:
ggpairs(deportistas[,c('Ferritina', 'IMC', 'Genero')],
        lower = list(continuous = "smooth"), diag = list(continuous = "barDiag"),
        axisLabels = "none")

corrplot(cor(deportistas[,c('Ferritina', 'IMC', 'Altura')]), method = "number", tl.col = "black")
```

Ejercicios propuestos

Ejercicio 1

Antes de que comience la construcción de un puente se pasa por una serie de etapas de producción, entre las que destaca su diseño. Esta fase se compone a su vez de varias actividades, por lo que suele ser de interés la predicción del tiempo de diseño a nivel de planificación presupuestaria. En el fichero **puentes** hay información sobre los proyectos de construcción de 45 puentes. A partir de dicha información trata de valorar el tiempo *Time* que se tarda en diseñar un puente en base a:

- Superficie de cubierta de puente (en miles de pies cuadrados), variable *DArea*
- Coste de construcción (en miles de dólares), variable *CCost*
- Número de planos estructurales, variable *DWGS*
- Longitud del puente (en pies), variable *Length*
- Número de tramos, variable *Spans*

Realiza el análisis indicando con todo detalle las características del modelo que vayas a emplear, las suposiciones que has de hacer y la validez de tus conclusiones. Con el modelo elegido responde a las siguientes preguntas

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

b) ¿Cuál será el tiempo estimado según tu modelo para la construcción de un puente con los predictores en su valor promedio?¿Y cuál sería el intervalo de confianza para el promedio de tiempo predicho? ¿Y si se trata de un nuevo puente?

c) Uno de los constructores indica que, en su experiencia, se tarda lo mismo en construir un puente de 1,2

o 3 tramos, y algo más en construir puentes de más de tres tramos ¿Podrías construir un modelo de regresión para comprobar la hipótesis del constructor ?¿Te parece acertada dicha hipótesis en función de la bondad de ajuste?

Ejercicio 2

En el banco de datos **diabetes**, que contiene datos sobre la mortalidad por dicha enfermedad se pretende estudiar el efecto del hábito tabáquico *TABAC* sobre la edad de diagnóstico de la diabetes *EDATDIAG* . Justifica la elección de variables explicativas de entre las disponibles:

- Mortalidad por diabetes, variable *MORT*
- Tiempo de vida en meses tras el diagnóstico, variable *TEMPSVIU*
- Edad del paciente, variable *EDAT*
- Índice de masa corporal, variable *BMI*
- Resultado del electrocardiograma, variable *ECG*
- Antecedentes coronarios, variable *CHD*
- Presión arterial sistólica y diastólica, variables *SBP* y *DBP*, respectivamente

a) Ajusta un modelo simple para contestar la pregunta de investigación. Indica la bondad del ajuste e interpreta el efecto.

b) Los resultados del modelo anterior sugieren alguna simplificación de la variable explicativa? Si es así realizala.

c) Valora qué variables de la base de datos deberían ser consideradas como potenciales confusores y evalúa la posible confusión causada por cada una de ellas. ¿Cuál es tu modelo final?

d) Caso de ser lícito considerar la variable *EDAT* como potencial confusor, analiza su efecto.

Ejercicio 3

Usando la base de datos **deportistas**, valora e interpreta la existencia de interacción entre *MCMagra* y *Genero* en la explicación de la variable *PrctGrasa*

Ejercicio 4

En el banco de datos **Boston** del paquete de R MASS, que contiene datos sobre los suburbios de Boston, queremos analizar el precio medio de la vivienda *medv* respecto del estatus de la población *lstat*. Para ello:

a) Ajusta una recta de regresión a los datos y representa el ajuste ¿Qué comentarías?

b) Ajusta un modelo parabólico directamente y mediante polinomios ortogonales, compara numérica y gráficamente los dos modelos entre sí y con el modelo lineal (comandos `anova` y `AIC`)

c) ¿Mejoraría el modelo con un polinomio de orden superior? Inténtalo usando el comando poly y representa el ajuste del modelo polinómico elegido.

Ejercicio 5

Las organizaciones profesionales de contables, ingenieros, etc., realizan encuestas regularmente entre sus miembros para conseguir información relativa a los salarios, las pensiones y las condiciones de empleo. Uno de los resultados de estas encuestas es la llamada curva de salario, que relaciona el sueldo con los años de experiencia. La curva salarial muestra el salario "típico' ' de los profesionales con un determinado número de años de experiencia. Es de gran interés para los miembros de la profesión, pues les gusta saber dónde están situados entre sus pares. También es útil para los departamentos de personal de las empresas, para realizar ajustes de sueldos o para contratar a nuevos profesionales. Modeliza la curva salarial, con los datos de la base **salarios**.

Ejercicio 6

Usando la base **diabetes**,

a) Ajusta el mejor modelo posible para predecir la edad al diagnóstico, usando el comando `regsubset` y basándote en el criterio de mínimo Akaike.

b) Describe el modelo que has seleccionado.

Ejercicio 7

En la base **deportistas**, pretendemos ajustar el mejor modelo predictivo del porcentaje de grasa, usando las variables disponibles. Elige entre los dos métodos `regsubsets` y `step` ¿Cuál has elegido y por qué?

Ejercicio 8

Usando la base de datos **Puentes**, ajusta el mejor modelo predictivo del coste de construcción *CCost*. ¿Qué variable de las disponibles tiene mayor capacidad predictiva?