Estadística
Facultad de Ingeniería – Maestría en Explotación de Datos y Gestión del Conocimiento
Universidad Austral

Regresión Lineal

Contenido

Ejercicio 1

Francis Galton, un científico del SXIX, comparó las alturas de los padres con las de sus hijos en pulgadas. Los datos de este estudio se encuentran en el paquete "UsingR", como el objeto 'galton'.

Tener en cuenta que para cargar los datos hay que descargar, instalar y cargar en memoria el paquete “UsingR”. Para hacer eso pueden correr las siguientes líneas de código:

```
install.packages("UsingR")

library(UsingR)

data(galton)
```

En base a esos datos, se pide lo siguiente:

a) Usando los datos de 'galton'[1], encontrar la media, el desvío estándar y la correlación entre las alturas de los padres y sus hijos.
b) Centrar las variables 'parent' (padres) y 'child' (hijos), y verificar que las medias son 0 (aclaración: tener en cuenta que el resultado de la media puede no dar 0 exacto debido al error de redondeo que cometen todas las máquinas, lo que se busca observar es que las medias sean 0, o en su defecto, lo más cercanas a 0 posible). Redondear los resultados con dos posiciones decimales.
c) Normalizar la altura de padres e hijos. Verificar que las variables normalizadas tienen media 0, desvío 1 y ver la correlación.

Ejercicio 2

Cargar en memoria los datos 'father.son', en base a esos datos se pide lo siguiente:

a) Obtener el modelo linear en el que la altura del hijo (sheight) es la variable resultado, y la altura del padre (fheight) es el predictor.
b) Obtener el intercepto y la pendiente de dicho modelo.
c) Graficar y superponer la línea de regresión estimada. (Nota: se pueden usar los gráficos del paquete “ggplot2”)
d) En base al modelo del ejercicio anterior, centrar las variables de padre e hijo, estimar un nuevo modelo sin tener en cuenta el intercepto y verificar si la estimación de la pendiente es la misma que en el modelo del ejercicio anterior.
e) Tener en cuenta que el coeficiente principal es igual a la sumatoria de x*y, dividido la sumatoria de x^2
f) Normalizar los datos de padre e hijo y ver si la pendiente del modelo normalizado es igual a la correlación.
g) Utilizando el modelo estimado en el ejercicio 2.1.1, predecir cuál va a ser la altura del hijo, si la altura del padre fuera de 63 pulgadas.
h) Volviendo al modelo del punto 2.1.1, graficar la altura del hijo en el eje horizontal, y en el eje vertical los residuos.

---

[1] *Nota: la función data() permite cargar en memoria datos, que por ejemplo se encuentren dentro de un paquete*.

i) Hallar el R cuadrado para este modelo.

Ejercicio 3

Considere un conjunto de datos en el que el desvío estándar de la variable resultado es el doble que el desvío estándar del predictor, y la correlación de las variables=0.3. ¿Cuál es el valor estimado de la pendiente de dicho modelo?

a) Considere el ejercicio anterior, si la variable resultado tiene media =1, y el predictor tiene media =0.5, ¿cuál va a ser el valor del intercepto?
b) Volviendo a la pregunta "a": ¿cuál sería la pendiente estimada si la variable resultado fuera el predictor, y viceversa?

Ejercicio 4

Responder si la siguiente afirmación es verdadera o falsa y justificar: Si la variable predictora tiene media 0, el intercepto estimado mediante una regresión lineal, ¿es la media del resultado?

Ejercicio 5

Volver a estimar un modelo lineal con la altura del hijo como variable resultado y la altura del padre como predictor. Obtener el p valor, y plantear el test de hipótesis correspondiente.

a) Interpretar ambos parámetros, centrando el intercepto.
b) Predecir la altura del hijo si la altura del padre fuera de 80 pulgadas, explicar por qué recomendaría o no esta predicción.

Ejercicio 6

Cargar los datos 'mtcars'. Estimar un modelo de regresión lineal con las millas por galón (mpg) como variable salida, y los caballos de fuerza (hp) como predictor.

a) Graficar un scatterplot del nuevo modelo[2].
b) Testear la hipótesis de no linealidad entre los caballos de fuerza y las millas por galón.
c) Predecir las millas por galón que se van a consumir si los caballos de fuerza fueran =111.

Ejercicio 7

La tabla "States.csv", presenta información de cada uno de los Estados de EEUU, dentro de la misma, la columna "msat" indica la puntuación promedio de cada estado en los

[2] Nota: Una forma de graficar scatterplots es usando ggplot2

exámenes SAT de matemática, y la columna "expense" indica los gastos por alumno en el año. Luego de cargar los datos en memoria se pide lo siguiente:

a) Estimar un modelo que exprese la nota del SAT de matemática en función de los gastos por alumno.

b) Verificar si dicha estimación cumple con los supuestos de Gauss Markov.

Ejercicio 8

(ejercicio adaptado del libro: Introducción a la econometría. Un enfoque moderno, J. M. Woolridge, 4ta ed., capítulo 2.

El archivo ceosal2.csv contiene los siguientes datos:

1. salary: 1990 salario, $1000s
2. age: edad en años
3. college: =1 si tiene educación universitaria
4. grad: =1 si tiene educación terciaria
5. comten: antigüedad en la empresa
6. ceoten: años como CEO de la empresa
7. sales: ventas de la empresa del año 1990, en millones
8. profits: ganancias de la empresa del año 1990, en millones
9. mktval: valores de mercado, fines de 1990, en millones.
10. lsalary: log(salary)
11. lsales: log(sales)
12. lmktval: log(mktval)
13. comtensq: comten^2
14. ceotensq: ceoten^2
15. profmarg: ganancias como % de las ventas

Se pide:

a) Hallar la antigüedad dentro de la empresa promedio, y el salario promedio.

b) Hallar el número de CEOs que llevan entre un año y cinco en el cargo.

c) Estimar un modelo de regresión simple que exprese el logaritmo del salario en función de la antigüedad en el cargo de ceo, e indique la variación en el salario ante un aumento de un año en el cargo.

d) Estimar un modelo de regresión lineal simple en el cual se expresen los años de antigüedad en el cargo de CEO en función de la antigüedad en la empresa.

Ejercicio 9

(ejercicio adaptado del libro: Introducción a la econometría. Un enfoque moderno, J. M. Woolridge, 4ta ed., capítulo 2)

Utilice los datos de "sleep75.csv"[3] para analizar si existe una relación inversa entre las horas de sueño por semana y las horas de trabajo pagado por semana. Cualquiera de las variables puede usarse como la variable dependiente. Estime el modelo:

$$sleep = b_0 + b_1 totwork + \varepsilon_t$$

Donde sleep corresponde a minutos de sueño por semana durante la noche y totwrk corresponde al total de minutos de trabajo por semana.

a) Dé sus resultados en forma de ecuación, además de la cantidad de observaciones y la R^2. ¿Qué significa el intercepto de la ecuación?
b) Si totwrk aumenta 2 horas, ¿cuánto se estima que disminuirá sleep?

Ejercicio 10

(ejercicio adaptado del libro: Introducción a la econometría. Un enfoque moderno, J. M. Woolridge, 4ta ed., capítulo 2)

Use la base de datos "wage2.csv" para estimar una regresión simple que explique el salario mensual (wage) en términos de la puntuación del coeficiente intelectual (IQ).

a) Determine el promedio muestral del salario y de IQ. ¿Cuál es la desviación estándar muestral de IQ? (La puntuación del coeficiente intelectual está estandarizada, de manera que el promedio de la población es 100 y la desviación estándar es 15.)
b) Ahora, estime un modelo en el que cada aumento de un punto en IQ tenga un mismo efecto porcentual sobre wage. Si IQ aumenta 15 puntos, ¿cuál es el aumento porcentual pronosticado para wage?

[3] Nota: el archivo usa como separador ";", con read.csv2() se soluciona.