Estadística no paramétrica & Bayesiana

VMO

19 de mayo 2019

Contents

1. Ejercicio ([1] Introducción)

1. Teniendo en cuenta los datos de Maggie Simpson, responde las siguientes preguntas:

a. ¿Cuál fue su puntuación media?

b. ¿Cuáles fueron el primer y tercer cuartil de sus puntuaciones?

c. De acuerdo con la prueba de rango con signo de Wilcoxon de una muestra, ¿son sus puntajes significativamente diferentes de un puntaje neutral de 3?

d. ¿Es útil el resultado del intervalo de confianza de la prueba para responder la pregunta anterior?

e. En general, ¿cómo resumiría sus resultados? Asegúrese de abordar la implicación práctica de sus puntuaciones en comparación con una puntuación neutral de 3.

f. ¿Estos resultados reflejan lo que usted esperaría de ver el gráfico de barras?

```
Input =("
  Speaker          Rater  Likert
 'Maggie Simpson'   1         3
 'Maggie Simpson'   2         4
 'Maggie Simpson'   3         5
 'Maggie Simpson'   4         4
 'Maggie Simpson'   5         4
 'Maggie Simpson'   6         4
 'Maggie Simpson'   7         4
 'Maggie Simpson'   8         3
 'Maggie Simpson'   9         2
 'Maggie Simpson'  10         5
")

datos <- read.table(textConnection(Input),header=TRUE)
```

2. Brian Griffin quiere evaluar el nivel de educación de los estudiantes en su curso de escritura creativa para adultos. Quiere saber el nivel de educación medio de su clase, y si el nivel de educación de su clase es diferente del nivel de licenciatura típico.

Brian usó la siguiente tabla para codificar sus datos.

Instructor	Student	Education
Brian Griffin	a	3

Instructor	Student	Education
Brian Griffin	b	2
Brian Griffin	c	3
Brian Griffin	d	3
Brian Griffin	e	3
Brian Griffin	f	3
Brian Griffin	g	4
Brian Griffin	h	5
Brian Griffin	i	3
Brian Griffin	j	4
Brian Griffin	k	3
Brian Griffin	l	2

Para cada uno de los siguientes, responda la pregunta y muestre el resultado de los análisis que usó para responder la pregunta.

a. ¿Cuál era el nivel medio de educación? (¡Asegúrese de informar el nivel de educación, no solo el código numérico!)

b. ¿Cuáles fueron el primer y tercer cuartil para el nivel educativo?

c. De acuerdo con la prueba de Wilcoxon de una muestra, ¿son los niveles de educación significativamente diferentes de los niveles de un bachiller típico?

d. ¿Es útil el resultado del intervalo de confianza de la prueba para responder la pregunta anterior?

e. En general, ¿cómo resumiría los resultados? Asegúrese de abordar las implicaciones prácticas.

f. Grafica los datos de Brian tal que te ayude a visualizar los datos.

g. ¿Los resultados reflejan lo que usted esperaría de mirar el gráfico?

2. Ejercicio ([4] NonLinearLS)

Considere la siguiente ecuación:

$$Y = \frac{\epsilon}{1 + e^{\beta_1 X_1 + \beta_2 X_2}}$$

$$log(Y) = -log(1 + e^{\beta_1 X_1 + \beta_2 X_2}) + log(\epsilon)$$

La segunda ecuación es la versión transformada que usaremos para la estimación.

- Usando como semilla =1, simule 100 valores de X_1 y X_2 con distribución uniforme entre 0 y 1.
- Usando el modelo propuesto, simule valores de Y donde $a = 0.8$ y $b = 0.5$.
- El ruido tiene distribución normal con media cero y varianza 1.

Aplicando estimación de mínimos cuadrados no lineales paraestimar el modelo con los datos simulados. Debe obtener los siguientes resultados:

```
##
## Formula: log(Y) ~ -log(1 + exp(a * X1 + b * X2))
##
## Parameters:
##   Estimate Std. Error t value Pr(>|t|)
## a   0.8470     0.2697   3.141  0.00223 **
```

```
## b   0.5962     0.2718   2.193  0.03065 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6517 on 98 degrees of freedom
##
## Number of iterations to convergence: 3
## Achieved convergence tolerance: 2.299e-06
```

3. Ejercicio ([6] GAM & [7] MARS & [10] RegBayes)

Usando los datos (`dataRegExp.dta`) que se encuentran en DataLectures, debes estimar el modelo:

$$ly2 = female + age + age2 + eyears + e_pos + white + black + casado + public$$

Donde

ly2: logaritmo del ingreso female: 1 si es mujer age: edad age2: edad al cuadrado eyears: años de educación e_pos: 1 si tiene postgrado white: blanco black: indígena, negro o mulato casado: si está casado (hombre o mujer) public: si trabaja en el sector público.

Para la estimación recuerda filtrar los datos para asalariados (`datos$pe28 >=6500 & datos$pe28 <= 9900`).

El modelo de regresión lineal múltiple, debe arrojar los siguientes resultados:

```
##
## Call:
## lm(formula = ly2 ~ female + age + age2 + eyears + e_pos + white +
##     black + casado + public, data = datos)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.9132 -0.3295 -0.0404  0.3086  2.2487
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.414e-01  9.270e-02   2.605 0.009244 **
## female      -8.205e-02  1.983e-02  -4.138 3.61e-05 ***
## age          2.052e-02  4.670e-03   4.393 1.16e-05 ***
## age2        -1.095e-04  5.447e-05  -2.010 0.044501 *
## eyears       6.779e-02  2.518e-03  26.920  < 2e-16 ***
## e_pos        2.937e-01  5.546e-02   5.297 1.27e-07 ***
## white        1.188e-01  3.269e-02   3.632 0.000286 ***
## black       -1.088e-01  5.154e-02  -2.110 0.034925 *
## casado       1.299e-01  2.026e-02   6.411 1.69e-10 ***
## public       1.740e-01  2.046e-02   8.508  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4977 on 2802 degrees of freedom
## Multiple R-squared:  0.373,  Adjusted R-squared:  0.371
## F-statistic: 185.2 on 9 and 2802 DF,  p-value: < 2.2e-16
```

1. Replique la estimación del modelo de regresión lineal múltiple. Interprete los resultados de cada coeficiente.

2. Estime un modelo GAM con grados de libertad 4 y 5 (Por ejemplo, `s(age,df = 4)`). Interprete el resultado del coeficiente con suavizamiento.
3. Ahora, ¿es el modelo planteado el mejor posible?

- Usa la estimación MARS para obtener el *mejor* modelo usando este enfoque, escribe el modelo de resultado.
- ¿Qué variables quedaron fuera?
- Analiza la gráfica de resultados (usando `plotmo`).

4. Contrastemos los resultados del ejercico anterior. Usando regresión bayesiana, ¿cuál es el mejor modelo? (asume una apriori `BIC` para los parámetros y una auniforme para la selección del modelo)

- Usa la estimación `bas.lm` para obtener el *mejor* modelo usando este enfoque, escribe el modelo de resultado.
- ¿Qué variables quedaron fuera? (comaprado con la regresión lineal múltiple)

4. Ejercicio: Bootstrap

Usaremos la base de datos `iris` que tiene características de plantas. Usaremos el algoritmo `kmeans` para generar tres grupos de la siguiente manera:

```
library(datasets)
head(iris)
```

```
##   Sepal.Length Sepal.Width Petal.Length Petal.Width Species
## 1          5.1         3.5          1.4         0.2  setosa
## 2          4.9         3.0          1.4         0.2  setosa
## 3          4.7         3.2          1.3         0.2  setosa
## 4          4.6         3.1          1.5         0.2  setosa
## 5          5.0         3.6          1.4         0.2  setosa
## 6          5.4         3.9          1.7         0.4  setosa
```

```
set.seed(20)
irisCluster <- kmeans(iris[, 3:4], 3, nstart = 20)
irisCluster
```

```
## K-means clustering with 3 clusters of sizes 50, 52, 48
##
## Cluster means:
##   Petal.Length Petal.Width
## 1     1.462000    0.246000
## 2     4.269231    1.342308
## 3     5.595833    2.037500
##
## Clustering vector:
##   [1] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
##  [36] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
##  [71] 2 2 2 2 2 2 2 3 2 2 2 2 2 3 2 2 2 2 2 2 2 2 2 2 2 2 2 2 3 3 3 3 3
## [106] 3 2 3 3 3 3 3 3 3 3 3 3 3 3 2 3 3 3 3 3 2 3 3 3 3 3 3 3 3 3 2 3
## [141] 3 3 3 3 3 3 3 3 3 3
##
## Within cluster sum of squares by cluster:
## [1]  2.02200 13.05769 16.29167
##  (between_SS / total_SS =  94.3 %)
##
## Available components:
```

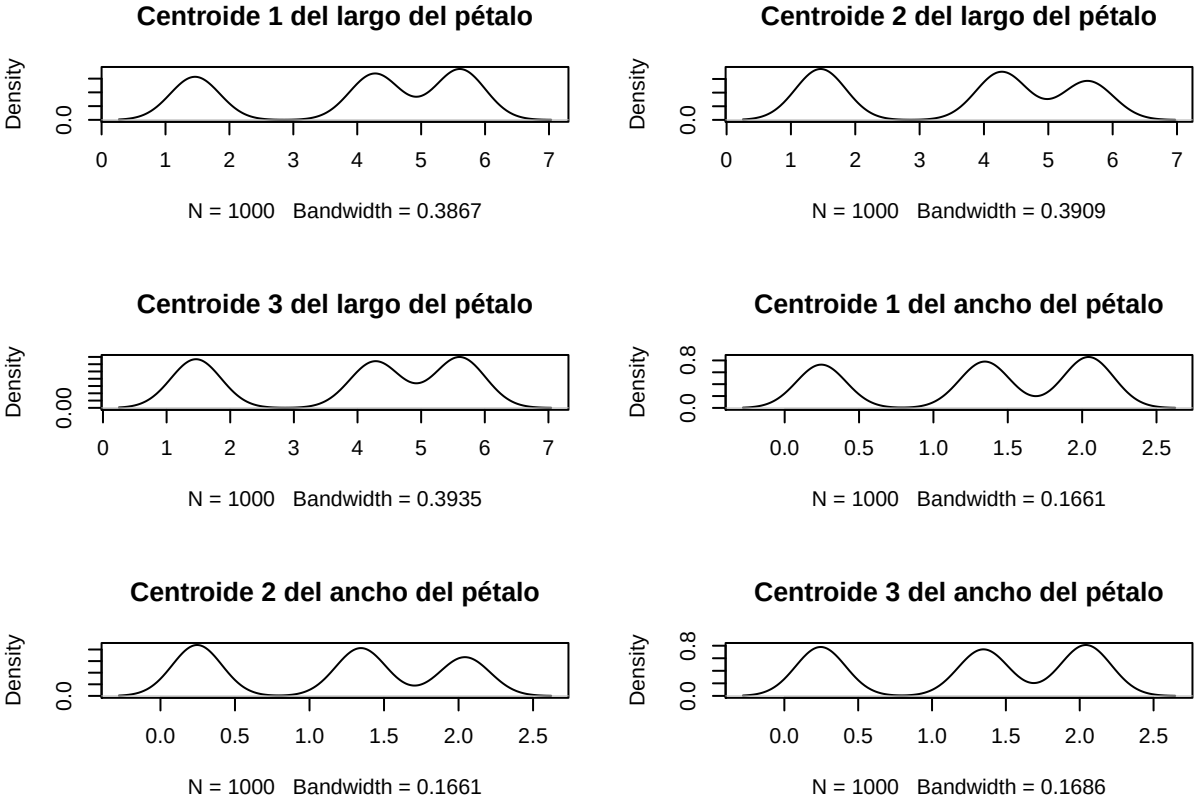
```
##
## [1] "cluster"      "centers"      "totss"        "withinss"
## [5] "tot.withinss" "betweenss"    "size"         "iter"
## [9] "ifault"
```

Como puedes ver, los promedios de cada grupo generado es:

```
irisCluster$centers
```

```
##   Petal.Length Petal.Width
## 1     1.462000    0.246000
## 2     4.269231    1.342308
## 3     5.595833    2.037500
```

Usa bootstrap para obtener ls distribución de probabilidad de los centroides. La configuración es: el tamño de la muestra en cada iteración es 500. Además, fija 1000 muestras bootstrap. Debes obtener resultados como los siguientes:



interpreta los resultados obtenidos, ¿puedes confiar en la primera estimación? (la de `irisCluster$centers`)