Práctica 1

***Alumno**: Francisco Javier Piqueras Martínez*
***Asignatura**: Modelado Estadístico de Datos*
***Fecha de entrega**: 16 de diciembre de 2019*

Índice

1. Descripción del documento 3
2. Ejercicios 3
2.1. Ejercicio 1 3
2.2. Ejercicio 2 4
2.3. Ejercicio 3 5
2.4. Ejercicio 4 5
2.5. Ejercicio 5 8

1. Descripción del documento

Este documento consiste en la realización de la Práctica 1 de la asignatura de MED.

2. Ejercicios

2.1. Ejercicio 1

Enunciado: *Se ha realizado un estudio para ver si influye la metodología docente a la hora de aprobar. Para ello **75** estudiantes han recibido la metodología 1 y **25** la metodología 2. De cada estudiante se ha registrado si al final aprobaban (1) o no (2). Los datos experimentales se dan en la tabla siguiente, donde el número de individuos con perfil aprobar = 1 y metodología = 1 es 35, con perfil aprobar = 1 y metodología = 2 es 15, con perfil aprobar = 2 y metodología = 1 es 40 y con perfil aprobar = 2 y metodología = 2 es 10. ¿Hay diferencias estadísticamente significativas entre las dos metodologías?*

aprobar	*metodologia*
1	1
...	...
1	1
1	2
...	...
1	2
2	1
...	...
2	1
2	2
...	...
2	2

Resolución:

Para saber si hay diferencias estadísticamente significativas entre las personas que han aprobado o no para cada una de las metodologías, se va a realizar un estudio de diferencia de proporciones poblacionales entre la metodología 1 y la metodología 2.

Puesto que la variable respuesta "aprobar" solo puede tomar dos valores, así como la variable explicativa "metodología", y como el tamaño muestral es grande (>5), vamos a aplicar el **test z de diferencia de proporciones**, enmarcado en un esquema D←D cuyo parámetro poblacional θ de interés es la diferencia de proporciones poblacionales entre π_1 y π_2 ($\theta = \pi_1 - \pi_2$). Para ello se considera el estadístico $\widehat{\Theta}$ dado por la v.a. $\widehat{\Pi_1} - \widehat{\Pi_2}$ (es decir, $\widehat{\Theta} = \widehat{\Pi_1} - \widehat{\Pi_2}$).

En primer lugar, se muestran los cálculos detallados para el intervalo de confianza al 95% dado por:

$$IC95\%(\pi_1 - \pi_2) = \left(\frac{35}{75} - \frac{15}{25} \mp \sqrt{\frac{35}{75}\left(1 - \frac{35}{75}\right)\frac{1}{75} + \frac{15}{25}\left(1 - \frac{15}{25}\right)\frac{1}{25}}\right) = (-0.246196, -0{,}020404)$$

Y para el contraste de hipótesis dado por:

$$z = \frac{\frac{35}{75} - \frac{15}{25}}{\sqrt{\frac{35+15}{75+25}(1 - \frac{35+15}{75+25})(\frac{1}{75} + \frac{1}{25})}} = -1.1547$$

Con lo que al ser $|z| = 1.1547 < 1.96$ se acepta la hipótesis nula o de igualdad y se concluye en que no existe diferencias significativas entre las poblaciones. Es decir, **no hay diferencia estadísticamente significativa entre ambas metodologías**.

2.2. Ejercicio 2

Enunciado: *En el modelo de regresión lineal, se define la matriz H (matriz "hat") como aquella matriz que pone el sombrero a la y, es decir que* $\hat{y} = Hy$*, entonces se verifica que H es simétrica e idempotente.*

a) Verdadero
b) Falso

Resolución:

Verdadero. H es simétrica e idempotente.

En primer lugar, se definen los conceptos simétrica e idempotente aplicado a matrices. Se dice que una matriz es simétrica cuando es igual a su traspuesta. Se dice que una matriz es idempotente cuando es igual a el producto por sí misma.

Por el método mínimos cuadrados, sabemos:

$$\beta = (X^T X)^{-1} X^T y \qquad \hat{y} = X\beta = X(X^T X)^{-1} X^T y$$

Por lo tanto:

$$\hat{y} = Hy \rightarrow \boldsymbol{H} = \frac{\hat{y}}{y} = \frac{X(X^T X)^{-1} X^T y}{y} = \boldsymbol{X}(\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X})^{-\mathbf{1}}\boldsymbol{X}^{\boldsymbol{T}}$$

Demostración de que H es simétrica:

$$H = H^T \rightarrow H^T = (X(X^T X)^{-1} X^T)^T = (X^T)^T [(X^T X)^{-1}]^T X^T = X[(X^T X)^T]^{-1} X^T = X(X^T X)^{-1} X^T = H$$

Demostración de que H es idempotente:

$$H = HH \rightarrow HH = X(X^TX)^{-1}X^TX(X^TX)^{-1}X^T = X(X^TX)^{-1}(X^TX)(X^TX)^{-1}X^T$$
$$= X(X^TX)^{-1}IX^T = X(X^TX)^{-1}X^T = H$$

2.3. Ejercicio 3

Enunciado: *En el modelo de regresión lineal, se define la matriz H (matriz "hat") como aquella matriz que pone el sombrero a la y, es decir que* $\hat{y} = Hy$, *entonces se verifica que los elementos* h_{ii} *de la diagonal de H, vienen dados por* $h_{ii} = x_i^T(X^TX)^{-1}x_i$, *siendo* $x_i^T = \left(1\ x_{i1} \ldots x_{ip}\right)$.

a) Verdadero
b) Falso

Resolución:

Verdadero.

Asumiendo, tal y como hemos calculado en el ejercicio anterior:

$$H = X(X^TX)^{-1}X^T$$

Vamos a asumir que $D = (X^TX)^{-1}$.

$$H = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1p} \\ 1 & x_{21} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{pmatrix} * \begin{pmatrix} d_{11} & d_{12} & \cdots & d_{1(p+1)} \\ d_{21} & d_{22} & \cdots & d_{2(p+1)} \\ \vdots & \vdots & \ddots & \vdots \\ d_{(p+1)1} & d_{(p+2)2} & \cdots & d_{(p+1)(p+1)} \end{pmatrix} * \begin{pmatrix} 1 & 1 & \cdots & 1 \\ x_{11} & x_{21} & \cdots & x_{n1} \\ \vdots & \vdots & \ddots & \vdots \\ x_{1p} & x_{2p} & \cdots & x_{np} \end{pmatrix}$$
$$= \begin{pmatrix} x_1^t d_1 & x_1^t d_2 & \cdots & x_1^t d_{p+1} \\ x_2^t d_1 & x_2^t d_2 & \cdots & x_2^t d_{p+1} \\ \vdots & \vdots & \ddots & \vdots \\ x_n^t d_1 & x_n^t d_2 & \cdots & x_n^t d_{p+1} \end{pmatrix} * \begin{pmatrix} 1 & 1 & \cdots & 1 \\ x_{11} & x_{21} & \cdots & x_{n1} \\ \vdots & \vdots & \ddots & \vdots \\ x_{1p} & x_{2p} & \cdots & x_{np} \end{pmatrix}$$

Por lo tanto, los elementos de la matriz H que pertenecen a la diagonal vienen dados por la siguiente fórmula:

$$\boldsymbol{h_{ii}} = \begin{pmatrix} x_i^t d_1 & x_i^t d_2 & \cdots & x_i^t d_{p+1} \end{pmatrix} * x_i = x_i^t * \begin{pmatrix} d_1 & d_2 & \cdots & d_{p+1} \end{pmatrix} * x_i = \boldsymbol{x_i^t} * \left(\boldsymbol{X^TX}\right) * \boldsymbol{x_i}$$

2.4. Ejercicio 4

Enunciado: *Dado el código en R, se pide rellenar el mayo número de 'xxx' posible:*

```
rm(list=ls())

datos=read.table('c_d_1.txt',header=T)
attach(datos)

ind1=which(exp==1)

ind2=which(exp==2)
n1=length(rta[ind1]); n1
n2=length(rta[ind2]); n2
tapply(rta,exp,mean)
tapply(rta,exp,sd)

t.test(rta[ind1],rta[ind2],var.equal=TRUE)
```

proporciona el siguiente resultado

```
> n1=length(rta[ind1]); n1
[1] 7
> n2=length(rta[ind2]); n2
[1] 10
> tapply(rta,exp,mean)
       1        2
25.85714 26.20000
> tapply(rta,exp,sd)
       1        2
9.856108 8.866917


        Two Sample t-test

data:  rta[ind1] and rta[ind2]
t = -0.075009, df = 15, p-value = 0.9412
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -10.085497   9.399783
sample estimates:
mean of x mean of y
 25.85714  26.20000
```

A continuación se escribe el siguiente código:

```
exp2=1*(exp==2)
summary(lm(data = datos,formula = rta ~ exp2))
```

que proporciona el siguiente resultado.

	Estimate	Std. Error	t value	$Pr(>\|t\|)$
(Intercept)	xxx	xxx	xxx	xxx
exp2	xxx	xxx	xxx	xxx

Tabla 2: Coeficientes de RL con $p = 1$ sin información rellenada

Residual standard error: xxx on xxx degrees of freedom

Multiple R-squared: xxx, Adjusted R-squared: xxx

F-statistic: xxx on xxx and xxx, p-value: xxx

Resolución:

	Estimate	Std. Error	T value	Pr(> |t|)
(Intercept)	**25.85714**	**3.5057**	**7.376**	**0.000000321**
Exp2	**0.3429**	**4.572**	**0.075**	**0.9412**

Residual standard error: xxx on **15** degrees of freedom
Multiple R-squared: xxx, Adjusted R-squared: xxx
F-statistic: **0.005626** on **1** and **15**, p-value**: 0.9412**

Intercept → (X=0), por lo tanto, exp1.

Estimate

La media de exp1 es 25.85714

exp2: $|\widehat{y_1} - \widehat{y_2}| = |25.85714 - 26.2| = |-0.34286| = 0.34286$

Asumiendo homocedasticidad ($\sigma_1^2 = \sigma_2^2$):

$$\hat{s}_c^2 = \frac{(n_1 - 1)\hat{s}_1^2 - (n_2 - 1)\hat{s}_2^2}{(n_1 - 1) + (n_2 - 1)} = 86.0305$$

Std. Error:

Intercept→ $\sqrt{\hat{s}_c^2(\frac{1}{n_1})} = 3.5057$

exp2 → $\sqrt{\hat{s}_c^2(\frac{1}{n_1} + \frac{1}{n_2})} = 4.572$

t-Student:

exp2 → $t = \frac{\widehat{y_1} - \widehat{y_2}}{4.572} = 0.075$

Intercept → $t = \frac{\widehat{y_1}}{3.5057} = 7.376$

p-value:

exp2 → $pval = 2\big(1 - pt(|0.075|, 15)\big) = 0.9412$

Intercept → $pval = 2\big(1 - pt(|7.376|, 15)\big) = 0.000000321$

Grados de libertad:

$$df = (n_1 - 1) + (n_2 - 1) = 15$$

f-statistic:

$$fstatistic = t^2 = 0.005626$$

2.5. Ejercicio 5

Enunciado: *Dado el código en R, se pide rellenar el mayo número de 'xxx' posible:*

```
rm(list=ls())

datos=read.table('c_n_1.txt',header=T)
attach(datos)

ind1=which(exp==1);

ind2=which(exp==2);
ind3=which(exp==3);
n1=length(rta[ind1]); n1
n2=length(rta[ind2]); n2
n3=length(rta[ind3]); n3
tapply(rta,exp,mean); tapply(rta,exp,sd)
summary(aov(rta~factor(exp)))
```

proporciona el siguiente resultado

```
> n1=length(rta[ind1]); n1
[1] 7
> n2=length(rta[ind2]); n2
[1] 10
> n3=length(rta[ind3]); n3
[1] 5
> tapply(rta,exp,mean); tapply(rta,exp,sd)
       1        2        3
25.85714 26.20000 22.60000
       1        2        3
9.856108 8.866917 8.876936
             Df Sum Sq Mean Sq F value Pr(>F)
factor(exp)   2   46.7   23.35   0.276  0.762
Residuals    19 1605.7   84.51
```

A continuación se escribe el siguiente código:

```
exp2=1*(exp==2)
exp3=1*(exp==3)
summary(lm(data=datos, formula=rta ~ exp2+exp3))
```

que proporciona el siguiente resultado donde se pide rellenar el mayor número posible de valores marcados con xxx.

	Estimate	Std. Error	t value	$Pr(>\|t\|)$
(Intercept)	xxx	xxx	xxx	xxx
exp2	xxx	xxx	xxx	xxx
exp3	xxx	xxx	xxx	xxx

Tabla 3: Coeficientes de RL con $p = 2$ sin información rellenada

Residual standard error: xxx on xxx degrees of freedom

Multiple R-squared: xxx, Adjusted R-squared: xxx

F-statistic: xxx on xxx and xxx, p-value: xxx

Resolución*:*

	Estimate	Std. Error	T value	Pr(> |t|)
(Intercept)	**25.85714**	**3.4746**	**7.442**	**0.0000000482**
Exp2	**0.3429**	**4.5303**	**0.076**	**0.940**
Exp3	**3.2571**	**5.3828**	**-0.605**	**0.552**

Residual standard error: xxx on **19** degrees of freedom
Multiple R-squared: xxx, Adjusted R-squared: xxx
F-statistic: xxx on **2** and **19**, p-value: xxx

Intercept → (X=0), por lo tanto, exp1.

Estimate

La media de exp1 es 25.85714

exp2: $|\widehat{y_1} - \widehat{y_2}| = |25.85714 - 26.2| = |-0.34286| = 0.34286$

exp3: $|\widehat{y_1} - \widehat{y_3}| = |25.85714 - 22.6| = |-3.25714| = 3.25714$

Asumiendo homocedasticidad ($\sigma_1^2 = \sigma_2^2 = \sigma_3^2$):

$$\hat{s}_c^2 = \frac{(n_1 - 1)\hat{s}_1^2 - (n_2 - 1)\hat{s}_2^2 - (n_3 - 1)\hat{s}_3^2}{(n_1 - 1) + (n_2 - 1) + (n_3 - 1)} = 84.4807$$

Std. Error:

Intercept→ $\sqrt{\hat{s}_c^2(\frac{1}{n_1})} = 3.4746$

exp2 → $\sqrt{\hat{s}_c^2(\frac{1}{n_1} + \frac{1}{n_2})} = 4.5303$

exp3 → $\sqrt{\hat{s}_c^2(\frac{1}{n_1} + \frac{1}{n_2})} = 5.3828$

t-Student:

exp3 → $t = \frac{\widehat{y_1} - \widehat{y_3}}{5.3828} = -0.605$

exp2 → $t = \frac{\widehat{y_1} - \widehat{y_2}}{4.5303} = 0.076$

Intercept → $t = \frac{\widehat{y_1}}{3.4740} = 7.442$

<u>p-value</u>:

exp3 → $pval = 2\big(1 - pt(|-0.605|, 19)\big) = 0.552$

exp2 → $pval = 2\big(1 - pt(|0.076|, 19)\big) = 0.940$

Intercept → $pval = 2\big(1 - pt(|7.442|, 19)\big) = 0.0000000482$

<u>Grados de libertad:</u>

$$df = (n_1 - 1) + (n_2 - 1) + (n_3 - 1) = 19$$