Modelado estadístico de datos: Práctica 1

Emilio Letón y Elisa M. Molanes-López

1. (CALC) (2 puntos) Se ha realizado un estudio para ver si influye la metodología docente a la hora de aprobar. Para ello 50 estudiantes han recibido la metodología 1 y 50 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 *metodologia* = 1 es 35, con perfil *aprobar* = 1 y *metodologia* = 2 es 15, con perfil *aprobar* = 2 y *metodologia* = 1 es 40 y con perfil *aprobar* = 2 y *metodologia* = 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

Tabla 1: Datos observados del esquema aprobar ← metodologia

2. (CALC) (1 punto) En el modelo de regresión lineal, se define la matriz $\mathbf{H}$ (matriz "hat") como aquella matriz que pone el sombrero a la $\mathbf{y}$, es decir que $\widehat{\mathbf{y}} = \mathbf{H}\mathbf{y}$, entonces se verifica que $\mathbf{H}$ es simétrica e idempotente.

   a) Verdadero.

   b) Falso.

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

   a) Verdadero.

   b) Falso.

4. (CALC) (2 puntos) El siguiente código en R

```
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

Se pide rellenar el mayor número posible de valores marcados con xxx.

5. (CALC) (2 puntos) El siguiente código en R

```
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

6. (2 puntos) Se ha realizado un estudio para ver si el peso en kg (rta) de unos deportistas depende de su cintura en cm (exp1), del número de km de entrenamiento (exp2) y del tipo de entrenamiento (exp3=1: Body building, exp3=2: Fitness). Han participado en el estudio 26 individuos. Los datos experimentales están en el fichero c_ccd.txt alojado en el curso virtual y se muestran en la tabla 6.

   Se pide:

   - Interpretar los resultados del modelo de regresión lineal con todas las variables.
   - Repetir el análisis quitando las variables no significativas. ¿Qué sucede?
   - Crear una variable interacción entre exp1 y exp3 e incorporarla al modelo anterior. ¿Qué ocurre?
   - Elegir de los tres modelos anteriores el mejor. ¿Se cumplen las condiciones de aplicabilidad de la regresión lineal?
   - Elaborar otro enunciado para estos datos.

   En el documento que se entregue habrá que incluir el código utilizado.

rta	exp1	exp2	exp3
69.3	83	8	1
69.6	84	7	1
71.5	86.5	4	1
71.5	84.5	32	1
70.6	86.4	15	1
69.2	82.5	6	1
65	82	10	2
65.4	81.8	17	2
63.7	80	6	2
69	82.5	18	1
65.8	84	0	2
68.7	87.2	3	2
64.8	84	10	2
70	86	11	1
65.9	84.2	18	2
63.9	84	4	2
62.1	79	12	2
73.1	97.2	18	2
75.4	91	0	1
72.6	89.5	9	1
69.6	89.5	11	2
72.3	87.5	7	1
67.3	87.5	15	2
68	87.5	5	2
68.1	86.5	14	2
71.3	87	9	1

Tabla 4: Datos observados del esquema rta ← exp1, exp2, exp3