TP 1. – Adrian Jose Zapater Reig

1 . 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 numero 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?

Si:

$$|\pi_1 - \pi_2| > EE[\pi_1 - \pi_2] * z_{1-\frac{\propto}{2}}$$

Rechazamos H_0.

$$EE[\pi_1 - \pi_2] = \sqrt{(V}[\pi_1 - \pi_2])$$

$$V[\widehat{\Pi_1} - \widehat{\Pi_2}] = \frac{\pi_1(1-\pi_1)}{n_1} + \frac{\pi_2(1-\pi_2)}{n_2}$$

$$EE[\widehat{\Pi_1} - \widehat{\Pi_2}] = \sqrt{\frac{\pi_1(1-\pi_1)}{n_1} + \frac{\pi_2(1-\pi_2)}{n_2}}$$

$$z_{1-\frac{\propto}{2}} = 1.96$$

Vamos a utilizar el test de z de diferencia de 2 proporciones. Es recomendable para tamaños muéstrales grandes [Comprobación > 5]

Según el test z, podemos rechazar H_0 si:

$$|z| > z_{1-\frac{\propto}{2}}$$

$$z = \frac{\pi_1 - \pi_2}{EE[\widehat{\Pi_1} - \widehat{\Pi_2}]}$$

$$EE[\widehat{\Pi_1} - \widehat{\Pi_2}] = \sqrt{V[\widehat{\Pi_1} - \widehat{\Pi_2}]}$$

$$V[\widehat{\Pi_1} - \widehat{\Pi_2}] = \frac{\pi_1(1-\pi_1)}{n_1} + \frac{\pi_2(1-\pi_2)}{n_2}$$

$$\pi_1 = \frac{35}{75} = 0.4667$$

$$\pi_2 = \frac{15}{25} = 0.6$$

$$EE[\widehat{\Pi_1} - \widehat{\Pi_2}] = \sqrt{\frac{\pi_1(1-\pi_1)}{n_1} + \frac{\pi_2(1-\pi_2)}{n_2}}$$

Estimamos EE bajo H_0, por lo que $\pi_1 = \pi_2 = \pi_c$→ valor común

$$\pi_c = \frac{a_1 + a_2}{n_1 + n_1}$$

$$EE_0[\widehat{\Pi_1} - \widehat{\Pi_2}] = \sqrt{\frac{\pi_c(1-\pi_c)}{n_1} + \frac{\pi_c(1-\pi_c)}{n_2}} = \sqrt{\pi_c(1-\pi_c) * \left(\frac{1}{n_1} + \frac{1}{n_2}\right)} = 0.11547$$

$$EE_0[\widehat{\Pi_1} - \widehat{\Pi_2}] = 0.11547$$

$$|z| = \frac{\pi_1 - \pi_2}{EE\left[\widehat{\Pi_1} - \widehat{\Pi_2}\right]} = \frac{0.4667 - 0.6}{0.11547} = 1.1544$$

$$|z| = 1.1544 < 1.96$$

Por lo que se acepta la H_0, y se concluye que no existe diferencia de las proporciones $\pi_1 = \pi_2$.

2. 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 simetrica e impotente.

$$\hat{y} = Hy$$

H es simétrica e idempotente

Linear model → $y = X\beta + \epsilon$

Sabemos de mínimos cuadrados que dado:

$$(y - X\beta)^2 = y^T y - 2\beta^T X^T y + \beta^T X^T X\beta$$

Derivamos por β e igualamos a 0:

$$-2X^T y + 2X^T X\beta = X^T y + X^T X\beta = 0$$
$$\beta = (X^T X)^{-1} X^T y$$

Aproximamos a $\hat{y}$, mediante: $\hat{y} = X\hat{\beta} \;\approx\; \hat{y} = Hy$

Sabemos de mínimos cuadrados que:

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

Se deduce que:

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

Partimos de esta ultima igualdad.

Simetrica:

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

Idempotente:

$$H^2 = H$$
$$H^2 = (\,X(X^T X)^{-1} X^T)(X(X^T X)^{-1} X^T)$$

Si X^T tiene p filas y m columnas, X tiene p columnas y m filas, por lo que se pueden multiplicar para obtener:

$$H^2 = X(X^T X)^{-1}(X^T X)(X^T X)^{-1} X^T$$
$$I = (X^T X)^{-1}(X^T X)$$
$$H^2 = XI(X^T X)^{-1} X^T = H$$
$$\therefore \; H^2 = H$$

3. 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 y□ = Hy, entonces se verifica que los

elementos h_{ii} de la diagonal de H vienen dados por $h_{ii} = x_i^T (X^T X)^{-1} x_i$, siendo $x_i^T = (1\ x_{i1}\ .\ .\ .\ x_{ip})$

Sabemos de mínimos cuadrados que dado:

$$(y - X\beta)^2 = y^T y - 2\beta^T X^T y + \beta^T X^T X\beta$$

Derivamos por β e igualamos a 0:

$$-2X^T y + 2X^T X\beta = X^T y + X^T X\beta = 0$$
$$\beta = (X^T X)^{-1} X^T y$$

Aproximamos a $\hat{y}$, mediante: $\hat{y} = X\hat{\beta} \ \approx \ \hat{y} = Hy$

Sabemos de mínimos cuadrados que:

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

Se deduce que:

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

Partimos de esta ultima igualdad.

$$H = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1p} \\ 1 & x_{21} & \cdots & x_{2p} \\ \vdots & \cdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{pmatrix} * \begin{pmatrix} q_{11} & q_{12} & \cdots & q_{1p+1} \\ q_{21} & q_{22} & \cdots & q_{2p+1} \\ \vdots & \cdots & \ddots & \vdots \\ q_{p+11} & q_{p+12} & \cdots & q_{p+1p+1} \end{pmatrix} * \begin{pmatrix} 1 & 1 & \cdots & 1 \\ x_{11} & x_{21} & \cdots & x_{n1} \\ \vdots & \cdots & \ddots & \vdots \\ x_{1p} & x_{2p} & \cdots & x_{np} \end{pmatrix}$$

$$x_1 = (1 \quad x_{11} \quad \cdots \quad x_{1p})$$

$$x_1{}^T = \begin{pmatrix} 1 \\ x_{11} \\ \vdots \\ x_{1p} \end{pmatrix}$$

Posicion (1,1) de la matriz resultado

$$q_1 = \begin{pmatrix} q_{11} \\ q_{21} \\ \vdots \\ q_{p+11} \end{pmatrix}$$

$$H = \begin{pmatrix} x_1^t * q_1 & x_1^t * q_2 & \cdots & x_1^t * q_{p+1} \\ x_2^t * q_1 & x_2^t * q_2 & \cdots & x_2^t * q_{p+1} \\ \vdots & \cdots & \ddots & \vdots \\ x_n^t * q_1 & x_n^t * q_2 & \cdots & x_n^t * q_{p+1} \end{pmatrix} * X^T$$

$$h_{11} = \left(x_1^t * q_1 \quad x_1^t * q_2 \quad \cdots \quad x_1^t * q_{p+1}\right) * x_1$$

$$h_{ii} = \left(x_i^t * q_1 \quad x_i^t * q_2 \quad \cdots \quad x_i^t * q_{p+1}\right) * x_i$$

$$h_{ii} = \ x_i{}^T \left(x_i^t * q_1 \quad x_i^t * q_2 \quad \cdots \quad x_i^t * q_{p+1}\right) * x_i = x_i{}^T (X^T X)^{-1} x_i$$

4. Se pide rellenar el mayor nú mero posible de valores marcados con xxx

	Estimate	St. Error	t- Student	P(t<)
y	25.85714	3.5057	7.376	0.000000321
Exp2	0.3429	4.572	0.075	0.9412

RSE	
Degrees of freedom	15
R^2	
Adjusted R^2	
F-stat	
P-value	0.9412

- y-Intercept Estimate

  Puesto que el intercepts es la media cuando X=0, que en este caso es la media de exp1

- Estimate exp2 – slope

$$|\overline{y_1} - \overline{y_2}| = 0.3429$$

- St. Error- exp2

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

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

$$t = \frac{\overline{y_1} - \overline{y_2}}{\sqrt{\hat{s}_c^2 * \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} = 0.075$$

- St. Error – exp2

$$t = \frac{\overline{y_1} - \overline{y_2}}{\sqrt{\hat{s}_c^2 * \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} = 0.075 \quad \therefore \ St.error = \sqrt{\hat{s}_c^2 * \left(\frac{1}{n_1} + \frac{1}{n_2}\right)} = 4.507$$

- T-student-y

$$t = \frac{\overline{y_1}}{\sqrt{\hat{s}_c^2 * \left(\frac{1}{n_1}\right)}} = 7.376$$

- St. Error – y

$$\sqrt{\hat{s}_c^2 * \left(\frac{1}{n_1}\right)} = 3.5057$$

- P-value – exp2

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

- P-value- y

$$2\big(1 - pt(|7.376|, 15)\big) = 0.000000231$$

- Degrees of freedom

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

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

	Estimate	St. Error	t- Student	P(t<)
y	25.85714	3.4740	7.442	0.000000321
Exp2	0.3429	4.5303	0.076	0.9412
Exp3	-3.2571	5.3828	-0.605	

RSE	
Degrees of freedom	19
R^2	
Adjusted R^2	
F-stat	
P-value	0.7615

- y-Intercept Estimate

  Puesto que el intercepts es la media cuando X=0, que en este caso es la media de exp1

- Estimate exp2 – slope

$$|\overline{y_1} - \overline{y_2}| = 0.3429$$

- Estimate exp3 – slope

$$|\overline{y_1} - \overline{y_3}| = 3.2571$$

- St. Error- exp2

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

$$\hat{s}_c^2 = \frac{(n_1 - 1)\hat{s}_1^2 + (n_2 - 1)\hat{s}_2^2 + (n_3 - 1)\hat{s}_3^2}{(n_1 - 1) + (n_2 - 1) + (n_2 - 1)} = 84.5083$$

$$t = \frac{\overline{y_1} - \overline{y_2}}{\sqrt{\hat{s}_c^2 * \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} = 0.076$$

- St. Error – exp2

$$t = \frac{\overline{y_1} - \overline{y_2}}{\sqrt{\hat{s}_c^2 * \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} = 0.075 \quad \therefore \ St.error = \sqrt{\hat{s}_c^2 * \left(\frac{1}{n_1} + \frac{1}{n_2}\right)} = 4.507$$

- St. Error- exp3

Asumimos homocedasticidad ($\sigma_1^2 = \sigma_3^2$):

$$\hat{s}_c^2 = \frac{(n_1 - 1)\hat{s}_1^2 + (n_2 - 1)\hat{s}_2^2 + (n_3 - 1)\hat{s}_3^2}{(n_1 - 1) + (n_2 - 1) + (n_2 - 1)} = 84.5083$$

$$t = \frac{\overline{y_1} - \overline{y_3}}{\sqrt{\hat{s}_c^2 * \left(\frac{1}{n_1} + \frac{1}{n_3}\right)}} = -0.605$$

- St. Error – exp3

$$t = \frac{\overline{y_1} - \overline{y_3}}{\sqrt{\hat{s}_c^2 * \left(\frac{1}{n_1} + \frac{1}{n_3}\right)}} = -0.605 \quad \therefore \ St.error = \sqrt{\hat{s}_c^2 * \left(\frac{1}{n_1} + \frac{1}{n_3}\right)} = 5.3828$$

- T-student-y

$$t = \frac{\overline{y_1}}{\sqrt{\hat{s}_c^2 * \left(\frac{1}{n_1}\right)}} = 7.442$$

- St. Error – y

$$\sqrt{\hat{s}_c^2 * \left(\frac{1}{n_1}\right)} = 3.4740$$

- P-value – exp2

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

- P-value – exp3

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

- P-value- y

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

- Degrees of freedom

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

6. Se ha realizado un estudio para ver si el peso en kg (rta) de unos deportistas depende de su cintura en cm (exp1), del numero 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. ¿Que sucede?
- Crear una variable interacción entre exp1 y exp3 e incorporarla al modelo anterior. ¿Que 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.

```
#Cargamos los datos:
rm(list=ls())
detach(datos)

datos=read.table('data_sets/c_ccd.txt', header = TRUE)
attach(datos)

summary(datos)

model.lr = lm(data = datos, formula = rta~exp1+exp2+exp3)
summary(model.lr)

# Call:
#   lm(formula = rta ~ exp1 + exp2 + exp3, data = datos)
#
# Residuals:
#   Min      1Q  Median      3Q     Max
# -1.7642 -0.5996  0.1003  0.4577  1.4069
#
# Coefficients:
#             Estimate   Std. Error t value  Pr(>|t|)
# (Intercept) 23.35297    3.78489    6.17   3.28e-06 ***
#   exp1        0.60220    0.04326   13.92 2.18e-12 ***
#   exp2        0.01173    0.02346    0.50    0.622
#   exp3       -4.16043    0.32004  -13.00 8.43e-12 ***
#   ---
#   Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#
# Residual standard error: 0.8111 on 22 degrees of freedom
# Multiple R-squared:  0.9469,           Adjusted R-squared:  0.9397
# F-statistic: 130.8 on 3 and 22 DF,  p-value: 3.551e-14

# Conclusiones:
# Mean rta: 68.60
# Para la primera columna tenemos 4 filas que estudiaremos a continuación:
# La primera fila (Intercept) indica el valor esperado de rta utilizando el valor de la media de exp1, 2 y 3.
# Desde la segunda a la cuarta fila tenemos el gradiente de la ecuación de cada exp.
```

Observaciones:
EXP1 y 3 tienen un p-valor muy bajo por lo que existe una fuerte correlación entre estas variables y el resultado.
EXP3 tiene una correlación negativa, mientras que EXP1 es positiva.
EXP2 tiene un p-valor alto, por lo que añade ruido al modelo.

```
model.lr2 = lm(data = datos, formula = rta~exp1+exp3)
summary(model.lr2)

# Call:
#   lm(formula = rta ~ exp1 + exp3, data = datos)
#
# Residuals:
#   Min      1Q   Median      3Q      Max
# -1.83742 -0.57356  0.05238  0.44248  1.29768
#
# Coefficients:
#   Estimate Std. Error t value Pr(>|t|)
# (Intercept) 23.50042    3.71137   6.332 1.84e-06 ***
#   exp1        0.60196    0.04254  14.150 7.72e-13 ***
#   exp3       -4.16392    0.31471 -13.231 3.07e-12 ***
#   ---
#   Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#
# Residual standard error: 0.7977 on 23 degrees of freedom
# Multiple R-squared:  0.9463,           Adjusted R-squared:  0.9417
# F-statistic: 202.7 on 2 and 23 DF,  p-value: 2.469e-15

#aplicamos factor porque exp3 es categorica
F = factor(exp3)
model.lr3 = lm(data = datos, formula = rta~exp1+exp3)
summary(model.lr3)
```

Los p-value decrecen un poco, posiblemente por eliminar el ruido que produce EXP2.
No se aprecia un cambio fuerte en otro valor.