

Parte teórica

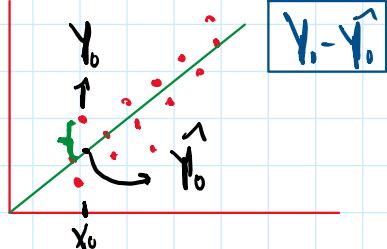
1. Determine el valor de verdad de las siguientes afirmaciones.

- (a) Bajo los supuestos del modelo de regresión lineal múltiple se cumple que la respuesta media estimada  $\hat{Y}_0 \sim N(\mathbb{E}[Y|x_0], \sigma^2 x_0(X'X)^{-1}x_0)$ , donde además,  $\hat{Y}_0$  es un estimador insesgado para  $Y_0$ . Falso
- (b) El error de predicción  $\hat{Y}_0 - Y_0$  tiene una varianza asociada dada por  $\sigma^2 x_0(X'X)^{-1}x_0$ , al igual que  $\hat{Y}_0$ , de aquí que si se sabe que  $x_0 = [1, x_{01}, x_{02}, \dots, x_{0k}]'$  es un punto en que no se comete extrapolación, entonces es correcto afirmar que  $x_0(X'X)^{-1}x_0 < 1$ .
- (c) El procedimiento analítico \*Shapiro-Wilk\*, cuyo juego de hipótesis está dado por  $H_0 : \varepsilon_i \sim N(0, \sigma^2)$  vs  $H_1 : \varepsilon_i \not\sim N(0, \sigma^2), i = 1, \dots, n$  permite determinar la normalidad de los residuales del modelo de regresión.
- (d) Una observación atípica está separada del resto de las observaciones en su valor de respuesta  $Y$  aunque no afecta los resultados del ajuste. Su evaluación se realiza a través del residual estandarizado  $|d_i| > 3$ .

(b). Falso.  $\hat{Y}_0 - Y_0, Y_0 - \hat{Y}_0$

$\text{Var}[\hat{Y}_0 - Y_0] = \text{Var}[Y_0] + \sigma^2 x_0(X'X)^{-1}x_0'$   
Son variables independientes

$$\hat{Y}_0 \sim N(\mathbb{E}[\hat{Y}_0], \sigma^2 x_0(X'X)^{-1}x_0')$$



$$\text{Var}[Y_0] + \text{Var}[\hat{Y}_0] = \text{Var}[Y_0] + \sigma^2 x_0(X'X)^{-1}x_0' = \sigma^2 I + \sigma^2 x_0(X'X)^{-1}x_0' = \sigma^2 I + \hat{Y}_0(X'X)^{-1}\hat{Y}_0'$$

$$\boxed{x_0 = [1, x_{01}, x_{02}, \dots, x_{0k}]} : \text{Extrapolación oculta: } x_0(X'X)^{-1}x_0' < \max_{1 \leq i \leq n} h_{ii} \sim \text{Diagonal } H$$

$$\boxed{h_{ii} < 1 : x_0(X'X)^{-1}x_0' < h_{ii} < 1} \text{ Verdadero.}$$

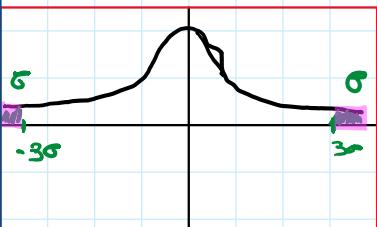
(c). Falso

(i). Se quiere determinar la normalidad de los errores y no de los residuales. (ii). Los residuales me ayudan a determinar la normalidad de los errores.

Hipótesis correcta:  $H_0: \varepsilon_i \sim \text{Normal} \quad i=1, \dots, n$   
 $H_1: \varepsilon_i \not\sim \text{Normal}$

(d). Falso.  $|d_i| > 3$

$$d_i = \frac{\varepsilon_i}{\sqrt{MSE}}$$



• Observación influencial: Hala el modelo en su dirección

• (e). Se cumple que una observación  $i$  es de balance si está definida en el espacio de la respuesta  $Y$  y se cumple que  $h_{ii} > 2p/n$ , afectando estadísticas como el  $R^2$  y los errores estándar de los coeficientes del modelo. ✓

• (f). Se cumple que  $D_i > 1, |DFBETAS_{j(i)}| > (2/\sqrt{n})$  y  $|DFFITS_i| > (2\sqrt{p}/n)$  simultáneamente para toda observación categorizada como influencial en un conjunto de datos.

(e). Falso. (i). Las observaciones de balance no se definen respecto al espacio de la respuesta sino de las covariadas.

(f). Falso.  $D_i > 1, |DFBETAS_{j(i)}| > 2/\sqrt{n}$   
 $|DFFITS_i| > 2\sqrt{p}/n$

• Se cumple al menos  $\frac{1}{3}$  de los criterios

2. Seleccione las expresiones adecuadas que se muestra a continuación, interpretelas y corrija las expresiones incorrectas.

a.  $d_i = \frac{\varepsilon_i}{\sqrt{MSE}}$  ✓

b.  $r_i = \frac{d_i}{\sqrt{1-h_{ii}}}$

c.  $DFBETAS_{j(i)} = \frac{\beta_j - \hat{\beta}_{j(i)}}{\sqrt{MSE_{(i)} r_{jj}}}$

d.  $DFFITS_i = \frac{\hat{y}_i - y_{i(0)}}{\sqrt{MSE_{(i)}}}$

$h_{ii}$

$\hat{y}_i$

$\beta_j - \hat{\beta}_{j(i)}$

$MSE_{(i)}$

$r_{jj}$

$y_{i(0)}$

$\hat{y}_i - y_{i(0)}$

$\sqrt{MSE_{(i)}}$

$\beta_j - \hat{\beta}_{j(i)}$

$MSE_{(i)}$

$\hat{y}_i - y_{i(0)}$

$\sqrt{MSE_{(i)}}$

$$\text{DFBETAS}_{j(i)} = \frac{\beta_j - \hat{\beta}_{j(i)}}{\sqrt{MSE_{ii} C_{jj}}} \xrightarrow{h_{ii}} \text{Diagonal principal } (X^T X)^{-1}$$

## Parte práctica

1. Verifique los supuestos del modelo de regresión, esto es,  $\varepsilon_i \stackrel{iid}{\sim} N(0, \sigma^2)$  a partir de los procedimientos apropiados para ello.

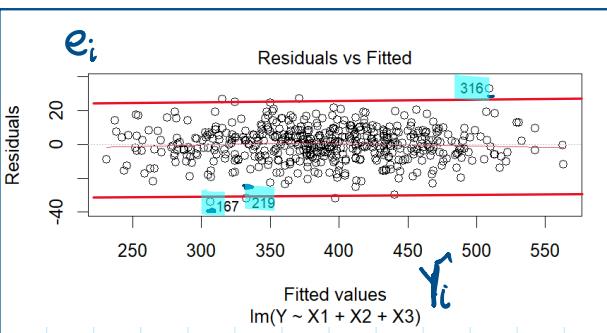
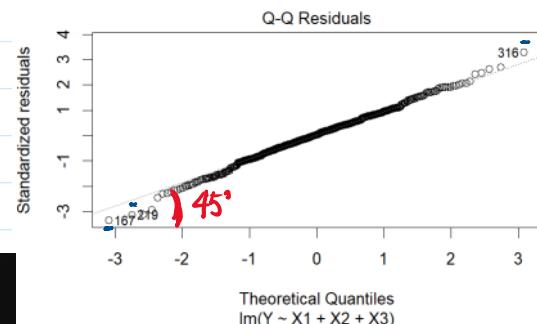
- (1). Independencia se asume en este curso
- (2). Media cero: Siempre se cumple
- (3). Distribución normal: (i). Shapiro; (ii). Gráfica.
- (4). Homocedasticidad: (i). Gráfica

- $H_0: \varepsilon_i \sim \text{Normal } \forall i$
- $H_1: \varepsilon_i \not\sim \text{Normal}$

```
Shapiro-Wilk normality test
data: modelo$residuals
W = 0.9957, p-value = 0.1877
```

$P_{\text{value}} > \alpha$

- A un nivel de significancia del 5%, se puede establecer que no existe evidencia suficiente para rechazar  $H_0$ , es decir, los errores tienen distribución normal.



- (ii). Atípicos:  $|D_{ii}| > 3$ ;  $|h_{ii}| > 3$

```
> atipicos_estandarizados
16 167 219 316
16 167 219 316
> atipicos_estudentizados
16 167 219 316
16 167 219 316
```

$D_{ii} > 3$ ,  $|DFBETAS_{j(i)}| > 2\sqrt{n}$   
 $|DFFITS| > 2\sqrt{p/n}$

•  $D_i$  (cooks):  $\hat{\beta}_i$ ; DFBETAS:  $\beta_j$ ; DFFITS (Ajuste  $\hat{y}_i$ )

```
> which(cooks > 1) # Verificar cooks
named integer(0)
```

```
> which(abs(DFFITS) > (2 * sqrt(p/n))) # Verificar DFFITS
1 16 22 29 45 75 107 109 113 122 154 167 173 214 219
1 16 22 29 45 75 107 109 113 122 154 167 173 214 219
222 236 247 276 289 294 316 317 356 362 453
222 236 247 276 289 294 316 317 356 362 453
```

3. Identifique puntos influenciales. Compare los criterios empleados para ello.

• No se identificaron observaciones influenciales

```
> summary(influencias)
Potentially influential observations of
  lm(formula = Y ~ X1 + X2 + X3) :

    dfb.1_ dfb.X1 dfb.X2 dfb.X3 dffit   cov.r   cook.d hat
1   -0.03 -0.08  0.12 -0.02 -0.18  0.97_*  0.01  0.01
8    0.04 -0.06  0.03 -0.06 -0.15  0.96_*  0.01  0.00
16   0.02 -0.10  0.17 -0.14 -0.28  0.94_*  0.02  0.01
29   0.06  0.05 -0.07 -0.13 -0.18  0.97_*  0.01  0.01
61   0.02 -0.03 -0.05  0.00 -0.11  0.98_*  0.00  0.00
113  0.01  0.13 -0.13 -0.07 -0.24  0.95_*  0.01  0.01
119 -0.08  0.01  0.12 -0.03 -0.16  0.97_*  0.01  0.01
158 -0.01  0.06  0.01 -0.08 -0.14  0.97_*  0.00  0.00
167 -0.21  0.08  0.00  0.23 -0.29  0.93_*  0.02  0.01
173  0.08 -0.06 -0.16  0.11  0.24  0.96_*  0.01  0.01
219 -0.23  0.24  0.18 -0.04 -0.33  0.94_*  0.03  0.01
254  0.07  0.01  0.03 -0.11  0.16  0.97_*  0.01  0.00
289  0.21 -0.19 -0.10 -0.05  0.25  0.96_*  0.02  0.01
311  0.01  0.00  0.00  0.00 -0.01  1.03_*  0.00  0.02
316 -0.23  0.09  0.23  0.15  0.32  0.93_*  0.03  0.01
362  0.16 -0.15  0.05 -0.15  0.24  0.97_*  0.01  0.01
```

4. Realice inferencia para  $\mathbf{x}_{01} = [1, 45.03, 80.88, 60.33]'$  y  $\mathbf{x}_{02} = [1, 77.08, 100, 13.76]'$  con su respectivo intervalo de predicción. Verifique primero si no se trata de un punto de extrapolación.

```
> which(abs(DFBetas) > 2/sqrt(n)) # Verificar DFBETAS
[1] 22 35 82 107 109 114 117 122 124 131 134
[12] 139 144 154 167 178 183 187 194 214 219 247
[23] 255 274 276 289 316 317 320 362 366 405 418
[34] 425 441 516 522 542 545 575 582 607 609 613
[45] 614 620 622 654 681 687 694 709 714 719 722
[56] 736 747 750 778 789 794 798 808 839 845 856
[67] 862 866 880 953 961 1001 1016 1030 1045 1075 1107
[78] 1113 1119 1134 1135 1173 1178 1181 1219 1236 1237 1247
[89] 1276 1280 1289 1294 1308 1316 1339 1397 1405 1435 1441
[100] 1454 1461 1516 1522 1529 1530 1534 1545 1559 1575
[111] 1582 1609 1614 1622 1624 1631 1634 1644 1654 1667 1673
[122] 1681 1683 1694 1722 1735 1736 1739 1747 1754 1755 1776
[133] 1794 1816 1817 1820 1838 1849 1856 1862 1921 1925 1941
[144] 1953
```

$$(i). \mathbf{X}_{01} = [1, 45.03, 80.88, 60.33]$$

$$\mathbf{X}_0 (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}_0^\top < \max_{i \in \mathcal{C}} h_{ii}$$

```
> ifelse(t(x01) %*% solve(t(X) %*% X) %*% x01 < max(Hat_values), "Pertenece a la region de diseño", "No pertenece")
[1]
[1,] "Pertenece a la region de diseño"
```

• No se comete extrapolación óptima

• Solo vamos a concluir respecto  $X_{01}$

$$\begin{cases} Y_0^1 = t_{n-p} \cdot \sqrt{MSE} L + X_0 (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}_0^\top J \\ Y_0^1 = t_{n-p} \cdot S.E(Y_0 - \hat{Y}_0) \end{cases}$$

$$Y_i = 70,80225 + 0,881 X_{i1} + 1,873 X_{i2} + 3,01 X_{i3}$$

$$Y_0^1 = 70,80225 + 0,881(45,03) + 1,873(80,88) + 3,01(60,33)$$

$$Y_0^1 = 145,8$$

$$\sqrt{MSE}; MSE = 10,18^2$$

```
fit      lwr      upr
1 445.887 425.8413 465.9327
```

11 15

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 70.80225   2.45665  28.82 <2e-16 ***
X1          0.88463   0.02848  31.06 <2e-16 ***
X2          1.87368   0.02673  90.41 <2e-16 ***
X3          3.04502   0.02789 109.19 <2e-16 ***
---
```

$$t_{n-p}; n-p = 500-1 = 496$$

$$\alpha/2 = 0,025$$

$$t = 1,96$$

```
Residual standard error: 10.18 on 496 degrees of freedom
Multiple R-squared:  0.9765, Adjusted R-squared:  0.9763
F-statistic: 6855 on 3 and 496 DF, p-value: < 2.2e-16
```