Tercer taller

miércoles, 18 de diciembre de 2024 2:00 p. m.

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

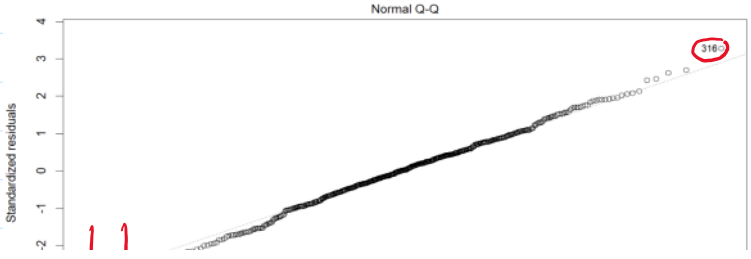
- (a) Bajo los supuestos del modelo de regresión lineal múltiple, se cumple que la respuesta media estimada $\hat{Y}_0 \sim N(E[Y|x_0], \sigma^2 x_0(X'X)^{-1}x_0')$ donde además, $\hat{Y}_0$ es un estimador insesgado para Y_0.
- (b) El error de predicción $\hat{Y}_0 - Y_0$ tiene una varianza asociada dada por $\sigma^2 x_0(X'X)^{-1}x_0$, al igual que $\hat{Y}_0$.
- (c) Si se sabe que $x_0 = [1, x_{01}, x_{02}, ..., x_{0k}]'$ es un punto en que no se comete extrapolación, entonces es correcto afirmar que $x_0(X'X)^{-1}x_0' < 1$.
- (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** $|d_i| > 3$.
- (e). Se cumple que una observación i es de balanceo si está definida en el espacio de la respuesta Y y se cumple que $h_{ii} > 2p/n$, controlando propiedades del modelo.
- (f). Una observación i es influencial si hala el modelo en su dirección, siendo inusual en el espacio de predictoras y la respuesta. Verdadera

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

a. $d_i = \frac{e_i}{\sqrt{MSE}}$ ✓ b. $r_i = \frac{d_i}{\sqrt{1-h_{ii}}}$

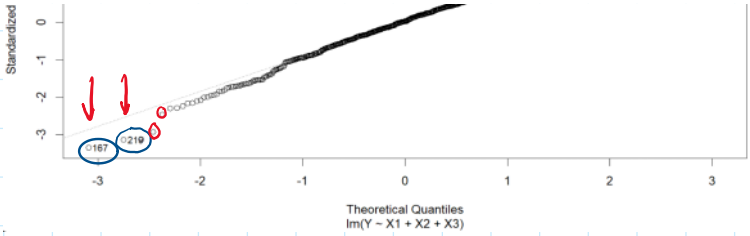
c. $DFBETAS_{j(i)} = \frac{\hat{\beta}_j - \hat{\beta}_{j(i)}}{\sqrt{MSE_{(i)} c_{jj}}}$ d. $DFFITS_i = \frac{\hat{y}_i - \hat{y}_{i(i)}}{\sqrt{MSE_{(i)}}}$

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




```
        Shapiro-Wilk normality test

data:  modelo$residuals
W = 0.9957, p-value = 0.1877
```



H_0: $\varepsilon_i \sim$ Normal

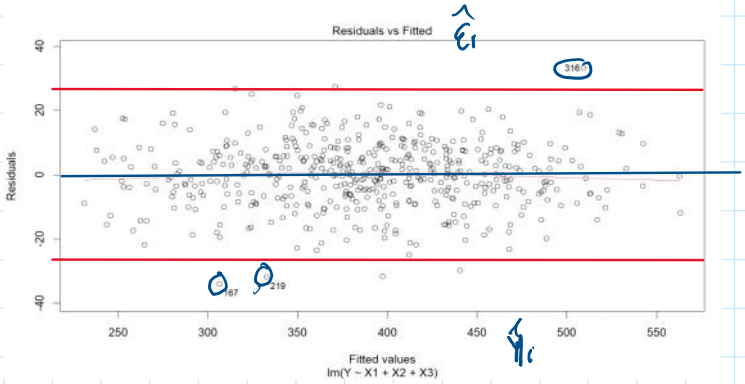
```
data:  modelo$residuals
W = 0.9957, p-value = 0.1877
```

$P_{val} < \alpha$ $\alpha = 0,05$

- No se rechaza H_0: se puede concluir a un nivel de significancia 5% que los errores tienen una $\sim$ Normal.

- Homogeneidad Varianza

- Verificar el supuesto de media cero



2. Identifique puntos atípicos y puntos de balanceo a través de un criterio gráfico y analítico. **¿Podrían ser estos puntos a su vez influenciales?**

$|d_i| > 3$; $|r_i| > 3$; Balanceo: $h_{ii} > 2p/n$: $2p/n < 1$

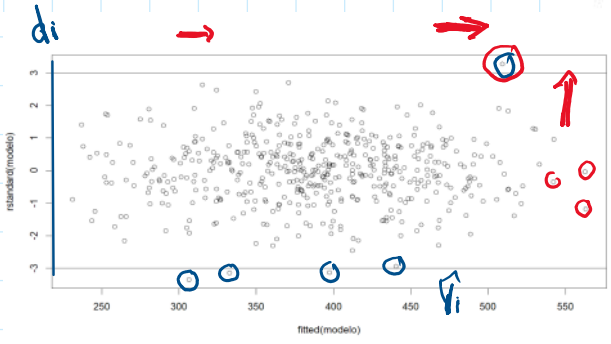
p = # parámetros (4): $\beta_0, \beta_1, \beta_2, \beta_3$
n = # muestra (500).

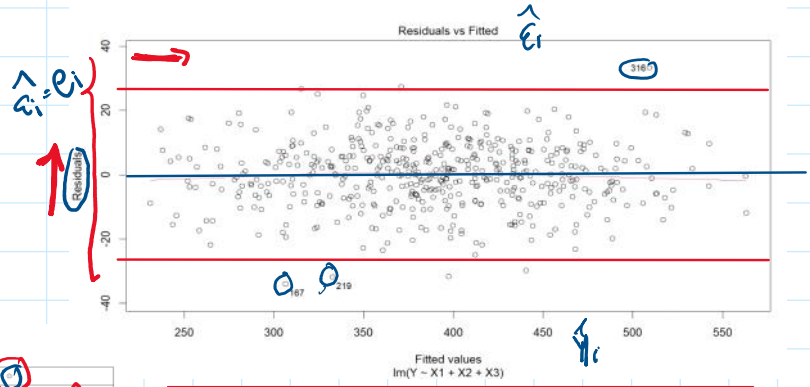
```
> balanceo
75 102 194 274 311
```

Pueden afectar R^2; y la estimación del error estándar asociado $\hat{\beta}_j$

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

Puntos atípicos





- Una observación influencial, se catalogaría como tal si es inusual en X y Y

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

A partir de 3 criterios: Cook's; DFBetas; DFFITS

- Cook's: $D_i > 1$
- DFBetas: $|DFBetas| > 2/\sqrt{n}$
- DFFITS: $|DFFITS| > 2\sqrt{p}$

Según los criterios anteriores, para que una observación sea influencial, no necesariamente debe ser catalogada como tal en todos los criterios anteriores.

- DFBetas: $|DFBetas| > 2/\sqrt{n}$
- DFFITS: $|DFFITS| > 2\sqrt{\frac{p}{n}}$

sea influencial, no necesariamente debe ser catalogada como tal en todos los criterios anteriores.

No hay influenciales

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

Según DFBetas

```
> 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 1535 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
```

```
> 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
```

Según DFFITS → y_i

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

$\underline{x}_0 (X'X)^{-1} \underline{x}_0' < \max\{h_{ii}\}$ ✓

Hiperplano (Ajustado)

- $\underline{x}_0 (X'X)^{-1} \underline{x}_0' < \max\{h_{ii}\}$ ✓ $\quad \underline{x}_0 = [1, \ldots, \ldots, 1]$

```
> summary(X)
  (Intercept)        X1              X2
 Min.   :1    Min.   : 21.06   Min.   : 27.57
 1st Qu.:1    1st Qu.:34.10    1st Qu.:43.76
 Median :1    Median :47.65    Median :63.08
 Mean   :1    Mean   :47.86    Mean   :62.94
 3rd Qu.:1    3rd Qu.:60.62    3rd Qu.:81.80
 Max.   :1    Max.   :76.19    Max.   :99.88
       X3
 Min.   :23.66
 1st Qu.:38.14
 Median :51.79
 Mean   :51.71
 3rd Qu.:65.36
 Max.   :81.00
```

X_{min}, X_{max}

- ifelse(..., ..., ...)

```
> ifelse(t(x01)%*%solve(t(X)%*%X)%*%x01 < max(Hat_values), "Per
tence a la region de diseno", "No pertenece")
     [,1]
[1,] "Pertence a la region de diseno"
> ifelse(t(x02)%*%solve(t(X)%*%X)%*%x02 < max(Hat_values), "Per
tence a la region de diseno", "No pertenece")
     [,1]
[1,] "No pertenece"
```

x_{01} ✓✓

x_{02} ✓ (No se puede realizar predicciones)

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

$\hat{y}_0$ | LI | LS

Intervalo predicción:

$\hat{y}_0 \pm t_{\alpha/2, n-p} \; s.e(\hat{y}_0 - y_0)$

$\hat{y}_0 \pm t_{\alpha/2, n-p} \sqrt{Var(\hat{y}_0 - y_0)}$

$$\hat{y}_0 \pm t_{\alpha/2, n-p} \sqrt{MSE[1] + \underline{x}_0 (X'X)^{-1} \underline{x}_0'}$$

El valor que podría tomar un Punto de predicción.