# CAPÍTULO 3

## ANOMALIAS EN REGRESIÓN Y MEDIDAS REMEDIALES

En este capítulo se estudiarán diversos diagnósticos de regresión que nos permitiran verificar si las suposiciones del modelo de regresión lineal se cumplen. Algunos de estos diagnósticos están basados en medidas que envuelven residuales y otros en plots de los residuales. También se discutirán posibles soluciones cuando las suposiciones no son satisfechas.

### 3.1 "Outliers", puntos de leverage alto y valores influenciales

Una observación $(y^*,x^*_1,\ldots\ldots x^*_p)$ es considerado un "outlier" si está bastante alejado de la mayoría de los datos sea en la dirección vertical o en la horizontal. Sin embargo, la mayoría de los textos llaman "outlier" a un valor alejado solamente en la dirección vertical y **punto de leverage alto** a una observación alejada en la dirección horizontal.
Una observación $(y^*,x^*_1,\ldots\ldots x^*_p)$ es considerado un **valor influencial** si su presencia afecta tremendamente el comportamiento del modelo. Por ejemplo, en el caso de regresión simple remover un valor influencial podría cambiar dramáticamente el valor de la pendiente.

Consideremos el siguiente conjunto de datos, consistente de 8 observaciones

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| X | 4 | 5 | 7 | 9 | 12 | 14 | 16 | 35 |
| Y | 6 | 7 | 12 | 15 | 18 | 21 | 28 | 65 |

La figura 3.1 muestra el plot de los datos. Notar que el punto O es un "outlier" y punto de leverage alto, pero a través de cálculos mostraremos que no es un valor influencial.

Primero, calcularemos la ecuación de regresión con el dato "outlier"

```
> l1=lm(y~x)
> summary(l1)

Call:
lm(formula = y ~ x)

Residuals:
   Min     1Q  Median     3Q    Max
-2.8825 -0.3140  0.4765  1.1130  1.4595

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -2.80152    1.03618  -2.704   0.0354 *
x            1.90600    0.06567  29.026 1.11e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.727 on 6 degrees of freedom
```

Multiple R-Squared: 0.9929, Adjusted R-squared: 0.9918
F-statistic: 842.5 on 1 and 6 DF, p-value: 1.108e-07

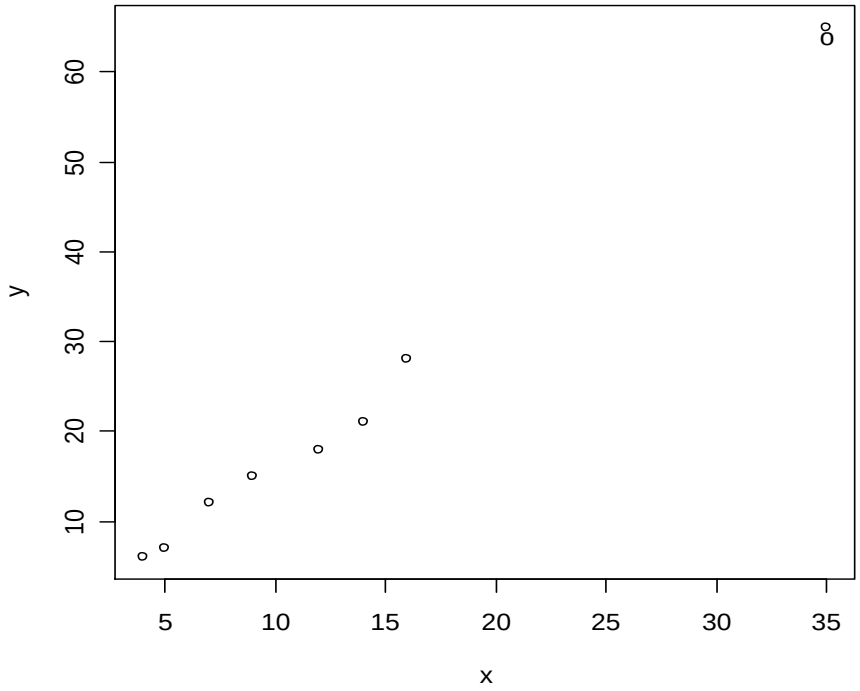


Figura 3.1. Ejemplo de una observación que es "outlier" y punto leverage alto pero que no es influencial.

Por otro lado la regresión sin el dato "outlier" es:

```
> x1=x[-8]
> y1=y[-8]
> l2=lm(y1~x1)
> summary(l2)

Call:
lm(formula = y1 ~ x1)

Residuals:
     1      2      3      4      5      6      7
 0.1034 -0.5818  1.0477  0.6773 -1.3784 -1.7489  1.8807

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -0.8443     1.3465  -0.627    0.558
x1            1.6852     0.1286  13.101 4.62e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Residual standard error: 1.442 on 5 degrees of freedom
Multiple R-Squared: 0.9717, Adjusted R-squared: 0.966
F-statistic: 171.6 on 1 and 5 DF, p-value: 4.625e-05

Notar que la pendiente y el $R^2$ han cambiado solo ligeramente. En consecuencia, la observación es un "outlier" y punto de leverage alto pero no es es influencial.

Supongamos ahora que al conjunto de datos anterior y al cual se le eliminó el "outlier", se le agrega el dato (35,22) que es considerado un punto de leverage alto. El plot del conjunto de datos es mostrado en la figura 3.2, donde la observación 0 representa el dato con leverage alto. Mostraremos que esta observación si resulta ser influencial.

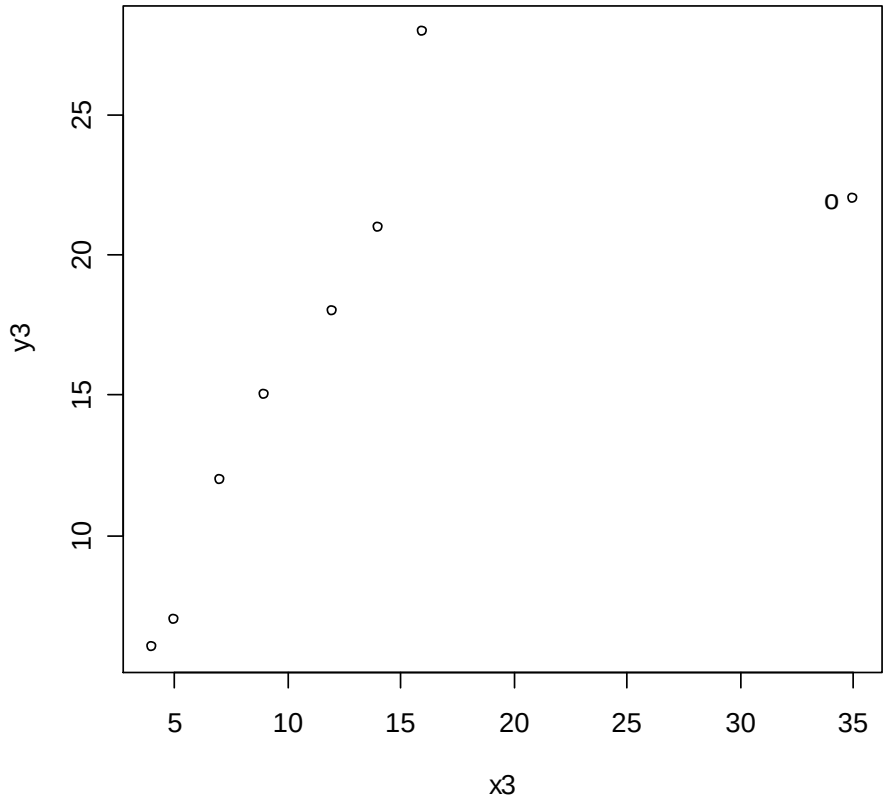


Figura 3.2. Ejemplo de una observación que es punto de leverage alto y que también es influencial.

La ecuación de regresión considerando el dato de leverage alto es

```
> l3=lm(y3~x3)
> summary(l3)

Call:
lm(formula = y3 ~ x3)
```

```
Residuals:
    Min      1Q  Median      3Q     Max
-5.7487 -5.1957 -0.1435  2.7556 10.1772

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   9.4642     3.6183   2.616   0.0398 *
x3            0.5224     0.2293   2.278   0.0629 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.03 on 6 degrees of freedom
Multiple R-Squared: 0.4638,     Adjusted R-squared: 0.3745
F-statistic: 5.191 on 1 and 6 DF,  p-value: 0.06295
```

Se puede observar el gran efecto sobre el $R^2$ que baja de 97.2% a 46.4% y un cambio drástico en la pendiente que cambia de 1.69 a 0.522..

En consecuencia un "outlier" vertical y/o punto de leverage alto puede ser influencial o no serlo. Por otro lado si una observación es influencial entonces es un "outlier" vertical o un punto de leverage alto.

## 3.2 Residuales y detección de "outliers".

Consideremos el modelo de regresión lineal múltiple $\mathbf{Y=XB+e}$ , donde E(**e**)=**0** y Var(**e**)=$\sigma^2$**I**.
Luego, $\widehat{\mathbf{Y}} = \mathbf{X}\widehat{\boldsymbol{\beta}}$ , pero como $\widehat{\boldsymbol{\beta}} = \mathbf{(X'X)^{-1}X'Y}$ , se tiene que $\widehat{\mathbf{Y}} = \mathbf{X(X'X)^{-1}X'Y} = \mathbf{HY}$ , ésta es la razón por la que a **H** se le llama la matriz HAT (sombrero), ya que actúa como una transformación de **Y** a $\widehat{\mathbf{Y}}$ . En particular, $\widehat{y}_i = \sum_{j=1}^{n} h_{ij} y_j$ , donde $h_{ij}$ es el elemento de la matríz H que está en la i-ésima fila y j-ésima columna.

Luego, el vector de residuales $\widehat{\mathbf{e}} = \mathbf{Y} - \widehat{\mathbf{Y}} = \mathbf{Y} - \mathbf{HY} = \mathbf{(I-H)Y}$ . En particular, $\widehat{e}_i = y_i - \sum_{j=1}^{n} h_{ij} y_j$ .

### 3.2.1 Media y Varianza del vector de residuales

Notar que

$$E(\widehat{\mathbf{e}}) = \mathbf{(I-H)}E(\mathbf{Y}) = \mathbf{(I-H)XB} = \mathbf{XB} - \mathbf{HXB} = \mathbf{XB} - \mathbf{X(X'X)^{-1}X'XB} = \mathbf{XB} - \mathbf{XB} = \mathbf{0}$$

Por otro lado,

$$Var(\widehat{\mathbf{e}}) = Var[\mathbf{(I-H)Y}] = \mathbf{(I-H)}Var(\mathbf{Y})\mathbf{(I-H)}' = \sigma^2\mathbf{(I-H)(I-H)}' = \sigma^2\mathbf{(I-H)^2} = \sigma^2\mathbf{(I-H)}$$

Aqui se ha usado el hecho que **I-H** es simétrica e idempotente, como se vió en la sección 2.2.3.

En particular, $Var(\widehat{e}_i) = \sigma^2(1-h_{ii})$. Esta varianza es estimada por $s^2(1\text{-}h_{ii})$.

Asimismo, $Cov(\widehat{e}_i, \widehat{e}_j) = -h_{ij}\sigma^2$

Notar que :

a) Tanto los errores $e_i$ como los residuales tienen media 0.
b) La varianza de los errores es constante, pero la de los residuales no lo es.
c) Los errores no están correlacionados, pero los residuales si lo están.

### 3.2.2 Residuales Estudentizados internamente

Para reducir el efecto de las varianzas de los residuales es más conveniente trabajar con versiones estandarizadas de ellos. Así, el ***residual estudentizado*** **internamente** se define por

$$r_i^* = \frac{\widehat{e}_i}{\sigma\sqrt{1-h_{ii}}} \qquad (3.1)$$

La covarianza de los residuales estudentizados es igual a

$$Cov(r_i^*, r_j^*) = Cov(\frac{\widehat{e}_i}{\sigma\sqrt{1-h_{ii}}}, \frac{\widehat{e}_j}{\sigma\sqrt{1-h_{jj}}}) = \frac{Cov(\widehat{e}_i, \widehat{e}_j)}{\sigma^2\sqrt{(1-h_{ii})(1-h_{jj})}} = \frac{-h_{ij}}{\sqrt{(1-h_{ii})(1-h_{jj})}}$$

En algunos programas estadísticos como MINITAB y el toolbox estadístico de MATLAB los $r_i^*$ son llamados **residuales estandarizados**.

### 3.2.3 Residuales estudentizados externamente

Supongamos que la i-ésima observación es eliminada del conjunto de datos y que se ajusta el modelo lineal con las n-1 observaciones que quedan. Sean $\widehat{\boldsymbol{\beta}}_{(i)}$ y $s^2_{(i)}$ las estimaciones de los parámetros del modelo y de la varianza de los errores respectivamente. Usando la siguiente identidad debido a Gauss

$$(\mathbf{X'}_{(i)}\,\mathbf{X}_{(i)})^{-1} = (\mathbf{X'X})^{-1} + \frac{(\mathbf{X'X})^{-1}\mathbf{x}_i\mathbf{x'}_i(\mathbf{X'X})^{-1}}{1-h_{ii}} \qquad (3.2)$$

donde $\mathbf{X}_{(i)}$ representa a la matriz $\mathbf{X}$ sin su i-ésima fila $\mathbf{x'}_i$., se puede establecer las siguientes relaciones entre $\hat{\boldsymbol{\beta}}$ y $\widehat{\boldsymbol{\beta}}_{(i)}$ y entre $s^2$ y $s^2_{(i)}$

i) $\hat{\boldsymbol{\beta}}_{(i)} = \hat{\boldsymbol{\beta}} - \frac{(\mathbf{X'X})^{-1}\mathbf{x}_i\hat{e}_i}{1-h_{ii}}$

ii) $s^2_{(i)} = \dfrac{n-p-1}{n-p-2}s^2 - \dfrac{\hat{e}_i^2}{(n-p-2)(1-h_{ii})}$

La identidad de Gauss es un caso particular de la **Identidad de Sherman-Morrison-Woodburry** (1950)

$$\left(\mathbf{A} \pm \mathbf{uv'}\right)^{-1} = \mathbf{A}^{-1} \mp \frac{\mathbf{A^{-1}uv'A^{-1}}}{1 \pm \mathbf{v'A^{-1}u}} \qquad (3.3)$$

donde **A** es una matriz cuadrada nosingular n x n, y **u** y **v** son dos vectores de dimensión n.
En efecto, puesto que $\mathbf{X}_{(i)}^{'}\mathbf{X}_{(i)} = \mathbf{X'X} - \mathbf{x}_i\mathbf{x'}_i$, donde $\mathbf{x'}_i$ es la i-ésima fila de **X**, se puede tomar **A**=**X'X** y **u**=**v**=$\mathbf{x_i}$. y se obtiene (3.2).

Si $\tilde{y}_i$ representa el valor estimado de la variable de respuesta para la i-ésima observación entonces $\tilde{y}_i = \mathbf{x'_i}\,\hat{\boldsymbol{\beta}}_{(i)}$. Como la i-ésima observación no fue usada en la estimación del modelo entonces $y_i$ y $\tilde{y}_i$ son independientes. Luego la varianza del residual $y_i - \tilde{y}_i$ está dada por

$$Var(y_i - \tilde{y}_i) = Var(y_i) + Var(\tilde{y}_i) = \sigma^2 + \sigma^2\mathbf{x'_i}\,(\mathbf{X'_{(i)}}\,\mathbf{X_{(i)}})^{-1}\mathbf{x}_i \quad (3.4)$$

Estimando $\sigma^2$ por $s^2_{(i)}$ y considerando que si $y_i$ no es un outlier entonces $E(y_i - \tilde{y}_i)=0$ se obtiene

$$t_i = \frac{y_i - \tilde{y}_i}{s_{(i)}\sqrt{1 + \mathbf{x'_i}\,(\mathbf{X}_{(i)}^{'}\mathbf{X_{(i)}})^{-1}\mathbf{x_i}}} \qquad (3.5)$$

$t_i$ es llamado un **residual estudentizado externamente** y tiene n-p-2 grados de libertad.

**Propiedad:** Relación entre el residual usual y el residual usando un modelo eliminando la i-ésima observación

$$y_i - \tilde{y}_i = \frac{\hat{e}_i}{1 - h_{ii}} \quad (3.6)$$

**Prueba:** Sustituyendo $\hat{\boldsymbol{\beta}}_{(i)} = (\mathbf{X}_{(i)}^{'}\mathbf{X}_{(i)})^{-1}\mathbf{X}_{(i)}^{'}\mathbf{y}_{(i)}$ en

$$y_i - \tilde{y}_i = y_i - \mathbf{x}_i'\tilde{\boldsymbol{\beta}}_{(i)} \qquad (3.7)$$

y usando luego la identidad de Gauss (3.2) se obtiene

$$y_i - \tilde{y}_i = y_i - \mathbf{x}_i'\left[(\mathbf{X'X})^{-1} + \frac{(\mathbf{X'X})^{-1}\mathbf{x_i}\mathbf{x}_i'(\mathbf{X'X})^{-1}}{1 - h_{ii}}\right]\mathbf{X}'_{(i)}\mathbf{y}_{(i)}$$

$$= y_i - \mathbf{x}_i'(\mathbf{X'X})^{-1}\mathbf{X}_{(\mathbf{i})}'\mathbf{y}_{(\mathbf{i})} - \frac{h_{ii}\mathbf{x}_\mathbf{i}'(\mathbf{X'X})^{-1}\mathbf{X}_{(\mathbf{i})}'\mathbf{y}_{(\mathbf{i})}}{1-h_{ii}}$$

$$= \frac{(1-h_{ii})y_i - (1-h_{ii})\mathbf{x}_\mathbf{i}'(\mathbf{X'X})^{-1}\mathbf{X}_{(\mathbf{i})}'\mathbf{y}_{(\mathbf{i})} - h_{ii}\mathbf{x}_\mathbf{i}'(\mathbf{X'X})^{-1}\mathbf{X}_{(\mathbf{i})}'\mathbf{y}_{(\mathbf{i})}}{1-h_{ii}}$$

$$= \frac{(1-h_{ii})y_i - \mathbf{x}_\mathbf{i}'(\mathbf{X'X})^{-1}\mathbf{X}_{(\mathbf{i})}'\mathbf{y}_{(\mathbf{i})}}{1-h_{ii}}$$

Si se usa luego el hecho que $\mathbf{X}_{(\mathbf{i})}'\mathbf{y}_{(\mathbf{i})} + \mathbf{x}_\mathbf{i} y_i = \mathbf{X'y}$, la anterior relación es equivalente a .

$$y_i - \tilde{y}_i = \frac{(1-h_{ii})y_i - \mathbf{x}_i'(\mathbf{X'X})^{-1}(\mathbf{X'y} - \mathbf{x}_\mathbf{i} y_i)}{1-h_{ii}}$$

como $\mathbf{x}_\mathbf{i}'(\mathbf{X'X})^{-\mathbf{1}}\mathbf{X'y} = \hat{y}_i$ se obtiene

$$y_i - \tilde{y}_i = \frac{(1-h_{ii})y_i - \hat{y}_i + h_{ii}y_i}{1-h_{ii}}$$

$$= \frac{y_i - \hat{y}}{1-h_{ii}} = \frac{\hat{e}_i}{1-h_{ii}}$$

Lo cual concluye la prueba.

Asímismo, se puede establecer la siguiente relación entre los distintos tipos de residuales

$$t_i = \frac{\hat{e}_i}{s_{(i)}\sqrt{1-h_{ii}}} = r_i^*\left(\frac{n-p-2}{n-p-1-r_i^{*2}}\right)^{1/2} \qquad (3.8)$$

## 3.3 Diagnósticos para detectar "outliers" y puntos de leverage alto

Ahora consideraremos diagnósticos basados en medidas y que servirán para detectar si una observación es un "outlier" o un punto de leverage alto. Los diagnósticos más básicos son:

Si $|h_{ii}|>2p/n$ (algunos usan 3p/n. Aquí p es el número de parámetros) entonces la i-ésima observación es considerado un "punto leverage" y pudiera ser influencial. Recordar que el promedio de los valores $h_{ii}$ es p/n

Si $|t_i|>2$ ( o si $|r_i|>2$) entonces la i-ésima observación es considerada un "outlier" y también puede ser influencial.

A continuación definiremos otros diagnósticos más sofisticados:

**i) La Distancia Cook (**Cook, 1977**):** Mide el cambio que ocurriría en el vector $\hat{\boldsymbol{\beta}}$ de coeficientes estimados de regresión (y por lo tanto en el valor ajustado de la variable de respuesta) si la i-ésima observación fuera omitida. Se calcula por

$$CD_i^2 = \frac{(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(i)})'\mathbf{X}'\mathbf{X}(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(i)})}{ps^2} = \frac{(\hat{Y} - \hat{Y}_{(i)})'(\hat{Y} - \hat{Y}_{(i)})}{ps^2} = r*_i^2 \frac{h_{ii}}{p(1-h_{ii})} \quad (3.9)$$

La primera igualdad resulta del hecho que

$$(\hat{\beta} - \hat{\beta}_{(i)})' X' X (\hat{\beta} - \hat{\beta}_{(i)}) = [X(\hat{\beta} - \hat{\beta}_{(i)}]'[X(\hat{\beta} - \hat{\beta}_{(i)}] = (\hat{Y} - \hat{Y}_{(i)})'(\hat{Y} - \hat{Y}_{(i)})$$

La segunda identidad resulta de la relación entre $\hat{\beta}$ y $\hat{\beta}_{(i)}$ mencionada en la sección 3.2.3, lo cual implica que

$$CD_i^2 = \frac{e_i^2}{(1-h_{ii})^2} \frac{x_i'(X'X)^{-1}x_i}{ps^2} = \frac{h_{ii}}{p(1-h_{ii})} \frac{e_i^2}{(1-h_{ii})s^2} = r_i^{*2} \frac{h_{ii}}{p(1-h_{ii})}$$

Notar que si el residual estandarizado es muy grande y si el valor leverage es grande, es decir si la observación está bien alejado en la dirección vertical y horizontal entonces su distancia Cook es bien grande. En general un $CD_i^2$ > 1 indica que la i-ésima observación es potencialmente influencial. Por su relación, que hay con los DFFITS, a ser descrito, mas adelante, una observación con $CD_i^2$ >4/n puede ser considerado influencial. Más formalmente, una observación con $CD_i^2$ > F(0.50,p,n-p) es considerado como un valor influencial, la razón es que $\boldsymbol{\beta}$ cae en un elipsoide de confianza centrado en $\hat{\boldsymbol{\beta}}$ de radio F($\alpha$,p,n-p). Aquí p es el número de coeficientes en el modelo. Sin embargo, si todos los $CD_i^2$ son menores que 1 es mejor plotear los valores $CD_i^2$ para detectar si hay observaciones con valores grandes comparados con los demás.

**ii) DFFITS (**Belsley, Kuh, y Welsch, 1980). Es similar a la Distancia Cook, excepto por un factor de escala y el remplazo de la varianza estimada $s^2$ por $s_{(i)}^2$, la varianza estimada del error excluyendo la i-ésima observación en los cálculos. Más precisamente,

$$DFFITS_i^2 = \frac{(\hat{\mathbf{y}} - \hat{\mathbf{y}}_{(i)})'(\hat{\mathbf{y}} - \hat{\mathbf{y}}_{(i)})}{s_{(i)}^2} = t_i^2 \frac{h_{ii}}{(1-h_{ii})} \quad (3.10)$$

En forma similar a lo realizado en la segunda identidad de la Distancia Cook, se obtiene que

$$DFFITS_i^2 = \frac{h_{ii} e_i^2}{(1-h_{ii})^2 s_{(i)}^2} = t_i^2 \frac{h_{ii}}{(i-h_{ii})}$$

Un $|DFFITS_i| > 2\sqrt{\frac{p}{n}}$ indica un posible valor influencial. Notar que

$$CD_i^2 = \frac{r_i^2}{pt_i^2} DFFITS_i^2 \qquad (3.11)$$

**iii) DFBETAS (**Belsley, Kuh, y Welsch, 1980**).** Mide la influencia de la i-ésima observación en cada uno de los coeficientes de regresión. Se calcula por

$$(DFBETAS)_{ji} = \frac{\beta_j - \beta_{j,(i)}}{s_{(i)}\sqrt{c_{jj}}} \qquad (3.12)$$

(i=1,..n, j=0,….p), donde $c_{jj}$ es el j-ésimo elemento de la diagonal de $\mathbf{(X'X)^{-1}}$.

Un $|DFBETAS|_{ji} > \frac{2}{\sqrt{n}}$ indica un posible valor influencial.

**iv) COVRATIO (**Belsley, Kuh, y Welsch, 1980**)**

Mide el efecto en la variabilidad de los coeficientes de regresión al remover la i-ésima observación. Se define por

$$COVRATIO_i = \frac{\det[s_{(i)}^2 (X'_{(i)} X_{(i)})^{-1}]}{\det[s^2 (X'X)^{-1}]} \qquad (3.13)$$

para i=1,….n. Donde det[A] significa el determinante de la matriz A. Usando propiedades de determinantes, se puede obtener la siguiente equivalente fórmula

$$(COVRATIO)_i = \left(\frac{s_{(i)}^2}{s^2}\right)^p \frac{1}{(1-h_{ii})} \qquad (3.14)$$

Si $(COVRATIO)_i > 1+3p/n$ o si $(COVRATIO)_i < 1-3p/n$ entonces la i-ésima observación tiene un valor influencial grande.

**Ejemplo 1:** Aplicar los diagnósticos de regresión al conjunto de datos **millaje.**

La siguiente es una lista completa de los diagnósticos para todas las observaciones obtenida usando SAS..

| Obs | Dep Var MPG | Predict Value | Std Err Predict | Residual | Std Err Residual | Student Residual | -2-1-0 1 2 | Cook's D |
|---|---|---|---|---|---|---|---|---|
| **1** | **65.4000** | **53.4146** | **1.267** | **11.9854** | **3.426** | **3.499** | **\|******\|** | **0.335** |
| **2** | **56.0000** | **49.7766** | **1.029** | **6.2234** | **3.505** | **1.776** | **\|***  \|** | **0.054** |
| **3** | **55.9000** | **49.7766** | **1.029** | **6.1234** | **3.505** | **1.747** | **\|***  \|** | **0.053** |
| 4 | 49.0000 | 45.3013 | 0.742 | 3.6987 | 3.577 | 1.034 | \|**  \| | 0.009 |
| **5** | **46.5000** | **50.2870** | **1.092** | **-3.7870** | **3.486** | **-1.086** | ****\|  \|** | **0.023** |
| 6 | 46.2000 | 45.3482 | 0.725 | 0.8518 | 3.580 | 0.238 | \|  \| | 0.000 |
| **7** | **45.4000** | **49.7766** | **1.029** | **-4.3766** | **3.505** | **-1.249** | ****\|  \|** | **0.027** |
| **8** | **59.2000** | **47.2349** | **1.213** | **11.9651** | **3.445** | **3.473** | **\|******\|** | **0.299** |
| **9** | **53.3000** | **47.2349** | **1.213** | **6.0651** | **3.445** | **1.760** | **\|***  \|** | **0.077** |
| 10 | 43.4000 | 41.9529 | 0.639 | 1.4471 | 3.596 | 0.402 | \|  \| | 0.001 |
| 11 | 41.1000 | 44.4650 | 0.646 | -3.3650 | 3.595 | -0.936 | *\|  \| | 0.006 |
| **12** | **40.9000** | **39.5790** | **1.263** | **1.3210** | **3.428** | **0.385** | **\|  \|** | **0.004** |
| 13 | 40.9000 | 38.8124 | 0.976 | 2.0876 | 3.520 | 0.593 | \|*  \| | 0.005 |
| 14 | 40.4000 | 44.4650 | 0.646 | -4.0650 | 3.595 | -1.131 | **\|  \| | 0.008 |
| 15 | 39.6000 | 45.6039 | 0.735 | -6.0039 | 3.578 | -1.678 | ***\|  \| | 0.024 |
| 16 | 39.3000 | 44.4650 | 0.646 | -5.1650 | 3.595 | -1.437 | **\|  \| | 0.013 |
| 17 | 38.9000 | 42.5103 | 0.605 | -3.6103 | 3.602 | -1.002 | **\|  \| | 0.006 |
| **18** | **38.8000** | **39.5790** | **1.263** | **-0.7790** | **3.428** | **-0.227** | **\|  \|** | **0.001** |
| 19 | 38.2000 | 42.5103 | 0.605 | -4.3103 | 3.602 | -1.197 | **\|  \| | 0.008 |
| 20 | 42.2000 | 38.4951 | 0.631 | 3.7049 | 3.598 | 1.030 | \|**  \| | 0.007 |
| 21 | 40.9000 | 38.0473 | 0.651 | 2.8527 | 3.594 | 0.794 | \|*  \| | 0.004 |
| 22 | 40.7000 | 42.5157 | 0.729 | -1.8157 | 3.579 | -0.507 | *\|  \| | 0.002 |
| 23 | 40.0000 | 37.8978 | 0.682 | 2.1022 | 3.589 | 0.586 | \|*  \| | 0.002 |
| 24 | 39.3000 | 40.2540 | 0.504 | -0.9540 | 3.618 | -0.264 | \|  \| | 0.000 |
| 25 | 38.8000 | 38.0856 | 0.647 | 0.7144 | 3.595 | 0.199 | \|  \| | 0.000 |
| **26** | **38.4000** | **38.8139** | **1.152** | **-0.4139** | **3.466** | **-0.119** | **\|  \|** | **0.000** |
| 27 | 38.4000 | 37.7657 | 0.849 | 0.6343 | 3.553 | 0.179 | \|  \| | 0.000 |
| 28 | 38.4000 | 38.0473 | 0.651 | 0.3527 | 3.594 | 0.098 | \|  \| | 0.000 |
| **29** | **29.5000** | **38.5108** | **0.623** | **-9.0108** | **3.599** | **-2.504** | *******\|  \|** | **0.038** |
| **30** | **46.9000** | **43.5274** | **1.173** | **3.3726** | **3.459** | **0.975** | **\|*  \|** | **0.022** |
| 31 | 36.3000 | 34.9973 | 0.659 | 1.3027 | 3.593 | 0.363 | \|  \| | 0.001 |
| 32 | 36.1000 | 39.0892 | 0.661 | -2.9892 | 3.592 | -0.832 | *\|  \| | 0.005 |
| 33 | 36.1000 | 39.2925 | 0.549 | -3.1925 | 3.611 | -0.884 | *\|  \| | 0.004 |
| 34 | 35.4000 | 36.0564 | 0.512 | -0.6564 | 3.617 | -0.181 | \|  \| | 0.000 |
| 35 | 35.3000 | 35.8061 | 0.649 | -0.5061 | 3.595 | -0.141 | \|  \| | 0.000 |
| 36 | 35.1000 | 39.4107 | 0.564 | -4.3107 | 3.609 | -1.194 | **\|  \| | 0.007 |
| 37 | 35.1000 | 37.8083 | 0.448 | -2.7083 | 3.625 | -0.747 | *\|  \| | 0.002 |
| 38 | 35.0000 | 37.9647 | 0.497 | -2.9647 | 3.619 | -0.819 | *\|  \| | 0.003 |
| 39 | 33.2000 | 34.1686 | 0.598 | -0.9686 | 3.603 | -0.269 | \|  \| | 0.000 |
| 40 | 32.9000 | 34.1686 | 0.598 | -1.2686 | 3.603 | -0.352 | \|  \| | 0.001 |
| 41 | 32.3000 | 30.8137 | 0.828 | 1.4863 | 3.558 | 0.418 | \|  \| | 0.002 |
| 42 | 32.2000 | 34.8852 | 0.512 | -2.6852 | 3.617 | -0.742 | *\|  \| | 0.002 |
| 43 | 32.2000 | 34.9947 | 0.465 | -2.7947 | 3.623 | -0.771 | *\|  \| | 0.002 |
| 44 | 32.2000 | 34.0747 | 0.524 | -1.8747 | 3.615 | -0.519 | *\|  \| | 0.001 |
| 45 | 32.2000 | 35.2763 | 0.576 | -3.0763 | 3.607 | -0.853 | *\|  \| | 0.004 |
| 46 | 31.5000 | 35.5677 | 0.478 | -4.0677 | 3.621 | -1.123 | **\|  \| | 0.004 |
| 47 | 31.5000 | 34.4756 | 0.454 | -2.9756 | 3.624 | -0.821 | *\|  \| | 0.002 |
| 48 | 31.4000 | 34.2879 | 0.491 | -2.8879 | 3.620 | -0.798 | *\|  \| | 0.002 |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 49 | 31.4000 | 34.9234 | 0.444 | -3.5234 | 3.626 | -0.972 | \| *\| | \| | 0.003 |
| 50 | 31.2000 | 30.9076 | 0.857 | 0.2924 | 3.551 | 0.082 | \| \| | \| | 0.000 |
| 51 | 33.7000 | 29.7337 | 0.610 | 3.9663 | 3.601 | 1.101 | \| \|** | \| | 0.007 |
| 52 | 32.6000 | 29.7337 | 0.610 | 2.8663 | 3.601 | 0.796 | \| \|* | \| | 0.004 |
| 53 | 31.3000 | 29.7337 | 0.610 | 1.5663 | 3.601 | 0.435 | \| \| | \| | 0.001 |
| 54 | 31.3000 | 29.3738 | 0.633 | 1.9262 | 3.598 | 0.535 | \| \|* | \| | 0.002 |
| **55** | **30.4000** | **23.9641** | **1.094** | **6.4359** | **3.485** | **1.847** | **\| \|*** | \|** | **0.067** |
| 56 | 28.9000 | 26.4784 | 0.715 | 2.4216 | 3.582 | 0.676 | \| \|* | \| | 0.004 |
| 57 | 28.0000 | 27.4881 | 0.670 | 0.5119 | 3.591 | 0.143 | \| \| | \| | 0.000 |
| 58 | 28.0000 | 31.4862 | 0.929 | -3.4862 | 3.533 | -0.987 | \| *\| | \| | 0.013 |
| 59 | 28.0000 | 29.7337 | 0.610 | -1.7337 | 3.601 | -0.481 | \| \| | \| | 0.001 |
| 60 | 28.0000 | 30.4341 | 0.853 | -2.4341 | 3.552 | -0.685 | \| *\| | \| | 0.005 |
| 61 | 28.0000 | 28.8107 | 0.737 | -0.8107 | 3.578 | -0.227 | \| \| | \| | 0.000 |
| 62 | 27.7000 | 27.1006 | 0.599 | 0.5994 | 3.603 | 0.166 | \| \| | \| | 0.000 |
| 63 | 25.6000 | 24.7507 | 0.902 | 0.8493 | 3.540 | 0.240 | \| \| | \| | 0.001 |
| 64 | 25.3000 | 23.2965 | 0.787 | 2.0035 | 3.567 | 0.562 | \| \|* | \| | 0.003 |
| 65 | 23.9000 | 23.4217 | 0.741 | 0.4783 | 3.577 | 0.134 | \| \| | \| | 0.000 |
| 66 | 23.6000 | 23.4906 | 0.679 | 0.1094 | 3.589 | 0.030 | \| \| | \| | 0.000 |
| **67** | **23.6000** | **24.0105** | **1.520** | **-0.4105** | **3.321** | **-0.124** | **\| \|** | \| | **0.001** |
| **68** | **23.6000** | **23.0093** | **0.628** | **0.5907** | **3.598** | **0.164** | **\| \|** | \| | **0.000** |
| 69 | 23.6000 | 22.8059 | 0.726 | 0.7941 | 3.580 | 0.222 | \| \| | \| | 0.000 |
| 70 | 23.6000 | 22.8684 | 0.684 | 0.7316 | 3.588 | 0.204 | \| \| | \| | 0.000 |
| 71 | 23.5000 | 20.7522 | 1.295 | 2.7478 | 3.415 | 0.805 | \| \|* | \| | 0.019 |
| **72** | **23.4000** | **19.9118** | **1.692** | **3.4882** | **3.237** | **1.077** | **\| \|**** | \| | **0.063** |
| 73 | 23.4000 | 22.7989 | 0.824 | 0.6011 | 3.559 | 0.169 | \| \| | \| | 0.000 |
| 74 | 23.1000 | 22.8458 | 0.781 | 0.2542 | 3.568 | 0.071 | \| \| | \| | 0.000 |
| **75** | **22.9000** | **18.7231** | **1.281** | **4.1769** | **3.421** | **1.221** | **\| \|**** | \| | **0.042** |
| 76 | 22.9000 | 19.2081 | 1.113 | 3.6919 | 3.479 | 1.061 | \| \|** | \| | 0.023 |
| 77 | 19.5000 | 18.6925 | 0.888 | 0.8075 | 3.543 | 0.228 | \| \| | \| | 0.001 |
| **78** | **18.1000** | **20.6117** | **2.033** | **-2.5117** | **3.035** | **-0.828** | **\| *\|** | \| | **0.061** |
| 79 | 17.2000 | 19.0194 | 1.082 | -1.8194 | 3.489 | -0.521 | \| *\| | \| | 0.005 |
| **80** | **17.0000** | **20.7779** | **1.593** | **-3.7779** | **3.287** | **-1.149** | **\| **\|** | \| | **0.062** |
| **81** | **16.7000** | **19.3010** | **1.871** | **-2.6010** | **3.137** | **-0.829** | **\| *\|** | \| | **0.049** |
| **82** | **13.2000** | **12.7102** | **1.636** | **0.4898** | **3.266** | **0.150** | **\| \|** | \| | **0.001** |

| Obs | Rstudent | Hat Diag H | Cov Ratio | Dffits | INTERCEP Dfbetas | VOL Dfbetas | HP Dfbetas | SP Dfbetas | WT Dfbetas |
|---|---|---|---|---|---|---|---|---|---|
| **1** | **3.7900** | **0.1204** | **0.5107** | **1.4021** | **1.1286** | **0.3421** | **1.1002** | **-1.0801** | **-1.2007** |
| **2** | **1.8014** | **0.0794** | **0.9408** | **0.5289** | **0.4117** | **0.1168** | **0.3936** | **-0.3937** | **-0.4227** |
| **3** | **1.7712** | **0.0794** | **0.9472** | **0.5200** | **0.4048** | **0.1148** | **0.3871** | **-0.3871** | **-0.4156** |
| 4 | 1.0346 | 0.0413 | 1.0383 | 0.2146 | 0.0142 | 0.0628 | 0.0176 | -0.0026 | -0.0793 |
| **5** | **-1.0877** | **0.0894** | **1.0852** | **-0.3408** | **-0.2762** | **-0.0706** | **-0.2633** | **0.2655** | **0.2751** |
| 6 | 0.2365 | 0.0394 | 1.1072 | 0.0479 | 0.0028 | 0.0098 | 0.0030 | 0.0001 | -0.0162 |
| **7** | **-1.2534** | **0.0794** | **1.0468** | **-0.3680** | **-0.2865** | **-0.0813** | **-0.2739** | **0.2739** | **0.2941** |
| **8** | **3.7569** | **0.1103** | **0.5120** | **1.3229** | **0.6042** | **-0.9409** | **0.4143** | **-0.5252** | **-0.2150** |
| **9** | **1.7852** | **0.1103** | **0.9771** | **0.6286** | **0.2871** | **-0.4471** | **0.1969** | **-0.2496** | **-0.1021** |
| 10 | 0.4002 | 0.0306 | 1.0897 | 0.0711 | -0.0108 | 0.0168 | -0.0098 | 0.0144 | -0.0103 |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 11 | -0.9352 | 0.0312 | 1.0407 | -0.1679 | -0.0799 | -0.0135 | -0.0724 | 0.0708 | 0.0925 |
| **12** | **0.3833** | **0.1195** | **1.2008** | **0.1412** | **-0.0790** | **-0.0946** | **-0.0883** | **0.0881** | **0.0772** |
| 13 | 0.5906 | 0.0714 | 1.1236 | 0.1637 | -0.0988 | 0.0457 | -0.0888 | 0.1039 | 0.0358 |
| 14 | -1.1327 | 0.0312 | 1.0135 | -0.2034 | -0.0968 | -0.0163 | -0.0877 | 0.0858 | 0.1121 |
| 15 | -1.6984 | 0.0405 | 0.9235 | -0.3488 | -0.2276 | -0.0140 | -0.2046 | 0.2117 | 0.2174 |
| 16 | -1.4468 | 0.0312 | 0.9620 | -0.2598 | -0.1236 | -0.0208 | -0.1121 | 0.1096 | 0.1432 |
| 17 | -1.0022 | 0.0274 | 1.0279 | -0.1683 | 0.0055 | -0.0228 | 0.0073 | -0.0152 | 0.0341 |
| **18** | **-0.2259** | **0.1195** | **1.2084** | **-0.0832** | **0.0465** | **0.0558** | **0.0520** | **-0.0519** | **-0.0455** |
| 19 | -1.1999 | 0.0274 | 0.9993 | -0.2014 | 0.0066 | -0.0272 | 0.0087 | -0.0182 | 0.0409 |
| 20 | 1.0302 | 0.0299 | 1.0267 | 0.1808 | -0.0646 | 0.0702 | -0.0573 | 0.0697 | 0.0059 |
| 21 | 0.7918 | 0.0318 | 1.0582 | 0.1434 | -0.0703 | 0.0342 | -0.0659 | 0.0752 | 0.0262 |
| 22 | -0.5048 | 0.0399 | 1.0934 | -0.1029 | -0.0643 | -0.0462 | -0.0627 | 0.0631 | 0.0693 |
| 23 | 0.5833 | 0.0349 | 1.0817 | 0.1108 | -0.0518 | 0.0359 | -0.0464 | 0.0551 | 0.0135 |
| 24 | -0.2621 | 0.0191 | 1.0834 | -0.0365 | -0.0010 | -0.0045 | 0.0005 | -0.0006 | 0.0053 |
| 25 | 0.1975 | 0.0314 | 1.0993 | 0.0356 | -0.0191 | -0.0029 | -0.0192 | 0.0209 | 0.0110 |
| **26** | **-0.1186** | **0.0995** | **1.1844** | **-0.0394** | **0.0149** | **0.0330** | **0.0191** | **-0.0174** | **-0.0204** |
| 27 | 0.1774 | 0.0540 | 1.1261 | 0.0424 | -0.0138 | 0.0283 | -0.0102 | 0.0140 | -0.0028 |
| 28 | 0.0975 | 0.0318 | 1.1019 | 0.0177 | -0.0087 | 0.0042 | -0.0081 | 0.0093 | 0.0032 |
| **29** | **-2.5951** | **0.0291** | **0.7192** | **-0.4492** | **0.1645** | **-0.1604** | **0.1481** | **-0.1778** | **-0.0220** |
| **30** | **0.9746** | **0.1032** | **1.1187** | **0.3306** | **0.2568** | **0.0056** | **0.2237** | **-0.2568** | **-0.1632** |
| 31 | 0.3605 | 0.0325 | 1.0941 | 0.0661 | -0.0372 | 0.0246 | -0.0336 | 0.0380 | 0.0167 |
| 32 | -0.8304 | 0.0327 | 1.0550 | -0.1528 | -0.0745 | -0.0869 | -0.0745 | 0.0754 | 0.0828 |
| 33 | -0.8828 | 0.0226 | 1.0379 | -0.1343 | -0.0712 | -0.0194 | -0.0619 | 0.0689 | 0.0567 |
| 34 | -0.1803 | 0.0197 | 1.0866 | -0.0255 | 0.0115 | -0.0017 | 0.0114 | -0.0124 | -0.0065 |
| 35 | -0.1399 | 0.0316 | 1.1009 | -0.0253 | 0.0075 | -0.0156 | 0.0056 | -0.0075 | 0.0010 |
| 36 | -1.1978 | 0.0238 | 0.9960 | -0.1871 | -0.0980 | -0.0197 | -0.0829 | 0.0950 | 0.0717 |
| 37 | -0.7449 | 0.0151 | 1.0452 | -0.0921 | -0.0066 | -0.0029 | -0.0007 | 0.0036 | 0.0052 |
| 38 | -0.8175 | 0.0185 | 1.0411 | -0.1122 | -0.0018 | 0.0484 | 0.0112 | -0.0038 | -0.0162 |
| 39 | -0.2672 | 0.0268 | 1.0918 | -0.0443 | 0.0172 | 0.0274 | 0.0221 | -0.0188 | -0.0260 |
| 40 | -0.3500 | 0.0268 | 1.0881 | -0.0581 | 0.0225 | 0.0359 | 0.0289 | -0.0246 | -0.0341 |
| 41 | 0.4155 | 0.0514 | 1.1127 | 0.0967 | -0.0773 | -0.0125 | -0.0742 | 0.0797 | 0.0576 |
| 42 | -0.7403 | 0.0197 | 1.0505 | -0.1048 | -0.0011 | -0.0458 | 0.0004 | 0.0027 | 0.0028 |
| 43 | -0.7693 | 0.0162 | 1.0438 | -0.0987 | 0.0024 | -0.0136 | 0.0083 | -0.0023 | -0.0116 |
| 44 | -0.5161 | 0.0206 | 1.0711 | -0.0749 | 0.0310 | 0.0333 | 0.0380 | -0.0332 | -0.0418 |
| 45 | -0.8513 | 0.0249 | 1.0441 | -0.1361 | 0.0128 | 0.0818 | 0.0318 | -0.0171 | -0.0542 |
| 46 | -1.1252 | 0.0171 | 1.0000 | -0.1485 | -0.0196 | 0.0088 | -0.0054 | 0.0187 | -0.0107 |
| 47 | -0.8192 | 0.0155 | 1.0377 | -0.1027 | 0.0263 | 0.0168 | 0.0341 | -0.0280 | -0.0375 |
| 48 | -0.7959 | 0.0180 | 1.0430 | -0.1079 | 0.0192 | -0.0439 | 0.0192 | -0.0182 | -0.0108 |
| 49 | -0.9715 | 0.0148 | 1.0187 | -0.1191 | 0.0061 | -0.0042 | 0.0138 | -0.0069 | -0.0182 |
| 50 | 0.0818 | 0.0550 | 1.1292 | 0.0197 | -0.0156 | -0.0056 | -0.0154 | 0.0162 | 0.0127 |
| 51 | 1.1029 | 0.0279 | 1.0144 | 0.1869 | -0.0247 | -0.0723 | -0.0442 | 0.0224 | 0.0905 |
| 52 | 0.7940 | 0.0279 | 1.0538 | 0.1345 | -0.0177 | -0.0521 | -0.0318 | 0.0162 | 0.0652 |
| 53 | 0.4326 | 0.0279 | 1.0848 | 0.0733 | -0.0097 | -0.0284 | -0.0173 | 0.0088 | 0.0355 |
| 54 | 0.5329 | 0.0300 | 1.0802 | 0.0937 | -0.0038 | 0.0428 | -0.0033 | -0.0008 | 0.0107 |
| **55** | **1.8767** | **0.0897** | **0.9350** | **0.5891** | **-0.3824** | **0.2079** | **-0.3056** | **0.3776** | **0.1730** |
| 56 | 0.6736 | 0.0383 | 1.0775 | 0.1345 | -0.0672 | 0.0567 | -0.0507 | 0.0648 | 0.0279 |
| 57 | 0.1416 | 0.0337 | 1.1033 | 0.0264 | -0.0114 | 0.0139 | -0.0097 | 0.0103 | 0.0079 |
| 58 | -0.9867 | 0.0647 | 1.0710 | -0.2595 | -0.0242 | 0.1519 | 0.0149 | 0.0245 | -0.0963 |
| 59 | -0.4790 | 0.0279 | 1.0818 | -0.0812 | 0.0107 | 0.0314 | 0.0192 | -0.0097 | -0.0393 |
| 60 | -0.6829 | 0.0545 | 1.0951 | -0.1640 | 0.0147 | 0.0954 | 0.0382 | -0.0143 | -0.0853 |
| 61 | -0.2252 | 0.0407 | 1.1091 | -0.0464 | 0.0170 | 0.0034 | 0.0207 | -0.0156 | -0.0283 |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 62 | 0.1653 | 0.0269 | 1.0950 | 0.0275 | -0.0157 | 0.0071 | -0.0136 | 0.0151 | 0.0114 |
| 63 | 0.2385 | 0.0610 | 1.1327 | 0.0608 | -0.0055 | -0.0121 | -0.0108 | 0.0030 | 0.0284 |
| 64 | 0.5592 | 0.0464 | 1.0967 | 0.1233 | -0.0241 | 0.0444 | -0.0217 | 0.0164 | 0.0376 |
| 65 | 0.1329 | 0.0411 | 1.1120 | 0.0275 | -0.0064 | 0.0038 | -0.0067 | 0.0049 | 0.0118 |
| 66 | 0.0303 | 0.0345 | 1.1056 | 0.0057 | -0.0009 | 0.0011 | -0.0008 | 0.0006 | 0.0018 |
| **67** | **-0.1228** | **0.1732** | **1.2899** | **-0.0562** | **0.0111** | **0.0513** | **0.0170** | **-0.0125** | **-0.0305** |
| 68 | 0.1631 | 0.0296 | 1.0982 | 0.0285 | -0.0067 | 0.0034 | -0.0054 | 0.0054 | 0.0092 |
| 69 | 0.2205 | 0.0395 | 1.1079 | 0.0447 | -0.0072 | 0.0229 | -0.0031 | 0.0046 | 0.0046 |
| 70 | 0.2026 | 0.0351 | 1.1033 | 0.0386 | -0.0071 | 0.0158 | -0.0040 | 0.0050 | 0.0065 |
| 71 | 0.8027 | 0.1258 | 1.1707 | 0.3045 | -0.0236 | 0.1381 | 0.0289 | 0.0196 | -0.0810 |
| **72** | **1.0786** | **0.2145** | **1.2597** | **0.5637** | **-0.1260** | **-0.2019** | **-0.0820** | **0.1434** | **0.0472** |
| 73 | 0.1678 | 0.0509 | 1.1227 | 0.0389 | -0.0044 | 0.0252 | -0.0005 | 0.0020 | 0.0006 |
| 74 | 0.0708 | 0.0457 | 1.1184 | 0.0155 | -0.0020 | 0.0092 | -0.0005 | 0.0011 | 0.0008 |
| **75** | **1.2250** | **0.1230** | **1.1039** | **0.4587** | **-0.0118** | **0.2285** | **-0.0020** | **-0.0218** | **0.0669** |
| 76 | 1.0621 | 0.0928 | 1.0932 | 0.3398 | -0.0327 | -0.0212 | -0.0519 | 0.0138 | 0.1490 |
| 77 | 0.2265 | 0.0591 | 1.1307 | 0.0567 | 0.0005 | 0.0121 | 0.0020 | -0.0042 | 0.0104 |
| **78** | **-0.8259** | **0.3097** | **1.4789** | **-0.5532** | **-0.2271** | **0.1223** | **-0.2680** | **0.2168** | **0.2449** |
| 79 | -0.5190 | 0.0877 | 1.1497 | -0.1610 | -0.0557 | -0.0540 | -0.0760 | 0.0613 | 0.0679 |
| **80** | **-1.1518** | **0.1903** | **1.2091** | **-0.5584** | **-0.1786** | **0.3170** | **-0.1807** | **0.1674** | **0.0829** |
| **81** | **-0.8274** | **0.2624** | **1.3839** | **-0.4935** | **-0.2092** | **-0.2396** | **-0.2846** | **0.2187** | **0.3211** |
| **82** | **0.1490** | **0.2007** | **1.3336** | **0.0747** | **0.0359** | **-0.0182** | **0.0340** | **-0.0386** | **-0.0086** |

***Nota: Las observaciones en negritas pueden ser influenciales según al menos uno de los diagnosticos que se describen a continuación:***

Una lista de las observaciones con al menos uno de los diagnosticos sobrepasandop los valores criticos es obtenida usando R (ver el laboratorio 10 del texto).

Potentially influential observations of
lm(formula = mpg ~ ., data = millaje) :

| | dfb.1_ | dfb.sp | dfb.wt | dfb.vol | dfb.hp | dffit | cov.r | cook.d | hat |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 1.13_* | -1.08_* | -1.20_* | 0.34_* | 1.10_* | 1.40_* | 0.51_* | 0.34 | 0.12 |
| 2 | 0.41_* | -0.39_* | -0.42_* | 0.12 | 0.39_* | 0.53 | 0.94 | 0.05 | 0.08 |
| 3 | 0.40_* | -0.39_* | -0.42_* | 0.11 | 0.39_* | 0.52 | 0.95 | 0.05 | 0.08 |
| 5 | -0.28_* | 0.27_* | 0.28_* | -0.07 | -0.26_* | -0.34 | 1.09 | 0.02 | 0.09 |
| 7 | -0.29_* | 0.27_* | 0.29_* | -0.08 | -0.27_* | -0.37 | 1.05 | 0.03 | 0.08 |
| 8 | 0.60_* | -0.53_* | -0.21 | -0.94_* | 0.41_* | 1.32_* | 0.51_* | 0.30 | 0.11 |
| 9 | 0.29_* | -0.25_* | -0.10 | -0.45_* | 0.20 | 0.63 | 0.98 | 0.08 | 0.11 |
| 12 | -0.08 | 0.09 | 0.08 | -0.09 | -0.09 | 0.14 | 1.20_* | 0.00 | 0.12 |
| 15 | -0.23_* | 0.21 | 0.22 | -0.01 | -0.20 | -0.35 | 0.92 | 0.02 | 0.04 |
| 18 | 0.05 | -0.05 | -0.05 | 0.06 | 0.05 | -0.08 | 1.21_* | 0.00 | 0.12 |
| 26 | 0.01 | -0.02 | -0.02 | 0.03 | 0.02 | -0.04 | 1.18_* | 0.00 | 0.10 |
| 29 | 0.16 | -0.18 | -0.02 | -0.16 | 0.15 | -0.45 | 0.72_* | 0.04 | 0.03 |
| 30 | 0.26_* | -0.26_* | -0.16 | 0.01 | 0.22_* | 0.33 | 1.12 | 0.02 | 0.10 |
| 55 | -0.38_* | 0.38_* | 0.17 | 0.21 | -0.31_* | 0.59 | 0.94 | 0.07 | 0.09 |
| 67 | 0.01 | -0.01 | -0.03 | 0.05 | 0.02 | -0.06 | 1.29_* | 0.00 | 0.17 |
| 72 | -0.13 | 0.14 | 0.05 | -0.20 | -0.08 | 0.56 | 1.26_* | 0.06 | 0.21_* |
| 75 | -0.01 | -0.02 | 0.07 | 0.23_* | 0.00 | 0.46 | 1.10 | 0.04 | 0.12 |
| 78 | -0.23_* | 0.22 | 0.24_* | 0.12 | -0.27_* | -0.55 | 1.48_* | 0.06 | 0.31_* |
| 80 | -0.18 | 0.17 | 0.08 | 0.32_* | -0.18 | -0.56 | 1.21_* | 0.06 | 0.19_* |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 81 | -0.21 | 0.22 | 0.32_* | -0.24_* | -0.28_* | -0.49 | 1.38_* | 0.05 | 0.26_* |
| 82 | 0.04 | -0.04 | -0.01 | -0.02 | 0.03 | 0.07 | 1.33_* | 0.00 | 0.20_* |

De acuerdo a los residuales estudentizados internamente o externamente, las observaciones 1, 8 y 29 son "outliers".

De acuerdo a los valores leverages $h_{ii}$, las observaciones 72, 78, 80, 81 y 82 son puntos leverages, pues tiene $h_{ii}>3p/n=.1829$

De acuerdo a la Distancia Cook no hay ninguna observación que tenga gran influencia pues todos los $CD_i^2$ son menores que 1, más aún son menores que F(0.50,5,77)=0.878. Sin embargo, las observaciones 1 y 8 tienen un $CD_i^2$ mucho mayor que las otras y deberían ser consideradas cuidadosamente.

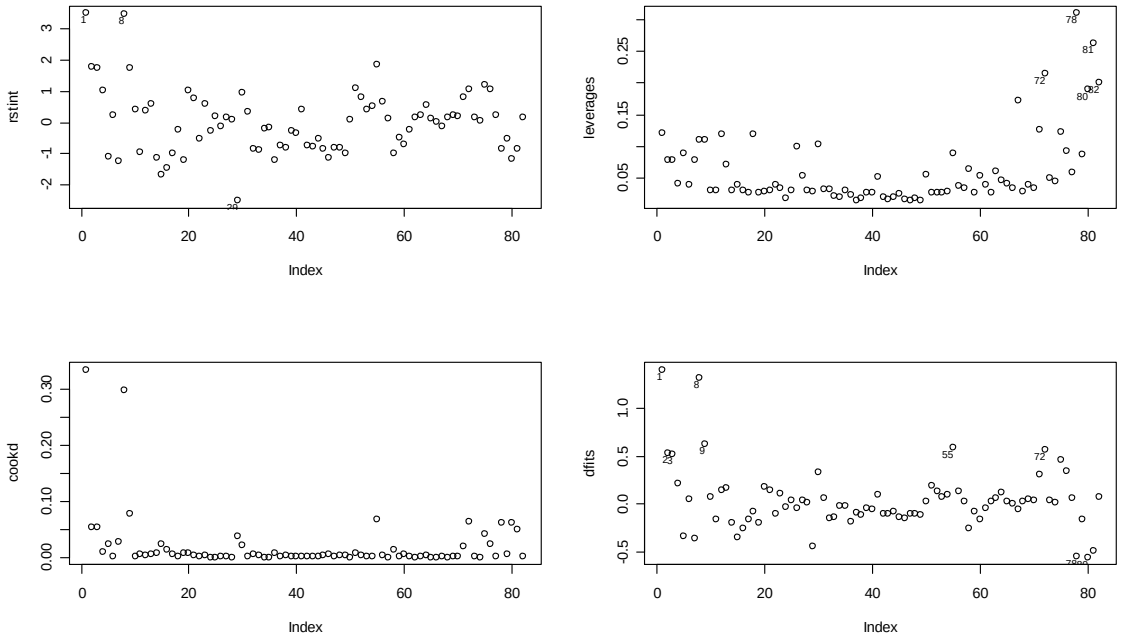


Figura 3.3. Plot de varios diagnosticos de regresión para cada observación del ejemplo 1.

De acuerdo al DFFITS serían influenciales las observaciones 1, 2, 3, 8, 9, 55, 72 ,78 y 80, pues su $|DFFITS_i|> 2\sqrt{\frac{5}{82}} =.49386$

De acuerdo al COVRATIO serían influenciales las observaciones 1, 8, 12, 18, 26, 29, 72, 78, 80, 81 y 82, pues su COVRATIO>1.1829 ó <0.8171 Los COVRATIO de las observaciones 12, 18, 26 y 29 están bastante cerca de los puntos de corte.

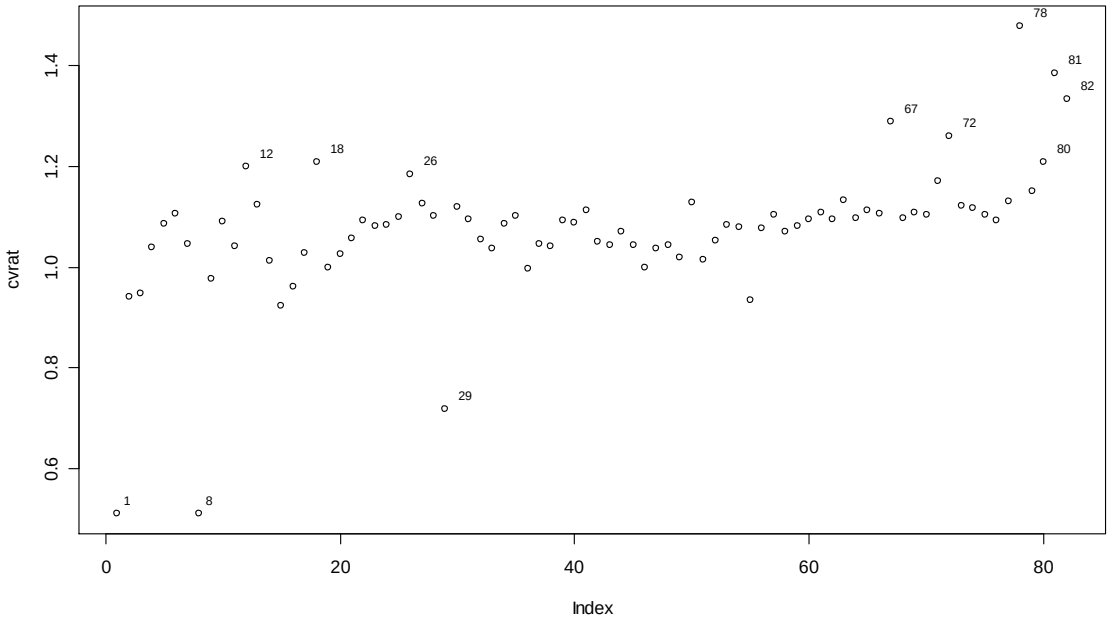


Figura 3.4. Plot de COVRATIOS para cada observación del ejemplo 1.

De acuerdo a los DFBETAS una observación es influencial si su valor absoluto es mayor $2/\sqrt{82}$ =0.22086. Las observaciones 1, 8, 9, 80 y 81 parecen afectar el comportamiento del coeficiente $\beta_1$, el valor DFBETAS de la observación 75 está muy cerca del punto de corte y no ha sido considerado. Las observaciones 1, 2, 3, 5, 7, 8, 30, 55, 78 y 81 afectan el comportamiento de $\beta_2$, en tanto que 1, 2, 3, 5, 7, 8, 9, 30, y 55 parecen tener influencia en $\beta_3$ y las observaciones 1, 2, 3, 5, 7, 78 y 81 afectan el comportamiento de $\beta_4$.

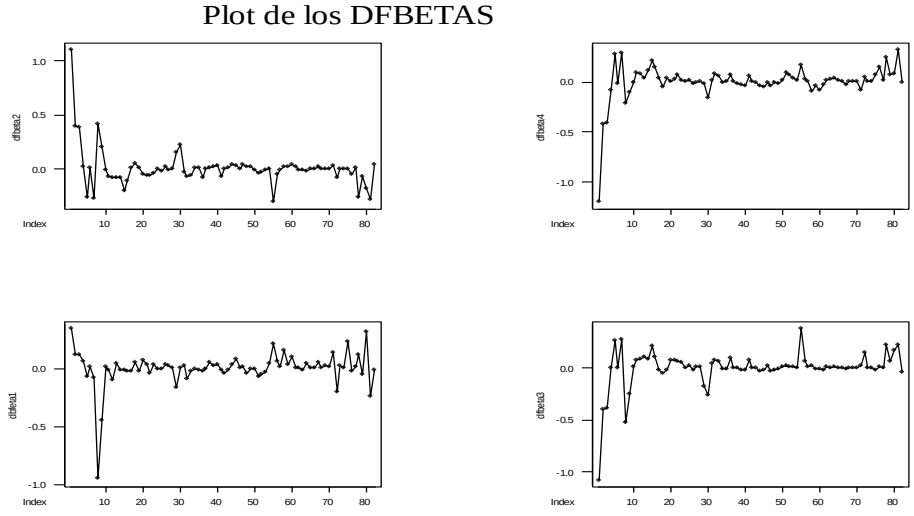


Figura 3.5. Plot de los DFBETAS para el conjunto datos millaje.

En conclusión , las observaciones 1, 2, 3, 5, 7, 8, 9, 30, 55, 72, 78,, 80, 81 y 82 parecen ser las mas influenciales.

## 3.4 Plot de Residuales para detectar el efecto de variables y casos influenciales

Existen ciertos plot de residuales que se usan para estudiar el efecto de añadir una nueva variable predictora en un modelo. Estos plots tambien permiten detectar la presencia de casos influenciales. Supongamos que queremos ver la importancia de la variable predictora $x_j$. Consideremos el modelo

$$Y=X_{-j}B_{-j}+\beta_j x_j +e$$

Donde $X_{-j}$ es la matriz X sin incluir la columna j. Se puede mostrar que

$$\hat{\beta}_j = \frac{\mathbf{x}_j'(\mathbf{I}-\mathbf{H}_{-j})\mathbf{Y}}{\mathbf{x}_j'(\mathbf{I}-\mathbf{H}_{-j})\mathbf{x}_j} \qquad (3.15)$$

Definamos los siguientes residuales

a) $\hat{e}_{Y/X_{-j}} = (I-H_{-j})Y$

b) $\hat{e}_{Y/x,X_{-j}} = (I-H)Y$

c) $\hat{e}_{x_j/X_{-j}} = (I-H_{-j})X_j$

en el caso a) se han considerado en el modelo todas las predictoras excepto $x_j$, en el caso b) están consideradas todas las variables predictoras y en el caso c) son los residuales de la regresión de $x_j$ versus las otras variables no consideradas en el modelo.

Hay cuatro tipo de plots de residuales que permiten ver el impacto de cada variable predictora $x_j$ en el modelo. Estos son:

a) Plot de Residuales versus las variables predictoras
b) Plot de regresión parcial (o plot de variable añadida)
c) Plot de residuales parciales
d) Plot de residuales parciales aumentados.

**a) Plot de residuales versus la variables predictoras.**

Aquí se plotea

$\hat{e}_{Y/x,X_{-j}}$ versus $x_j$

Si el modelo es adecuado los puntos se deberían alinear a lo largo de una franja horizontal. Si se observa algún patrón no lineal entonces la variable predictora debería ser transformada.

Para el ejemplo de Millaje estos son los plots.de residuales que resultan.

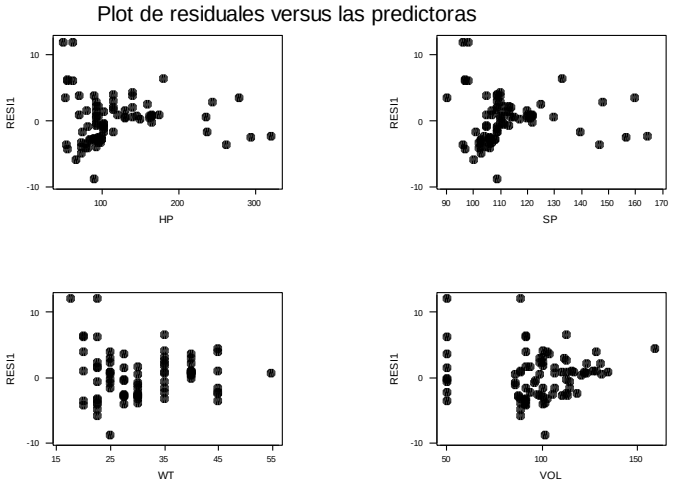


Figura 3.6. Plot de residuales versus las predictoras para el conjunto de datos millaje

Lo que más se destaca en estas gráficas es la presencia de varios valores "outliers" y puntos de leverage alto. También se observa que no todos los puntos se alinean uniformemente alrededor del eje horizontal 0, pero es dificil detectar la tendencia para usar una transformación nolineal.

## b) Plots de regresión parciales ( o plot de variable añadida)

Aquí se plotea los residuales $\widehat{e}_{Y/X_{-j}}$ versus $\widehat{e}_{x_j/X_{-j}}$

En el plot de regresión parcial se plotea los residuales de la regresión de y considerando todas las variables predictoras excepto $x_j$ versus los residuales de la regresión de $x_j$ contra todas las variables predictoras distintas a ella.

Si la variable $x_j$ entra al modelo en forma lineal entonces su plot de regresión parcial debería mostrar una tendencia lineal que pasa por el origen. Si se observa una tendencia no lineal habría que considerar una transformación de $x_j$. También se puede localizar a los puntos que afecta el cálculo del coeficiente de regresión correspondiente.

Consideremos por ejemplo el plot de regresión parcial para la variable HP del conjunto de datos Millaje, donde se asume que el modelo contiene ya las otras 3 variables predictoras.

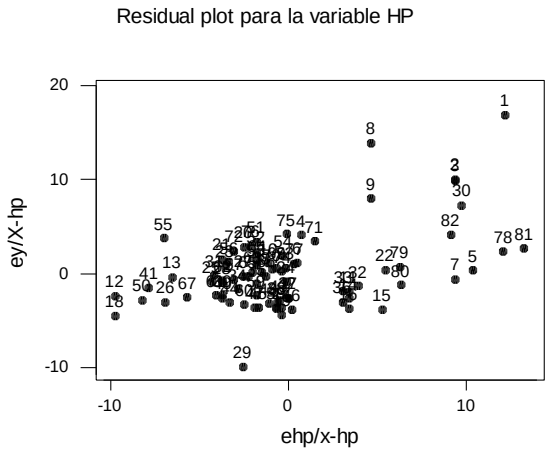


Figura 3.7. Plot de regresión parcial considerando la variable predictora HP

Se observan muchos valores influenciales y la tendencia lineal es bien pobre. Si usamos 1/HP en lugar de HP la cosa no mejora mucho.
En realidad el efecto de este plot se observa mejor si consideramos primero la regresión de MPG con la variable VOL, que es la que tiene menos correlación y si consideramos luego añadir WT. El plot de regresión parcial que se obtiene es

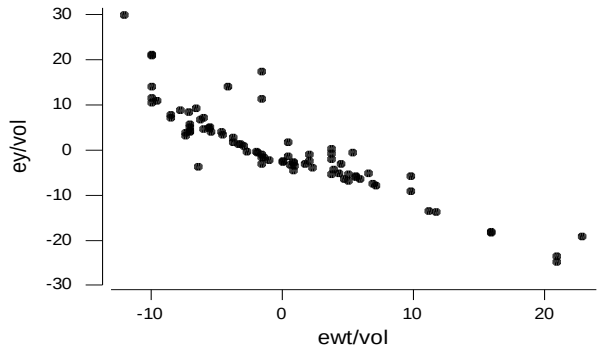


Figura 3.8. Plot de regresión parcial considerando la variable WT asumiendo que el modelo solo contiene a VOL.

Se puede observar que hay bastante linealidad en el plot, y que la linea estimada pasaría por el origen. Luego, se debería usar una regresión lineal múltiple con dos variables predictoras.
Consideremos la misma situación anterior pero en lugar de WT ahora queremos incluir HP. El plot de la Figura 3.7 ya no se ve tan lineal sino parece como una rama de una hipérbola equilatera.

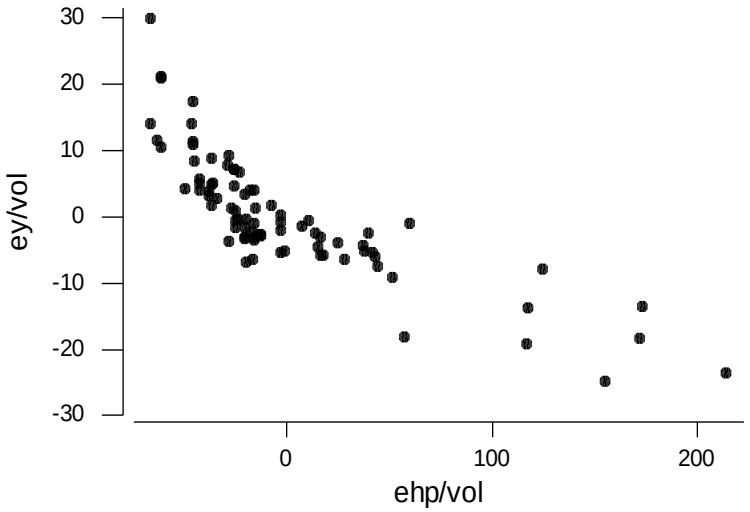


Figura 3.9. Plot de regresión parcial considerando la variable HP asumiendo que el modelo solo contiene a VOL.

Así que sería mejor usar 1/HP en lugar de HP en el modelo

**c) Plot de residuales parciales o de residuales más componente**

Aquí se plotea
$\hat{e}_{Y/x,X_{-j}} + x_j\beta_j$ versus $x_j$

Es más efectivo para detectar nolinealidad que el plot de regresión parcial No es muy adecuado para detectar casos influenciales.

**d) Plot de residuales parciales aumentados**

Aquí se plotea

$$\hat{e}_{y/X_{-j},x_j^2} + x_j\beta_j + x_j^2\beta_{jj} \text{ versus } x_j$$

Este plot fue propuesto por Mallows (1986) y es el más adecuado para cotejar si la variable $x_j$ debe entrar en forma cuadrática al modelo.

## 3.5 Plot de residuales para detectar Normalidad

La suposición de la normalidad de los errores es bién importante para el proceso de hacer inferencia en regresión lineal múltiple. Al igual que en regresión lineal simple esto puede ser cotejado haciendo un plot de normalidad para los errores estudentizados internamente.

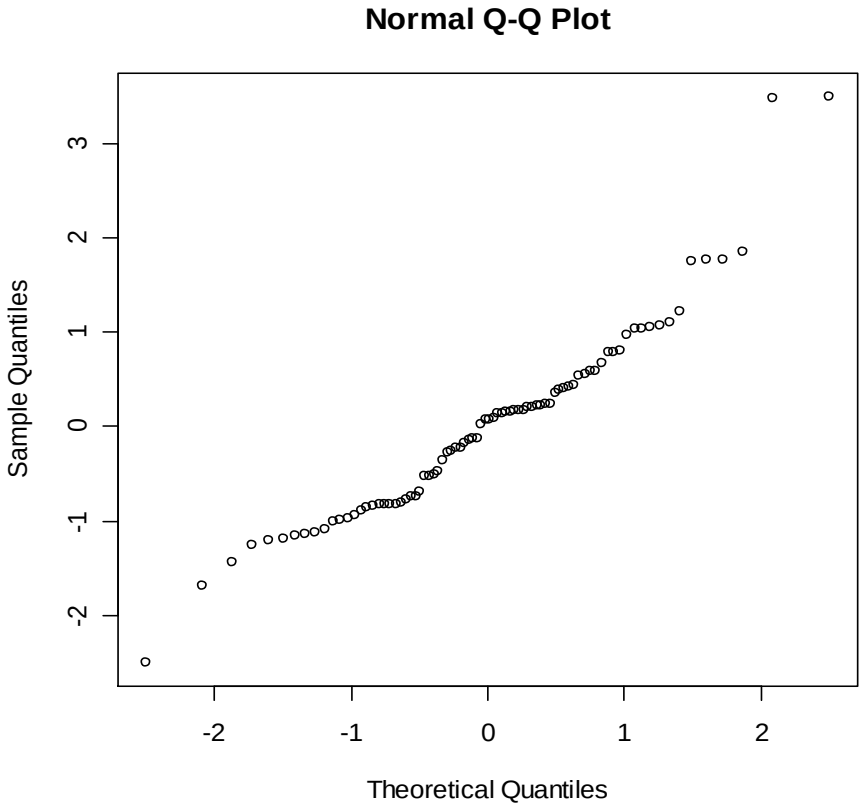


Figura 3.10. Plot de normalidad para los residuales del conjunto de datos millaje.

El plot de Normalidad consiste en un plot de los **scores normales** (estadísticos de orden normales) versus los residuales estandarizados ordenados. Los scores normales representan los valores esperados de observaciones ordenadas que provienen de una distribución normal estándar. El i-ésimo score normal es aproximado en forma bastante precisa por

$$z_{(i)} = \Phi^{-1}(\frac{i-3/8}{n+1/4})$$

donde Φ representa la función de distribución acumulada de una normal estándar y n (n>5) es el número de observaciones en la muestra.
Para que haya normalidad los puntos deben estar alineados alrededor de una recta que pasa por el origen. En la figura 3.8 se muestra el plot de normalidad de los residuales correspondientes a la regresión lineal multiple del conjunto de datos **millaje**, se observa que los puntos están bastante alineados pero se observan varios "outliers" en ambos extremos de la distribución.
Si la tendencia de los puntos es curveada entonces la distribución es asimétrica. El tipo de asimetría es determinada por el lado donde está la parte curveada. Un plot de normalidad que produce una curva en forma de S indica que la distribución tiene una cola pesada o liviana dependiendo de la forma de la S. Si la S es alargada entonces la cola es liviana.

También se podría aplicar una prueba noparamétrica como la de Kolmogorov-Smirnov o Shapiro-Wilks para detectar normalidad.

```
> l1=lm(mpg~.,data=millaje)
> resi=rstandard(l1)
> ks.test(resi,"pnorm")

        One-sample Kolmogorov-Smirnov test

data:  resi
D = 0.0881, p-value = 0.519
alternative hypothesis: two-sided

> boxplot(resi)
> shapiro.test(resi)

        Shapiro-Wilk normality test

data:  resi
W = 0.945, p-value = 0.001542
```

El "p-value" de la prueba de Kolmogorov-Smirnov es mayor que 0.05 por lo tanto se acepta la hipótesis de que hay normalidad de los residuales. Sin embargo, la prueba de Shapiro-Wilks indica que no hay normalidad puesto que el "p-value" de la prueba es pequeño.

## 3.6 Detectando varianza no constante

La suposición de que en el modelo de regresión lineal múltiple, los errores tienen varianza constante es importante para que los estimadores mínimos cuadráticos sean óptimos. Por lo general varianza no constante viene acompañado del hecho que no hay normailidad.

Para detectar si la varianza es constante o no se hace un plot de resiiduales estudentizados versus los valores ajustados $\hat{y}_i$'s. Si los puntos aparecen alineados arbitrariamente alrededor de una franja horizontal centrada en la linea horizontal en cero entonces hay indicación de varianza constante (homocedasticidad). Si los puntos forman algún tipo de patrón como el que se muestra en la figura 3.9 entonces indica una violación de la suposición de homocedasticidad. Aquí la varianza varía en forma proporcional a la media de la variable de respuesta Y. Este plot es típico cuando los errores siguen una distribución Poisson o log-normal.

Algunas veces la varianza puede variar de acuerdo a los variables de una variable predictora. Para detectar esta situación hay que hacer un plot de residuales versus cada variable predictora.
Si hay indicación de que la varianza poblacional $\sigma^2$ no es constante entonces hay dos remedios posibles:

i) Usar mínimos cuadrados ponderados donde los pesos que se usan son hallados en base a los datos tomados.
ii) Transformar la variable de respuesta Y usando tranfomación que estabiliza la varianza

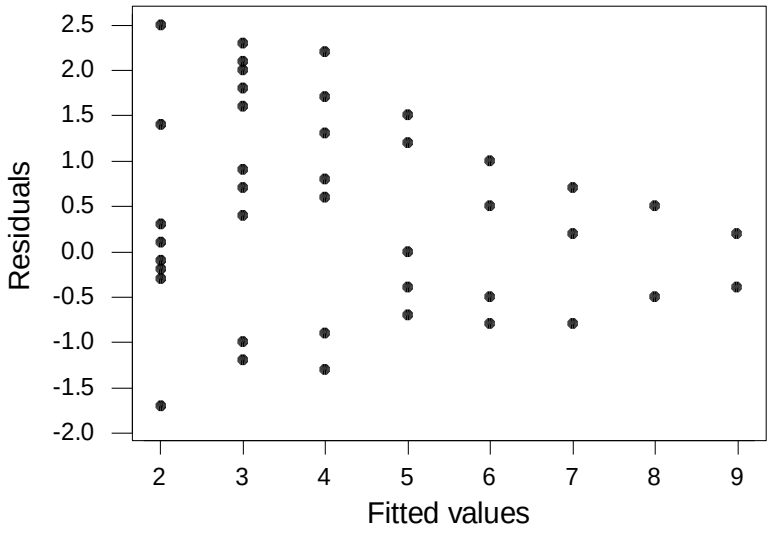


Figura 3.11. Este plot muestra que la varianza de los errores no es constante y que varia en forma proporcional a la media de la variable de respuesta

Las medidas remediales para varianza no constante serán discutidas en el capítulo 4 del texto.

## 3.7 Errores correlacionados en Regresión

Una de las suposiciones que se hace en regresión lineal es que Cov($e_i, e_j$)=E($e_i e_j$)=0 para $i \neq j$ .Es decir que los errores no se correlacionan entre si . Hay un caso en regresión cuando la variable predictora es tiempo, donde puede haber una dependencia del comportamiento con respecto al tiempo . Por ejemplo, las ventas de una compañia de ropas pueden seguir un patrón que depende de la epoca del año y pudiera ocurrir entonces que que E($e_i, e_{i+k}$) $\neq 0$ para un cierto k en este caso se dice que los errores tiene una correlación serial y estan autocorrelacionados . Si se gráfica los residuales versus la variable predictora tiempo y se observa mucho cambio de signo entonces la autocorrelación es negativa si el cambio de signo no es muy frecuente entonces la autocorrelacion es positiva.

**Ejemplo 2.** Consideremos las siguientes series de tiempo

| year | y1 | y2 | y3 |
|---|---|---|---|
| 1 | 10 | 15 | 5 |
| 2 | 20 | 10 | 10 |
| 3 | 35 | 18 | 15 |
| 4 | 50 | 12 | 18 |
| 5 | 12 | 24 | 20 |
| 6 | 24 | 15 | 32 |

| | | | |
|---|---|---|---|
| 7 | 40 | 32 | 37 |
| 8 | 50 | 18 | 39 |
| 9 | 14 | 40 | 42 |
| 10 | 25 | 20 | 35 |
| 11 | 40 | 55 | 30 |
| 12 | 60 | 25 | 27 |

cuyas graficas aparecen en la siguiente figura

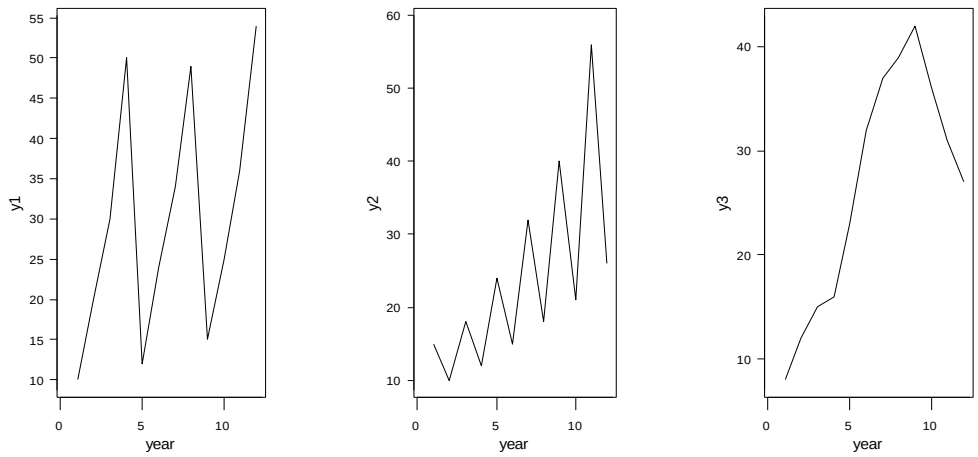


Figura 3.12. Gráfica de las 3 series de tiempo del ejemplo

En los dos primeros plots la autocorrelación es negativa y en la última es positiva
Los plots de residuales correspondientes se muestran en la figura 3.13

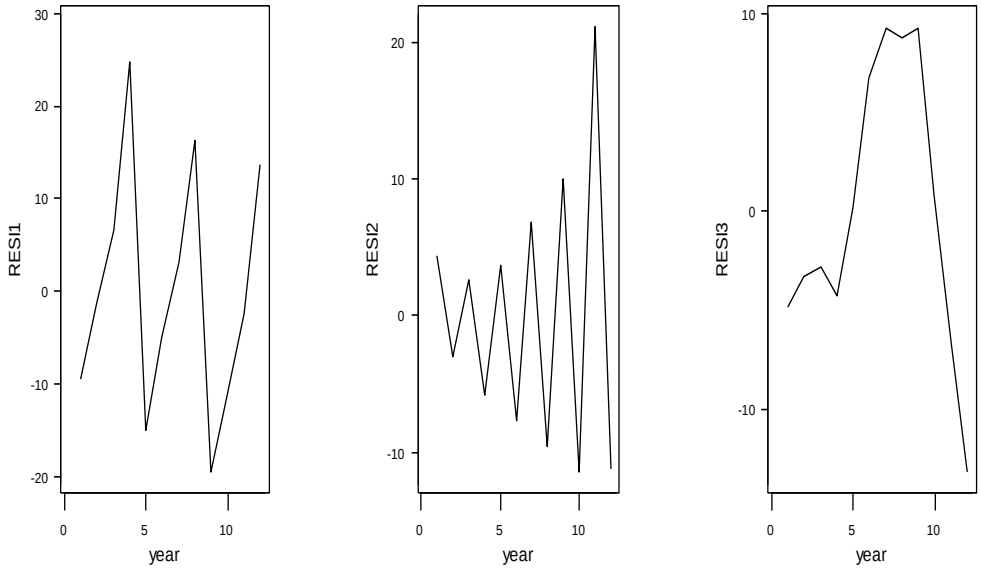


Figura 3.13. Plot de los residuales de las tres series de tiempo del ejemplo 2

Cuando los residuales cambian frecuentemente de signo hay autocorrelación negativa y si hay un conglomerado de residuales de un mismo signo antes de cambiar a otro entonces la autocrrelación es positiva. Lo anterior se puede observar más claramente si se plotea los residuales en el tiempo t versus los residuales en el tiempo t-1.

El Modelo autorregresivo de primer orden para los errores se define

$$e_t=\rho e_{t-1}-u_t$$

donde se supone que las $u_t$ son variables aleatorias distrbuidas normalmente con media 0 y varianza constante.

La prueba de Durbin-Watson se usa para detectar si hay correlación positiva. Es decir para probar Ho: $\rho=0$ vs Ha: $\rho>0$.
La prueba está dada por

$$D=\frac{\sum_{t=2}^{n}(e_t-e_{t-1})^2}{\sum_{t=1}^{n}e_t^2}$$

Existen tablas de la prueba, que dan valores límites DL y DU para que se rechaze la hipótesis nula. La tabla considera número de casos:(n), número de predictoras y tres niveles de significación. Las decisiones se toman así: Se rechaza Ho si D<DL, se acepta Ho si D>DU y la prueba no lleva a ninguna conclusión si DL<D<DU.

La prueba de Durbin-Watson no detecta autocorrelaciones de segundo orden o mayor. La librería **car** de R tiene una función **durbinWatsonTest** que calcula este estadístico; El estadistico es dado por muchos programas entre ellos MINITAB.

Para los datos cuyas graficas de residuales aparecen en la figura 3.11 se obtienen los siguientes resultados para el estadístico de Durbin-Watson.

```
> l1=lm(y1~year,data=corrdata)
> dw(l1$residuals)
[1] 1.835639
> l2=lm(y2~year,data=corrdata)
> dw(l2$residuals)
[1] 3.537003
> l3=lm(y3~year,data=corrdata)
> dw(l3$residuals)
[1] 0.4571476
>
```

Buscando en la tabla de Durbin-Watson con n=12 (aprox con n=15), k=1 y $\alpha$=0.5 resulta ser que $D_L$=1.08 y $D_U$=1.36, por lo tanto, no se rechaza la hipótesis nula en los dos primeros casos ya que DW=1.835 y DW=3.53 respectivamente son mayores que 1.36, y se concluye que no hay autocorrelación de primer orden entre los errores. En el último caso si se rechaza la hipótesis nula puesto que DW=.457 < 1.08 y se concluye que hay autocorrelación positiva de primer orden.

Si se desea probar una hipótesis de dos lados Ho: $\rho$=0, versus Ha:$\rho\neq 0$ entonces se rechaza Ho: si $D<D_L$ ó $4-D<D_L$, al nivel de significación de $2\alpha$. Si $D>D_U$ y $4-D>D_U$ entonces no se rechaza. Para cualquier otro valor de D la prueba no llega a ninguna conclusión.

Una regla práctica es que cuando el estadístico de Durbin-Watson sale cerca de 2 entonces es probable que no hay autocorrelación.

Si hubiera autocorrelación positiva de primer orden entonces una forma de resolver el problema sería considerar el modelo

$$y_t = \beta_o + \beta_1 t + \beta_2 y_{t-1} + e_t$$

donde $y_{t-1}$ son los valores de la variable de respuesta en el tiempo anterior. Estos modelos son llamados modelos de series de tiempo y son discutidos en textos especializados.

# Ejercicios

1. Probar las siguientes identidades

i) $\hat{\boldsymbol{\beta}}_{(\mathbf{i})} = \hat{\boldsymbol{\beta}} - \dfrac{(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x_i}\hat{e}_i}{1-h_{ii}}$

ii) $s^2_{(i)} = \dfrac{n-p-1}{n-p-2}s^2 - \dfrac{\hat{e}_i^2}{(n-p-2)(1-h_{ii})}$

2. Probar las siguientes relaciones

i) $t_i = \dfrac{\widehat{e}_i}{s_{(i)}\sqrt{1-h_{ii}}}$

ii) $t_i = r_i^*\left(\dfrac{n-p-2}{n-p-1-r_i^{*2}}\right)^{1/2}$

3. Probar la siguiente formula equivalente para calcular COVRATIOS

$$(COVRATIO)_i = \left(\frac{s^2_{(i)}}{s^2}\right)^p \frac{1}{(1-h_{ii})}$$

4. Usar el conjunto de datos **Fuel** con variable de respuesta: Fuel y las predictoras TAX, DLIC, INC y ROAD para responder a las siguientes preguntas. Los datos están disponibles en la página de internet del texto.

a) Cotejar las suposiciones del modelo de regresión multiple mediante un plot de residuales.
b) Determinar outliers y puntos con leverage alto usando los diagnosticos de regresión.
c) Usar plot de residuales para evaluar el efecto de añadir la segunda variable predictora con la correlación más alta con Fuel al modelo que ya tiene incluido la variable más altamente correlacionada con Fuel.

5. Usar el conjunto de datos **Headcirc** con variable de respuesta: headcirc (circunferencia de la cabeza del bebe) para responder a las siguientes preguntas. Los datos están disponibles en la página de internet del texto

a) Cotejar las suposiciones del modelo de regresión multiple mediante un plot de residuales.
b) Determinar outliers y puntos con leverage alto usando los diagnósticos de regresión.
c) Usar plot de residuales para evaluar el efecto de añadir la segunda variable predictora con la correlación más alta con Headcirc al modelo que ya tiene incluido la variable más altamente correlacionada con Headcir.

6. Usar el conjunto de datos **Grasa** con variable de respuesta: grasa (porcentaje de grasa en el cuerpo) para responder a las siguientes preguntas. Los datos están disponibles en la página de internet del texto.

a) Cotejar las suposiciones del modelo de regresión multiple mediante un plot de residuales.
b) Determinar outliers y puntos con leverage alto usando los diagnósticos de regresión.
c)Usar plot de residuales para evaluar el efecto de añadir al modelo con la variable predoctora más altamente correlacionada con grasa, la segunda variable más talamente correlacionada con grasa.

7. Supongamos que ajustamos un modelo de regresion múliple con intercepto y se define la distancia (cuadrada) Mahalanobis de $\mathbf{x_i}=(x_{i,1},\ldots x_{i,p})$ a $\bar{\mathbf{x}} = \sum_{i=1}^{n} \frac{\mathbf{x}_i}{n}$ por

$D_i^2 = \mathbf{(x_i - \bar{x})' C^{-1} (x_i - \bar{x})}$, donde $\mathbf{C^{-1}}$ es la inversa de la matriz de covarianza de las x's.

Establecer una relación entre los valores leverages $h_{ii}$ y $D_i^2$.