# Capítulo 7

# Residuales

## 7.1. Introducción

Los diagramas de dispersión son utilizados para tomar algunas decisiones antes de estimar un modelo de regresión. Por otro lado los *diagnósticos de regresión* son usados luego de estimar el modelo para verificar si la función media o los supuestos son consistentes con la data observada. Los estadísticos usados se calculan usando los residuales. Si el modelo estimado tiene un conjunto de residuales que no aparenta ser razonable entonces algunos aspectos del modelo, quizás la función media o los supuestos hechos sobre la función variancia, deben ser revisados. Se podría tener interés en la importancia de cada caso en el proceso de estimación. En algunos conjuntos de datos, las estadísticas pueden cambiar de manera sustancial si uno de los casos es omitido de la data. Tal caso es llamado *influencial*, y debe ser detectado.

## 7.2. Los residuales

El modelo básico de regresión lineal múltiple esta dado por:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e} \qquad \text{Var}(\mathbf{e}) = \sigma^2\mathbf{I} \tag{7.2.1}$$

donde $\mathbf{X}$ es una matriz conocida con $n$ filas y $p^{'}$ columnas para una función media incluye un intercepto. Se asume que $\mathbf{X}$ que tiene rango completo, es decir que $(\mathbf{X}^{'}\mathbf{X})^{-1}$ existe. Sin embargo lo anterior no representa una limitación importante en un modelo de regresión ya que siempre es posible eliminar términos de la función media tal que se logre tener rango completo.

El vector de $p^{'} \times 1$ parámetros desconocidos es $\boldsymbol{\beta}$ y el vector de errores no observables es $\mathbf{e}$. Se asume que estos errores tienen la misma variancia y no están correlacionados.

Para el modelo 7.2.1 los valores estimados $\hat{\mathbf{Y}}$ están dados por:

$$\hat{\mathbf{Y}} = \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}(\mathbf{X}^{'}\mathbf{X})^{-1}\mathbf{X}^{'}\mathbf{Y} = \mathbf{H}\mathbf{Y} \tag{7.2.2}$$

donde $\mathbf{H}$ es la matriz $n \times n$ definida por:

$$\mathbf{H} = \mathbf{X}(\mathbf{X}^{'}\mathbf{X})^{-1}\mathbf{X}^{'} \tag{7.2.3}$$

y es llamada la *matriz hat*, ya que transforma el vector de respuestas observadas $\mathbf{Y}$ en el vector de respuestas estimadas $\hat{\mathbf{Y}}$. El vector de residuales se define por:

$$\begin{aligned} \hat{\mathbf{e}} &= \mathbf{Y} - \hat{\mathbf{Y}} \\ &= \mathbf{Y} - \mathbf{X}(\mathbf{X}^{'}\mathbf{X})^{-1}\mathbf{X}^{'}\mathbf{Y} \\ &= (\mathbf{I} - \mathbf{H})\mathbf{Y} \end{aligned} \tag{7.2.4}$$

### 7.2.1. Diferencias entre ê y e

Los errores son variables aleatorias no observables, se asume que tienen media cero, variancia constante $\sigma^2$ y no se encuentran correlacionados. Los residuales $\hat{\mathbf{e}}$ son cantidades calculadas usando el modelo de regresión estimado. Su media y variancia son:

$$\mathrm{E}(\hat{\mathbf{e}}) = 0 \qquad \mathrm{Var}\,(\hat{\mathbf{e}}) = \sigma^2(\mathbf{I} - \mathbf{H}) \tag{7.2.5}$$

Los residuales también tienen media cero, sin embargo su varianza no es constante y se encuentran correlacionados. A partir de la ecuación 7.2.4 se puede observar que los residuales son combinaciones lineales de los valores observados. Si los últimos se encuentran normalmente distribuídos también lo estarán los residuales. Si se incluye el intercepto en el modelo entonces las suma de los residuales es cero, $\hat{\mathbf{e}}^{'}\mathbf{1} = \sum \hat{e}_i = 0$. La variancia del $i$-ésimo residual es:

$$\mathrm{Var}(\hat{e}_i) = \hat{\sigma}^2(1 - h_{ii}) \tag{7.2.6}$$

donde $h_{ii}$ es el $i$-ésimo elemento de la diagonal de $\mathbf{H}$ y es llamado *leverage*. Las medidas de diagnóstico están basadas en los residuales cuyo comportamiento debería ser parecido al de los errores. Lo anterior depende de la matriz hat que relaciona ambas cantidades y al mismo tiempo determina las variancias y covariancias de los residuales.

### 7.2.2. La matriz hat

La matriz $\mathbf{H}$ es simétrica de dimensión $n \times n$ y tiene muchas propiedades interesantes. Se cumple que $\mathbf{HX} = \mathbf{X}$ y también $(\mathbf{I}-\mathbf{H})\mathbf{X} = \mathbf{0}$. La matriz hat es *idempotente*, es decir $\mathbf{H}^2 = \mathbf{H}$. La covariancia entre los valores estimados y los residuales es:

$$\text{Cov}(\hat{\mathbf{Y}}, \hat{\mathbf{e}}) = \text{Cov}(\mathbf{HY}, (\mathbf{I} - \mathbf{H})\mathbf{Y}) = \sigma^2\mathbf{H}(\mathbf{I} - \mathbf{H}) = \mathbf{0}$$

Otro nombre para $\mathbf{H}$ es *proyección ortogonal* del espacio de columnas de $\mathbf{X}$. Los elementos $h_{ij}$ están dados por:

$$h_{ij} = \mathbf{x}_i^{'}(\mathbf{X}^{'}\mathbf{X})^{-1}\mathbf{x}_j = \mathbf{x}_j^{'}(\mathbf{X}^{'}\mathbf{X})^{-1}\mathbf{x}_i = h_{ji} \tag{7.2.7}$$

Se pueden obtener muchas relaciones útiles para los $h_{ij}$. Por ejemplo:

$$\sum_{i=1}^{n} h_{ii} = p^{'} \tag{7.2.8}$$

y si la función media incluye un intercepto:

$$\sum_{i=1}^{n} h_{ij} = \sum_{j=1}^{n} h_{ij} = 1 \tag{7.2.9}$$

Cada elemento de la diagonal $h_{ii}$ esta acotado debajo por $1/n$ y arriba por $1/r$, donde $r$ es el número de filas de $\mathbf{X}$ que son iguales a $\mathbf{x}_i$.

### 7.2.3. Residuales estudentizados internamente

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

$$r_i = \frac{\hat{e}_i}{\sqrt{\hat{\sigma}^2(1 - h_{ii})}} \tag{7.2.10}$$

### 7.2.4. Residuales estudentizados externamente

Suponga que la $i-$ésima observación es eliminada del conjunto de datos y que se ajusta el modelo lineal con las $n-1$ observaciones restantes obteniendo

$\hat{\boldsymbol{\beta}}_{(i)}$ el vector estimado de parámetros y $s^2_{(i)}$ la estimación de la varianza de los errores. El *residual estudentizado externamente* se define por:

$$t_i = \frac{y_i - \hat{y}_{(i)}}{s_{(i)}\sqrt{1 + \mathbf{x}_i'(\mathbf{X}_{(i)}'\mathbf{X}_{(i)})^{-1}\mathbf{x}_i}} \tag{7.2.11}$$

donde $\hat{y}_{(i)} = \mathbf{x}_i'\hat{\boldsymbol{\beta}}_{(i)}$. Se puede probar que:

$$y_i - \hat{y}_{(i)} = \frac{\hat{e}_i}{1 - h_{ii}} \tag{7.2.12}$$

### 7.2.5. Residuales y la matriz hat con pesos

Cuando $\text{Var}(\mathbf{e}) = \sigma^2\mathbf{W}^{-1}$ todos los resultados vistos en esta sección requieren cierta modificación. Una versión útil de la matriz hat esta dada por:

$$\mathbf{H} = \mathbf{W}^{1/2}\mathbf{X}(\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{W}^{1/2}\mathbf{X}' \tag{7.2.13}$$

y los leverages son los elementos de la diagonal de esta matriz. Los valores estimados están dados por $\hat{\mathbf{Y}} = \mathbf{X}\hat{\boldsymbol{\beta}}$, donde $\hat{\boldsymbol{\beta}}$ es el estimador por mínimos cuadrados ponderados.

Se definen los *residuales por mínimos cuadrados ponderados* por:

$$\hat{e}_i = \sqrt{w_i}(y_i - \hat{y}_i) \tag{7.2.14}$$

Cuando todos los pesos son iguales a uno, 7.2.14 se reduce a 7.2.4. Para los gráficos y cualquier medida de diagnóstico se debería usar 7.2.14 para definir los residuales. Algunos programas de computadora usan los residuales no ponderados en lugar de 7.2.14 por defecto.

No existe un nombre consistente para estos residuales. Por ejemplo, en R los residuales definidos en 7.2.14 son llamados *Pearson residuals* en algunas funciones y *weighted residuals* en otros.

## 7.3. Medidas de influencia

Se consideran algunos indicadores que sirven para detectar si una observación es considerado un posible valor influencial. Los indicadores más básicos son:

- Si $|h_{ii}| > \frac{2p^{'}}{n}$, donde $p^{'}$ es el número de parámetros, entonces la $i-$ésima observación podría ser influencial.
- Si $|t_i| > 2$ entonces la $i-$ésima observación podría ser influencial.

Los indicadores más sofisticados son:

- *La distancia de Cook* que mide el cambio que ocurriría en el vector de coeficientes estimados de regresión si la $i-$ésima observación fuese omitida. Se calcula por:

$$\mathcal{DC}_i^2 = \frac{(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(i)})^{'}\mathbf{X}^{'}\mathbf{X}(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(i)})}{p^{'}s^2} = r_i^2\frac{h_{ii}}{p^{'}(1-h_{ii})} \tag{7.3.1}$$

  Si $\mathcal{DC}_i^2 > 1$ entonces la $i-$ésima observación es potencialmente influencial.

- *DFFITS* es un indicador similar a la distancia de Cook y sus valores son obtenidos usando los residuales estudentizados externamente:

$$\mathcal{DFF}_i^2 = t_i^2\frac{h_{ii}}{(1-h_{ii})} \tag{7.3.2}$$

  Si $|\mathcal{DFF}_i| > 2\sqrt{\frac{p^{'}}{n}}$ indica un posible valor influencial.

- *DFBETAS* que mide la influencia de la $i-$ésima observación en cada uno de los coeficientes de regresión. Se calcula por:

$$\mathcal{DFB}_{ji} = \frac{\beta_j - \beta_{j(i)}}{s_{(i)}\sqrt{c_{jj}}} \tag{7.3.3}$$

  donde $c_{jj}$ es el $j-$ésimo elemento de la diagonal de $(\mathbf{X}^{'}\mathbf{X})^{-1}$. Si $|\mathcal{DFB}_{ji}| > \frac{2}{\sqrt{n}}$ la $i-$ésima observación es un posible valor influencial.

- *COVRATIO* que el efecto en la variabilidad de los coeficientes de regresión al eliminar la $i-$ésima observación. Se define por:

$$\mathcal{CR}_i = \frac{\det\left[s_{(i)}^2(\mathbf{X}_{(i)}^{'}\mathbf{X}_{(i)})^{-1}\right]}{\det\left[s^2(\mathbf{X}^{'}\mathbf{X})^{-1}\right]} = \left(\frac{s_{(i)}^2}{s^2}\right)^{p^{'}}\frac{1}{1-h_{ii}} \tag{7.3.4}$$

  Si este indicador se encuentra fuera del intervalo $(1 - 3\frac{p^{'}}{n}, 1 + 3\frac{p^{'}}{n})$ entonces la $i-$ésima observación podría ser un valor influencial.

**Data Autopartes**

El gerente de ventas de una empresa dedicada a la comercialización de autopartes desea desarrollar un método para pronosticar las ventas anuales de una región. En apariencia varios factores están relacionados con las ventas y de acuerdo con la opinión del gerente las variables que deben intervenir en el análisis son $Y$ = Ventas anuales (en millones de dólares), $X_1$ = Número de tiendas de venta al menudeo, $X_2$ = Número de autos registrados (en millones), $X_3$ = Ingreso personal (en millones de dólares), $X_4$ = Antiguedad promedio de los automóviles (en años) y $X_5$ = Número de supervisores. La información correspondiente a diez regiones elegidas al azar se encuentra en Autopartes. Las medidas de influencia para el modelo de regresión lineal múltiple estimado se muestran a continuación.

```
Influence measures of
         lm(formula = Y ~ X1 + X2 + X3 + X4 + X5) :

    dfb.1_  dfb.X1  dfb.X2   dfb.X3  dfb.X4   dfb.X5   dffit    cov.r   cook.d
1   0.0436 -0.1606  0.0942  0.21700 -0.1199 -0.04326  0.3554  7.24851 0.026617
2   0.0660 -0.1142  0.0981 -0.00242 -0.2455  0.22741 -0.4883  9.70111 0.050149
3  -0.0641  0.3025 -0.1387 -0.16202  0.2344 -0.21768  0.6129  3.35422 0.070324
4   0.9888  0.0293 -0.5189 -0.16190 -0.7459  0.21755  1.2787 12.51972 0.318740
5   0.1072  0.7263 -0.8431 -0.42878  0.1599  0.57422 -1.2034  5.88110 0.265856
6   0.0230  0.0117 -0.0194 -0.02339 -0.0103  0.00346 -0.0535 10.61620 0.000636
7  -2.1657 -1.9576  2.6056  1.53259  2.1447 -3.43876 -3.9681  0.00667 0.950454
8   0.9188  0.6131 -0.6191 -0.37171 -1.1429  0.41471 -1.8331  0.38561 0.413263
9  -0.7828 -0.8090  0.4040  0.52278  0.5607  1.25495  3.0442  0.00593 0.576230
10  0.2218  0.6220 -0.2171 -0.87839 -0.0972  0.14611  1.1524 18.62713 0.269665
     hat inf
1  0.436   *
2  0.584   *
3  0.401
4  0.796   *
5  0.696   *
6  0.474   *
7  0.662   *
8  0.581   *
9  0.545   *
10 0.824   *
```

```
      StudRes       Hat     CookD
7  -2.8362676 0.6618605 0.9749123
10  0.5319235 0.8243556 0.5192925
```

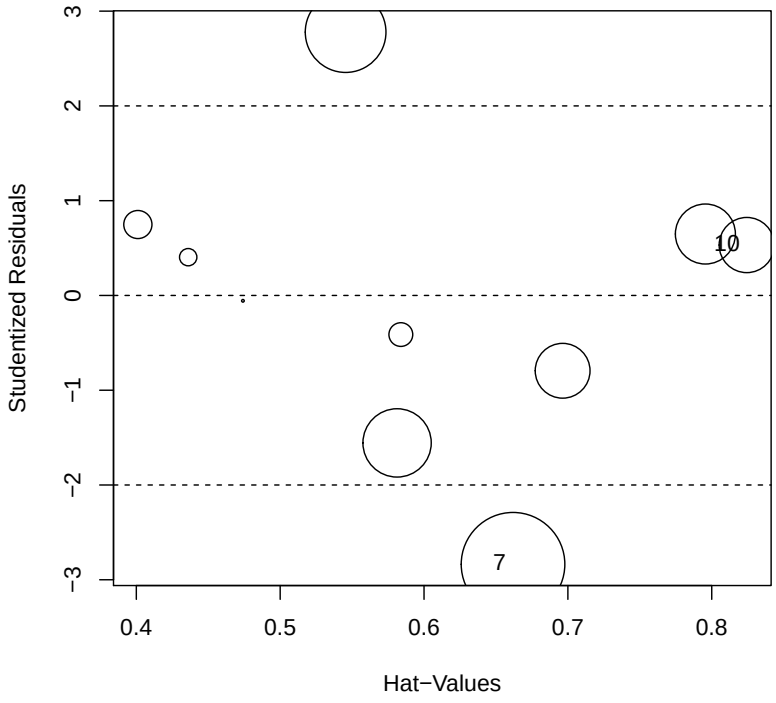


Figura 7.1: Gráfico de influencia para la data `Autopartes`

## 7.4. Verificación de supuestos

### 7.4.1. Supuesto de normalidad de los errores

El supuesto de normalidad de los errores permite realizar el proceso de inferencia en un problema de regresión lineal. Un método gráfico consiste en comparar los residuales con los *scores* o estadísticos de orden normales. El

$i-$ésimo score normal es aproximado por:

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

donde $\Phi$ representa la función de distribución acumulada de la normal estándar y $n > 5$ es el número de observaciones. Si el supuesto de normalidad se cumple los puntos en el gráfico deben estar alineados alrededor de una recta que pasa por el origen.

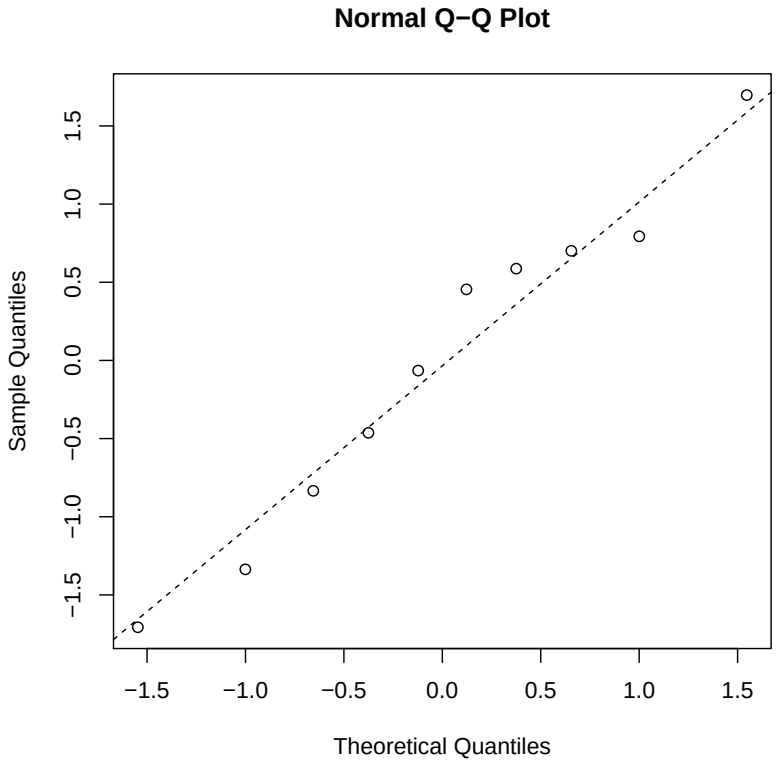


Figura 7.2: Gráfico de normalidad para la data Autopartes

### 7.4.2. Supuesto de homogeneidad de variancias

Un gráfico de residuales puede indicar que el supuesto de homogeneidad de variancias no se cumple. Es posible usar transformaciones para estabilizar

la variancia reemplazando $Y$ por $Y_T$ de tal forma que se logre tener variancia constante en la escala transformada. Una segunda opción es encontrar pesos que puedan ser usados para la estimación por mínimos cuadrados ponderados. Si existen repeticiones entonces las variancias intragrupos pueden ser usadas para aproximar los pesos.

Cook y Weisberg (1983) proporcionan una prueba diagnóstico para detectar variancia no constante. Suponer que $\text{Var}[Y|X]$ depende de un vector de parámetros desconocidos $\boldsymbol{\lambda}$ y de un conjunto de términos conocidos $Z$ . Se asume que:

$$\text{Var}[Y|X, Z = \mathbf{z}] = \sigma^2 \exp\{\boldsymbol{\lambda}'\mathbf{z}\} \tag{7.4.1}$$

Asumiendo que los errores están normalmente distribuidos, es posible establecer una *prueba de score* para $\boldsymbol{\lambda} = \mathbf{0}$

Si se tiene un conjunto inicial de pesos conocidos, entonces la prueba de score considera que:

$$\text{Var}[Y|X, Z = \mathbf{z}] = \frac{\sigma^2}{w} \exp\{\boldsymbol{\lambda}'\mathbf{z}\} \tag{7.4.2}$$

La hipótesis nula para la prueba de score es $\text{Var}[Y|X, Z = \mathbf{z}] = \sigma^2/w$ versus la alternante dada en 7.4.2. La prueba es exactamente la misma desarrollada anteriormente, excepto por el paso uno donde debe usarse la regresión por mínimos cuadrados ponderados con pesos $w$ y en los siguientes pasos deben usarse los residuales de Pearson dados en 7.2.14.

La prueba score para el modelo de regresión en la data `Autopartes` se muestra a continuación. La hipótesis de variancia constante es sostenible con la data.

```
> ncvTest(modelo1)

Non-constant Variance Score Test
Variance formula: ~ fitted.values
Chisquare = 0.1157834    Df = 1     p = 0.7336535
```

### 7.4.3. Supuesto de errores no correlacionados

Uno de los supuestos en el análisis de regresión lineal es que los errores no se encuentran correlacionados, es decir $\text{Cov}(e_i, e_j) = 0$ para $i \neq j$. Si en un gráfico de residuales y los valores estimados se observa una tendencia

cíclica es posible que los errores no cumplan con este supuesto. La prueba de *Durbin-Watson* mide el grado de correlación de un error con el anterior y el posterior a él. El estadístico es:

$$D = \frac{\sum_{i=2}^{n}(e_i - e_{i-1})^2}{\sum_{i=1}^{n} e_i^2} \tag{7.4.3}$$

El valor del estadístico $D$ se encuentra entre 0 (correlación positiva) y 4 (correlación negativa). Si $D$ se encuentra cerca de 2 los errores no se encuentran correlacionados.

```
> durbinWatsonTest(modelo1)

 lag Autocorrelation D-W Statistic p-value
   1    -0.004388905       1.96452   0.564
 Alternative hypothesis: rho != 0
```

## 7.5. Los residuales cuando el modelo no es correcto

Si el modelo estimado se basa en supuestos incorrectos se tendrán gráficos de residuales muy lejos de parecerse a gráficos nulos. La Figura 7.3 muestra algunos gráficos de residuales obtenidos con un modelo de regresión lineal simple.

El primer gráfico es un gráfico nulo que indica que no existen problemas con el modelo estimado. Las Figuras 7.3 (b) y (c) sugieren variancia no constante de la cantidad graficada en el eje horizontal. La curvatura observada en la Figura 7.3 (d) indica que se ha considerado una función media incorrecta.

### Data Gasolina2001

Si la función media y otros supuestos son correctos, entonces todos los posibles gráficos de residuales deben ser gráficos nulos. Lo anterior incluye gráficos de residuales versus cada variable predictora y versus los valores estimados tal como se muestra en la Figura 7.4 para la data `Gasolina2001`.

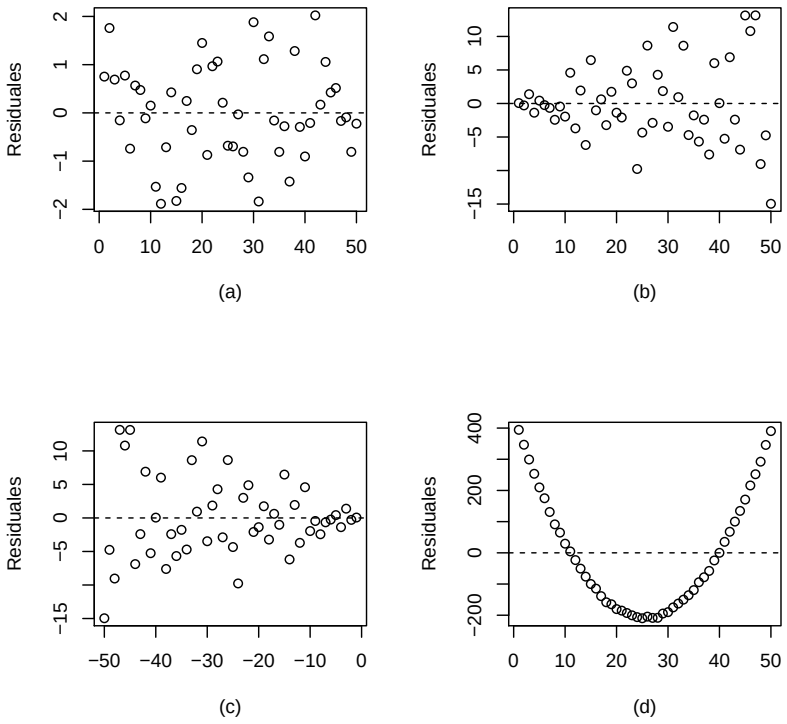


Figura 7.3: Gráficos de residuales

Ninguno de los gráficos en las Figuras 7.4(a)-(d) sugiere algún problema con los supuestos, sin embargo se observa un residual positivo relativamente grande para Wyoming y un residual negativo también grande para Alaska. En algunos de los gráficos, el punto para el distrito de Columbia se encuentra separado de los otros. Wyoming es un estado grande, se encuentra escasamente poblado y tiene un sistema de carreteras bien desarrollado. El conducir grandes distancias para cubrir necesidades vitales, como visitar al doctor, son comunes en este estado. Mientras Alaska también es grande, se encuentra escasamente poblado y la mayoría de las personas viven en áreas relativamente pequeñas cerca de las ciudades. Muchos de los lugares en Alaska no son accesibles por carreteras. Estas condiciones hacen que tenga un bajo consumo de combustible del que pudiera esperarse.

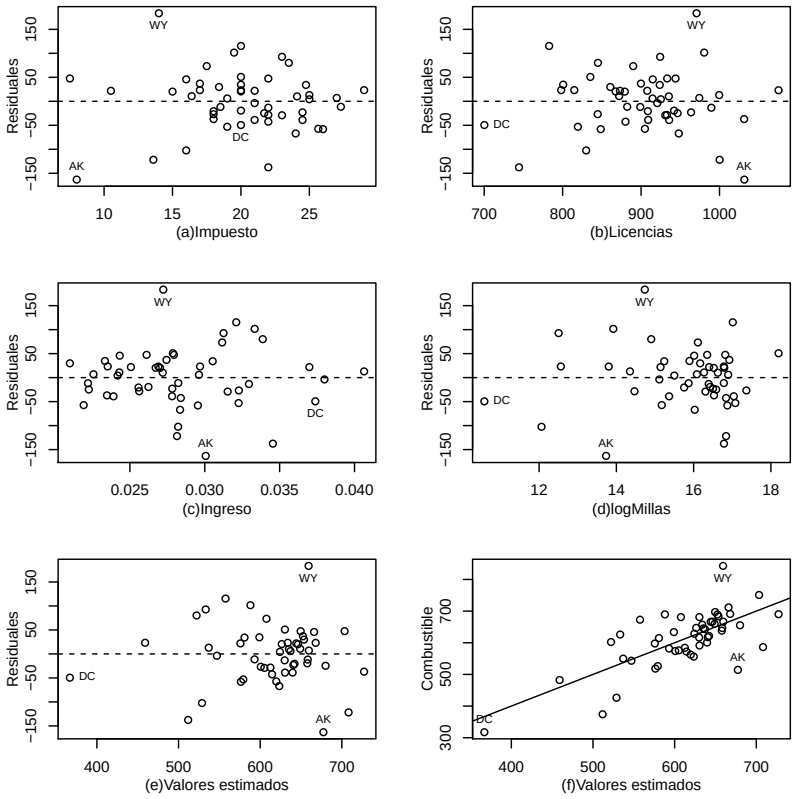


Figura 7.4: Gráfico de residuales para la data Gasolina2001

El distrito de Columbia es un área urbana bastante compacta con un tránsito bastante rápido, por lo que el uso de los autos tiende a ser mucho menor. Tiene un residual pequeño sin embargo constituye un valor inusual ya que se encuentra separado horizontalmente del resto de la data en alguno de los gráficos. El distrito de Columbia tiene un alto leverage, $h_{99} = 0{,}415$, mientras los otros dos estados son candidatos para ser outliers.

La Figura 7.4(e) es un gráfico de residuales versus los valores estimados. La mayoría de los programas de computación producen este gráfico ya que contiene información de todos los términos en la función media. Aparentemente se trataría de un gráfico nulo.

La Figura 7.4(f) es diferente de los anteriores ya que se trata de un gráfico de los valores observados versus los valores estimados. Si la funcion media y

los supuests se cumplen entonces los puntos en el gráfico deberían seguir la tendencia de la recta obtenida por mínimos cuadrados ordinarios.

## 7.6. Probando la curvatura

Se pueden establecer pruebas para ayudar a decidir si los gráficos de residuales como los de la Figura 7.4 corresponden efectivamente a gráficos nulos.

Suponga que se tiene un gráfico de residuales $\hat{e}$ versus una cantidad $U$ sobre el eje horizontal, donde $U$ podría ser un término o una combinación de ellos en la función media (si se trata de un término polinomial, por ejemplo $U = X_1^2$, entonces este procedimiento no es conveniente). Se estima la función media a la que se le agrega el término $U^2$ y se usa una prueba de curvatura para probar la significancia de $U^2$. Si $U$ corresponde a una variable predictora se usa la prueba usual $t$. Si $U$ es igual a los valores estimados entonces se usa la *prueba de no aditividad de Tukey*.

Se muestran las pruebas de curvatura y los gráficos de residuales en la Figura 7.5. En ninguna de estas pruebas se rechaza la hipótesis nula, por lo que no existe suficiente evidencia contra la función media.

Como segundo ejemplo considere nuevamente la data Naciones Unidas, `UN`, con variable respuesta log(`Fertilidad`) y dos predictores log(`PBIpp`) y `Purban`. Se considera que la función media:

$$\mathrm{E}[\log(\mathrm{Fertilidad})|\mathbf{X}] = \beta_0 + \beta_1 \log(\mathrm{PBIpp}) + \beta_2 \mathrm{Purban} \tag{7.6.1}$$

es apropiada para esta data. Los gráficos de residuales versus los dos términos y los valores estimados se muestran en la Figura 7.6. Sin las líneas de referencia la apariencia de los gráficos es satisfactoria. Sin embargo, las pruebas de curvatura dicen algo diferente. Para cada uno de los gráficos las pruebas estadísticas tienen un $p$-valor pequeño sugiriendo que la función media 7.6.1 no es adecuada para esta data.

```
           Test stat Pr(>|t|)
Impuesto      -1.077    0.287
Licencias     -1.922    0.061
Ingreso       -0.084    0.933
logMillas     -1.347    0.185
Tukey test    -1.446    0.148
```

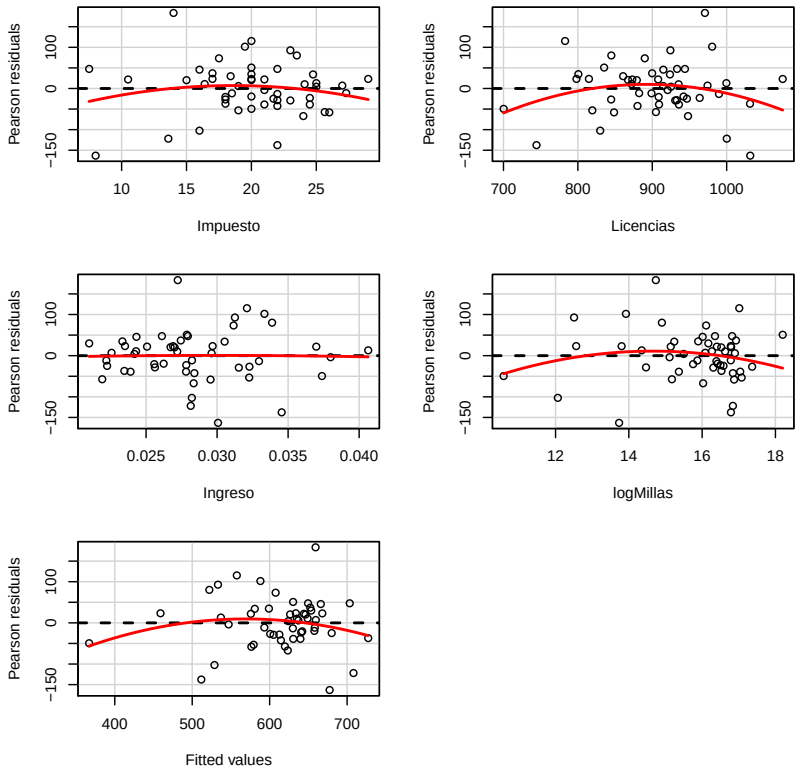


Figura 7.5: Gráfico de residuales para la data `Gasolina2001`

```
             Test stat Pr(>|t|)
log2(PBIpp)      3.219    0.002
Purban           3.368    0.001
Tukey test       3.653    0.000
```

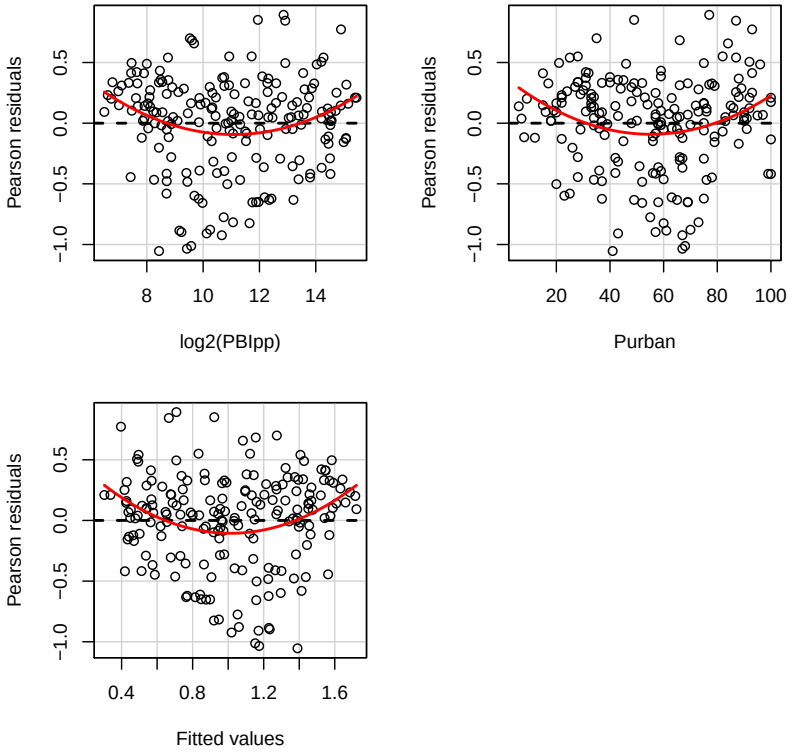


Figura 7.6: Gráfico de residuales para la data UN