# Modelo Lineal: Diagnóstico

**Verificación de Supuestos y Diagnóstico** Supongamos que tenemos una muestra $(y_i, \mathbf{x}_i), \quad i = 1, \ldots, n$ que cumple:

$$y_i = \mathbf{x}_i'\boldsymbol{\beta} + \epsilon_i$$

donde $\epsilon_i = N(0, \sigma^2)$ son independientes y estimamos por el método de mínimos cuadrados a $\boldsymbol{\beta}$ y realizamos todas las inferencias que necesitamos.

**¿Cómo verificamos todos los supuestos que hemos realizado?**

Los 4 supuestos que revisaremos son:

1. Linealidad: $E(Y) = \boldsymbol{X\beta}$
2. Homoscedasticidad: $Var(\epsilon_i) = \sigma^2 = cte.$
3. Normalidad: $\epsilon_i$ tienen distribución Normal
4. Independencia de los errores: $\epsilon_i$ independiente de $\epsilon_j$ si $i \neq j$.

Comencemos por considerar los residuos:

$$e_i = y_i - \hat{y}_i \quad i = 1, \ldots, n$$

Como sabemos

$$\mathbf{e} = (\mathbf{I} - \mathbf{P})\mathbf{Y}$$

por lo tanto

$$E(\mathbf{e}) = 0 \quad \Sigma_{\mathbf{e}} = \sigma^2(\mathbf{I} - \mathbf{P})$$

Por lo tanto, $V(e_i) = \sigma^2(1-p_{ii})$, con lo cual los residuos son heteroscedásticos.

Si además, los errores son normales, como hemos supuesto antes

$$\mathbf{e}_i \sim N(0, \sigma^2(1 - p_{ii}))$$

Observemos además, que los residuos no son independientes, en tanto:

$$Cov(e_i, e_j) = -\sigma^2 p_{ij}$$

Definimos otros residuos relacionados:

$$r_i = \frac{y_i - \hat{y}_i}{s\sqrt{(1 - p_{ii})}} \quad \text{residuo standarizado}$$

$$r_i^* = \frac{y_i - \hat{y}_i}{s_{(i)}\sqrt{(1 - p_{ii})}} \quad \text{residuo studentizado}$$

donde $s_{(i)}$ es el desvío standard muestral computado partir de una regresión ajustada sin la observación $i$.

Sea $\mathbf{X}_{(i)}$ la matriz $\mathbf{X}$ sin la $i-$ésima fila: $\mathbf{x}_i$. Probarán en la práctica que son ciertas las siguientes igualdades:

$$\mathbf{X}'_{(i)}\mathbf{X}_{(i)} = \mathbf{X}'\mathbf{X} - \mathbf{x}_i\mathbf{x}'_i$$

$$\left(\mathbf{X}'_{(i)}\mathbf{X}_{(i)}\right)^{-1} = (\mathbf{X}'\mathbf{X})^{-1} + \frac{(\mathbf{X}'\mathbf{X})^{-1}\,\mathbf{x}_i\mathbf{x}'_i\,(\mathbf{X}'\mathbf{X})^{-1}}{1 - p_{ii}}$$

con lo cual

$$\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}_{(i)} = (\mathbf{X}'\mathbf{X})^{-1}\,\mathbf{x}_i\,\frac{e_i}{1 - p_{ii}}$$

$$s^2_{(i)} = \frac{(n-p)s^2 - e_i^2(1-p_{ii})}{n-p-1}$$

## Distribución de los Residuos

A fin de estudiar la distribución de estos residuos podríamos graficar:

- Esquemas de Tallo y Hoja
- Histogramas

- Boxplots

De esta forma podríamos evaluar:

- simetría
- valores extremos
- valores centrales
- outliers
- posibles agrupamientos
- normalidad

```
summary(salida)

Call:
lm(formula = BIO ~ K + PH)

Residuals:
    Min      1Q  Median      3Q     Max
```

```
-679.25 -253.50  -95.44  259.45 1135.80


Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -506.7131   279.8016  -1.811   0.0773 .
K             -0.4871     0.2031  -2.398   0.0210 *
PH           411.9779    48.4954   8.495 1.15e-10 ***


Residual standard error: 401.1 on 42 degrees of freedom
Multiple R-squared: 0.6476,     Adjusted R-squared: 0.6308
F-statistic: 38.59 on 2 and 42 DF,  p-value: 3.074e-10

names(salida)
 [1] "coefficients"  "residuals"     "effects"       "rank"          "fitted.values" "assign"
 [8] "df.residual"   "xlevels"       "call"          "terms"         "model"


 names(lm.influence(salida))
[1] "hat"          "coefficients" "sigma"        "wt.res"
```

```
stem(salida$res/( 401.1*sqrt(1-lm.influence(salida)$hat)))

The decimal point is at the |

  -1 | 9
  -1 | 332
  -0 | 9888777766655555
  -0 | 332211111
   0 | 022
   0 | 677778
   1 | 04
   1 | 69
   2 | 01
   2 | 9


boxplot(salida$res/( 401.1*sqrt(1-lm.influence(salida)$hat)))

qqnorm(salida$res/( 401.1*sqrt(1-lm.influence(salida)$hat)))
```

## Chequeando la Normalidad

El QQ–plot es un gráfico de percentiles muestrales vs. percentiles teóricos (bajo una cierta distribución asumida $F$).

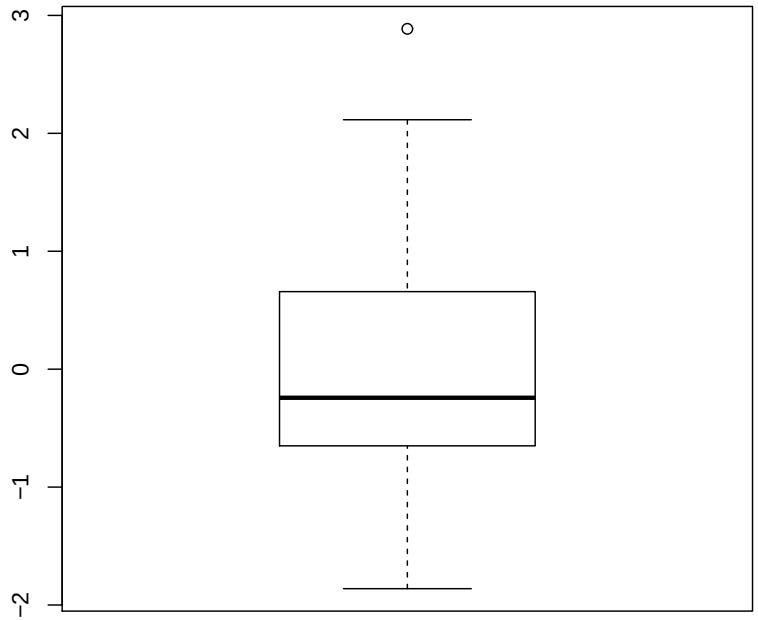



Figura 1: Boxplot de $r_i$: Datos de Biomasa

Si la muestra proviniese de una población con distribución $F$ los percentiles muestrales vs. los teóricos caerían aproximadamente sobre una recta a 45°.

Para esto ordenamos los residuos standarizados

$$r_{(1)} \leq r_{(2)} \leq \ldots \leq r_{(n)}$$

y graficamos los percentiles muestrales $\{1/(n+1), 2/(n+1), \ldots, n/(n+1)\}$ contra los percentiles teóricos de una $N(0,1)$ $\{\phi^{-1}(1/(n+1)), \phi^{-1}(2/(n+1)), \ldots, \phi^{-1}(n/(n+1))\}$ .

Si el gráfico se desviase de la recta, estaríamos encontrando evidencia contra la normalidad.

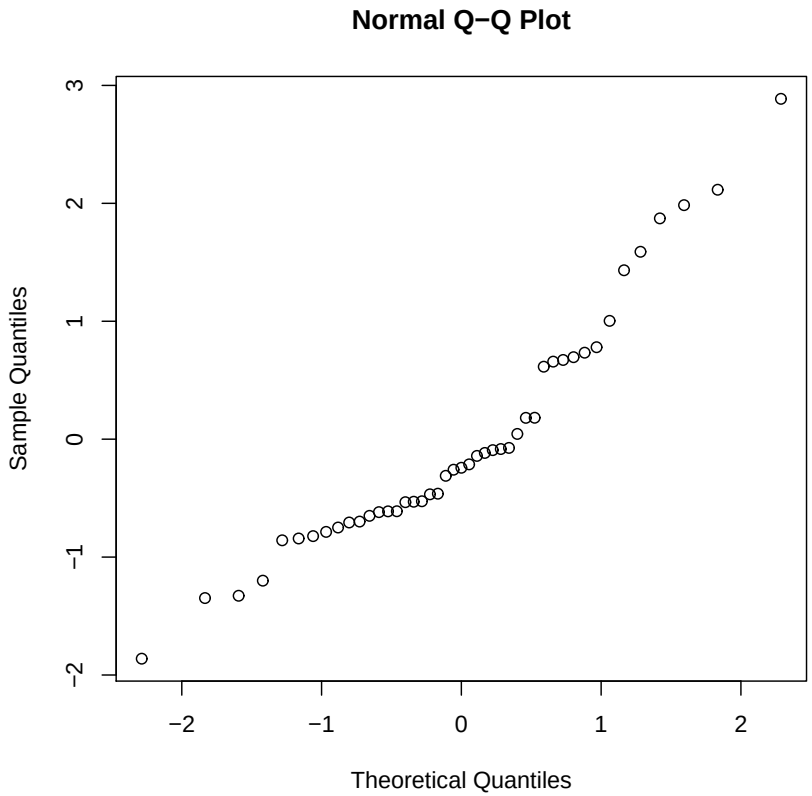



Figura 2: QQ-plot de $r_i$: Datos de Biomasa

## Linealidad

$\hat{y}_i \quad vs. \quad e_i$

Uno de los gráficos que se realiza después de realizar el ajuste es el de $\hat{y}_i \quad vs. \quad e_i$

**¿Qué esperamos observar?** Consideremos el modelo $y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_{p-1} x_{i(p-1)}$

Si quisiéramos hacer una regresión entre $e_i$ vs. $\hat{y}_i$ el estimador de la pendiente tendría como numerador:

$$\sum_{i=1}^{n} (e_i - \overline{e})(\hat{y}_i - \overline{y}) = \sum_{i=1}^{n} e_i \, \hat{y}_i = \mathbf{Y}(\mathbf{I} - \mathbf{P})\mathbf{P}\mathbf{Y} = 0$$

En cambio si la regresión la hiciésemos entre $e_i$ vs. $y_i$ el estimador de la pendiente tendría como numerador:

$$\sum_{i=1}^{n} (e_i - \overline{e})(y_i - \overline{y}) = \sum_{i=1}^{n} e_i \, y_i = \mathbf{Y}'(\mathbf{I} - \mathbf{P})\mathbf{Y} = \mathbf{Y}'(\mathbf{I} - \mathbf{P})(\mathbf{I} - \mathbf{P})\mathbf{Y} = \mathbf{e}'\mathbf{e} = \|\mathbf{e}\|^2$$

es decir, la suma de cuadrados de los residuos.

Más aún, el estimador del coeficiente correspondiente a la pendiente en este caso sería:

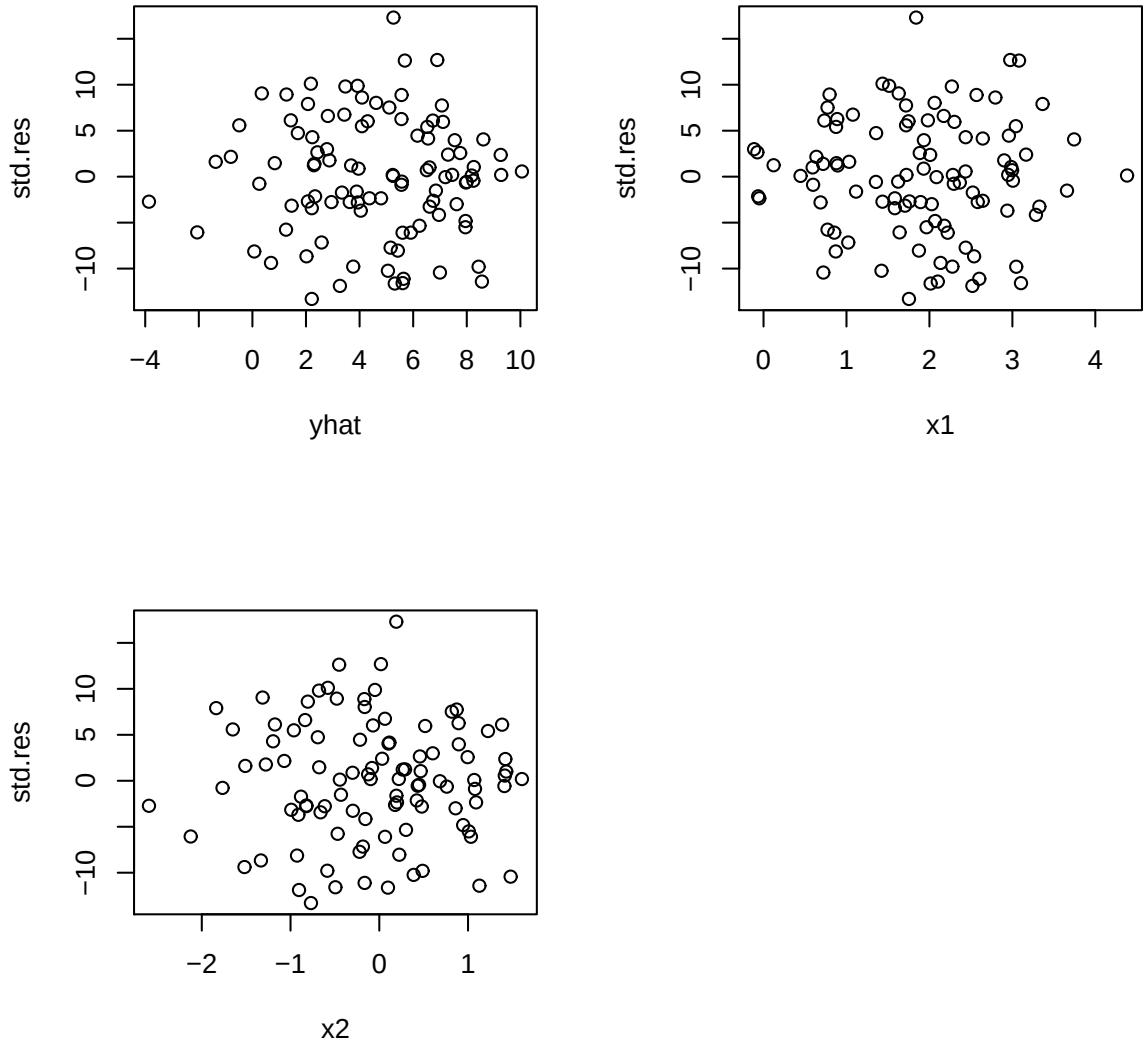



Figura 3: Linealidad: OK!!

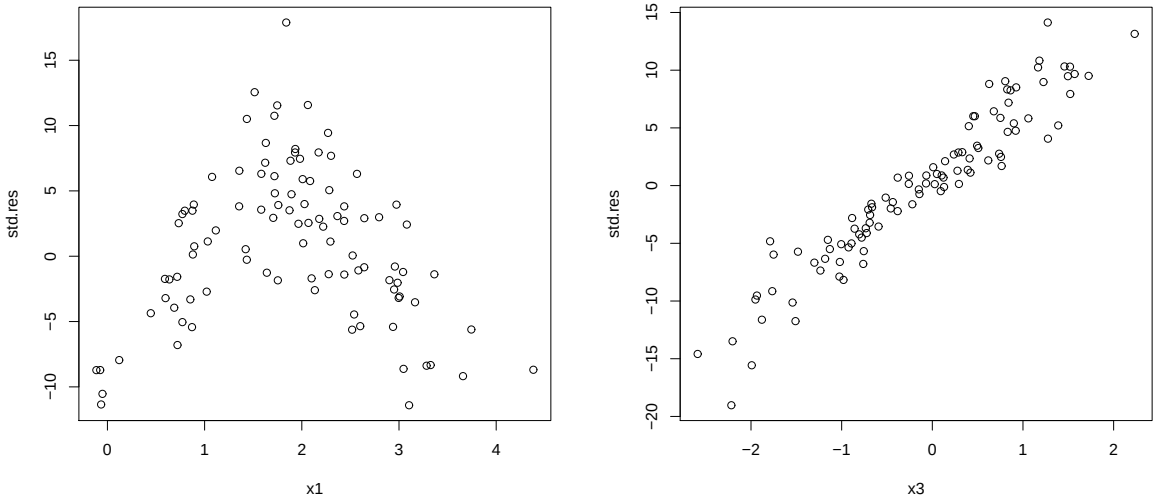
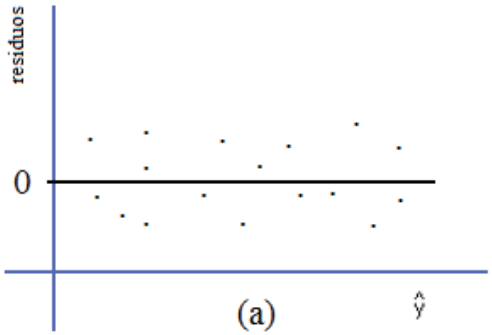
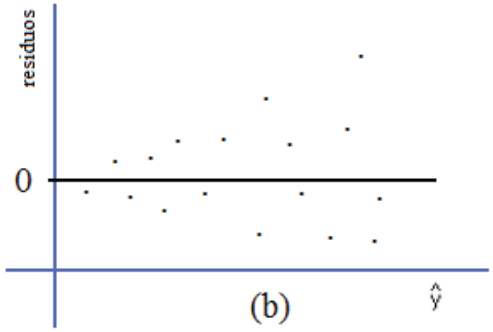
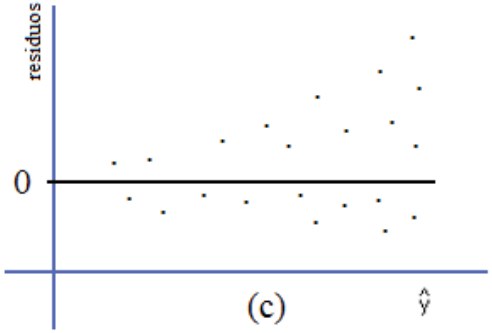
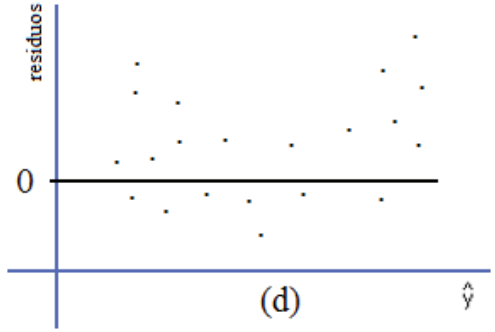



Figura 4: Linealidad: MAL!!




(a)
Representa la situación esperable si el modelo se cumple: una nube de residuos alrededor del 0 sin estructura.

(b) y (c)
Muestran gráficos en los que el supuesto de igualad de varianzas no se cumple.

(d) El supuesto de linealidad no se satisface.

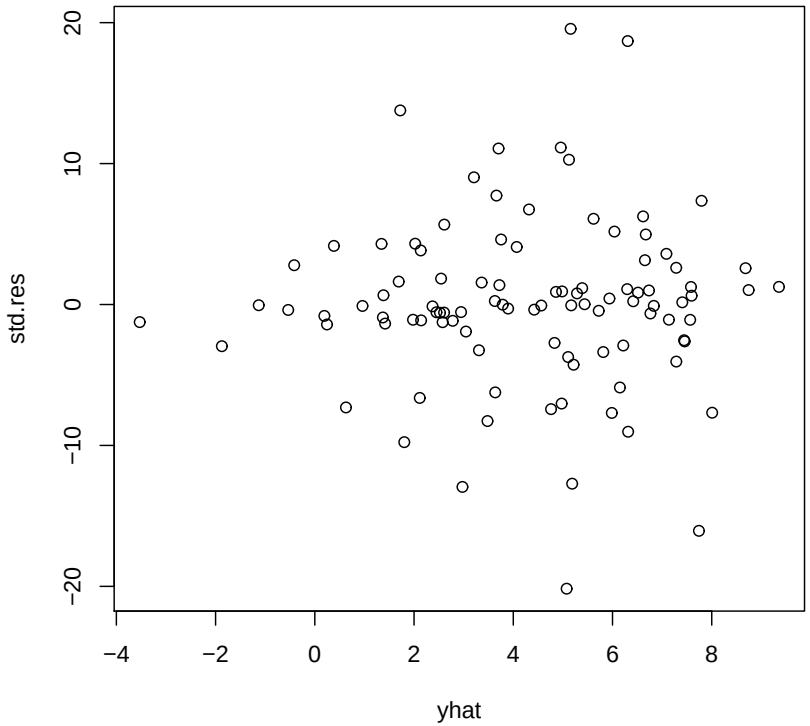



Figura 6: Boxplot de $r_i$: Heteroscedaticidad

$$\frac{\mathbf{e'e}}{\sum_{i=1}^{n}(y_i-\overline{y})^2}=\frac{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}{\sum_{i=1}^{n}(y_i-\overline{y})^2}=\frac{\sum_{i=1}^{n}(y_i-\overline{y})^2-\sum_{i=1}^{n}(\hat{y}_i-\overline{y})^2}{\sum_{i=1}^{n}(y_i-\overline{y})^2}=1-R^2$$

Es decir que esta pendiente sería 0 sólo en el caso de ajuste perfecto.

El caso (d) correspondería a un modelo inadecuado. Por ejemplo, supongamos que ajustamos $y_i=\beta_0+\beta_1 x_{i1}+\epsilon_i$, pero en realidad es:

$$y_i=\beta_0+\beta_1 x_{i1}+\beta_2 x_{i2}+\epsilon_i$$

Luego:

$$\begin{aligned} E(e_i) &= E(y_i-\hat{y}_i) \\ &= E(y_i-\hat{\beta}_0+\hat{\beta}_1 x_{i1}) \\ &= h+g x_{i1}+k x_{i2} \end{aligned}$$

o sea tanto $e_i$ como $y_i$ varían con $x_{i1}$.

### $e_i$ vs. cada variable regresora

Tengamos en cuenta que por las ecuaciones normales:

$$\sum_{i=1}^{n}(e_i-\overline{e})(x_{ij}-\overline{x}_{.j}) = \sum_{i=1}^{n} e_i\ (x_{ij}-\overline{x}_{.j}) = \sum_{i=1}^{n} e_i\ x_{ij} = \sum_{i=1}^{n}(y_i-\hat{\beta}_0-\hat{\beta}_1 x_{i1}+\ldots-\hat{\beta}_{p-1}x_{i(p-1)})x_{ij} = 0$$

De manera que, si el modelo elegido fuera correcto no debería aparecer ninguna estructura en el gráfico de $e_i$ vs. $x_{ij}$. Por lo tanto, los gráficos anteriores también nos sirven de guía en este caso.

Por ejemplo, si en el razonamiento anterior reemplazásemos $x_{i2}$ por $x_{i1}^2$ tendríamos:

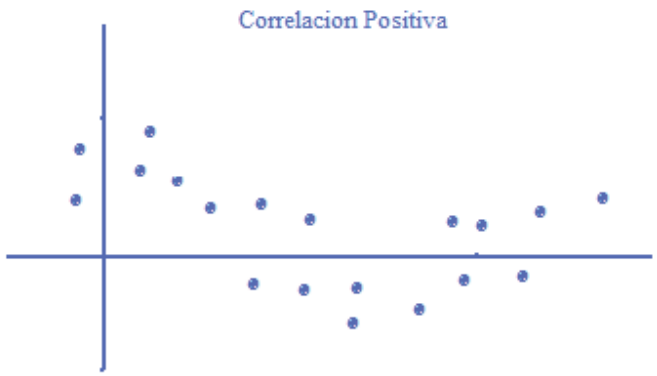
$$E(e_i) = h + g x_{i1} + k x_{i1}^2$$

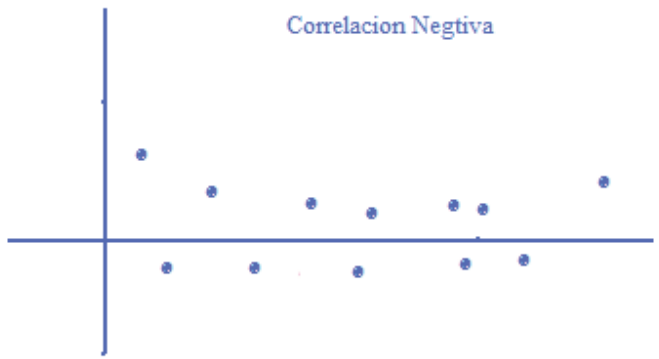
el gráfico quedaría cercano a una parábola.

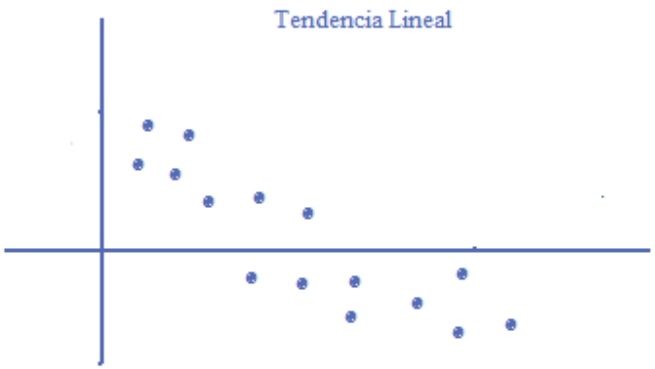
### $e_i$ vs. tiempo

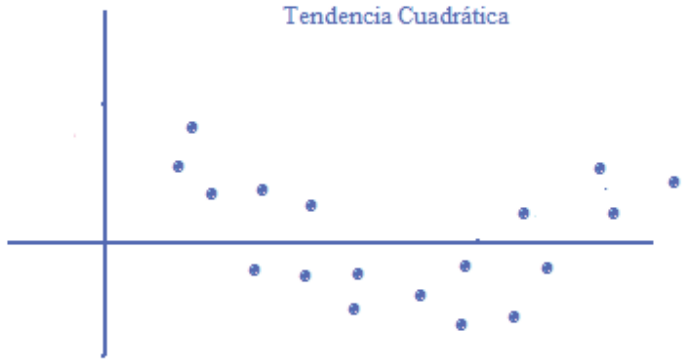
En principio cualquier factor podría influir en $Y$ y debería incluirse en la regresión como variable explicativa. Si un factor ha sido omitido, podría graficarse $e_i$ vs. factor y ver si hay alguna tendencia o patrón particular.

A veces con los datos se registra el tiempo o el orden en que han sido tomadas las mediciones. Puede ser de interés estudiar si los residuos tienen alguna dependencia en el tiempo.


Correlacion Positiva


Correlacion Negtiva


Tendencia Lineal


Tendencia Cuadrática