Universidad de Los Andes
Facultad de Ciencias Económicas y Sociales
Instituto de Estadística

# Métodos Estadísticos I

## Análisis de Residuos

**Prof. Douglas Rivas**

7 de julio de 2010

# Introducción

### Supuestos en el análisis de regresión lineal simple

- La relación entre $x$ e $y$ es lineal.
- $E(\varepsilon_i) = 0$,
- $Var(\varepsilon_i) = \sigma^2$,
- $Cov(\varepsilon_i, \varepsilon_j) = 0 \quad \forall i \neq j$
- $\varepsilon_i$ se distribuyen normal.

### En resumen

$$\varepsilon_i \sim NIID(0, \sigma^2)$$

# Introducción

## Supuestos en el análisis de regresión lineal simple

- La relación entre $x$ e $y$ es lineal.
- $E(\varepsilon_i) = 0$,
- $Var(\varepsilon_i) = \sigma^2$,
- $Cov(\varepsilon_i, \varepsilon_j) = 0 \quad \forall_{i \neq j}$
- $\varepsilon_i$ se distribuyen normal.

## En resumen

$$\varepsilon_i \sim NIID(0, \sigma^2)$$

# Introducción

## Supuestos en el análisis de regresión lineal simple

- La relación entre $x$ e $y$ es lineal.
- $E(\varepsilon_i) = 0$,
- $Var(\varepsilon_i) = \sigma^2$,
- $Cov(\varepsilon_i, \varepsilon_j) = 0 \quad \forall_{i \neq j}$
- $\varepsilon_i$ se distribuyen normal.

## En resumen

$$\varepsilon_i \sim NIID(0, \sigma^2)$$

# Introducción

## Supuestos en el análisis de regresión lineal simple

- La relación entre $x$ e $y$ es lineal.
- $E(\varepsilon_i) = 0$,
- $Var(\varepsilon_i) = \sigma^2$,
- $Cov(\varepsilon_i, \varepsilon_j) = 0 \quad \forall_{i \neq j}$
- $\varepsilon_i$ se distribuyen normal.

## En resumen

$$\varepsilon_i \sim NIID(0, \sigma^2)$$

# Introducción

## Supuestos en el análisis de regresión lineal simple

- La relación entre $x$ e $y$ es lineal.
- $E(\varepsilon_i) = 0$,
- $Var(\varepsilon_i) = \sigma^2$,
- $Cov(\varepsilon_i, \varepsilon_j) = 0 \quad \forall_{i \neq j}$
- $\varepsilon_i$ se distribuyen normal.

## En resumen

$$\varepsilon_i \sim NIID(0, \sigma^2)$$

# Introducción

## Supuestos en el análisis de regresión lineal simple

- La relación entre $x$ e $y$ es lineal.
- $E(\varepsilon_i) = 0$,
- $Var(\varepsilon_i) = \sigma^2$,
- $Cov(\varepsilon_i, \varepsilon_j) = 0 \quad \forall_{i \neq j}$
- $\varepsilon_i$ se distribuyen normal.

## En resumen

$$\varepsilon_i \sim NIID(0, \sigma^2)$$

# Introducción

## Supuestos en el análisis de regresión lineal simple

- La relación entre $x$ e $y$ es lineal.
- $E(\varepsilon_i) = 0$,
- $Var(\varepsilon_i) = \sigma^2$,
- $Cov(\varepsilon_i, \varepsilon_j) = 0 \quad \forall_{i \neq j}$
- $\varepsilon_i$ se distribuyen normal.

## En resumen

$$\varepsilon_i \sim NIID(0, \sigma^2)$$

# Introducción

## Supuestos en el análisis de regresión lineal simple

- La relación entre $x$ e $y$ es lineal.
- $E(\varepsilon_i) = 0$,
- $Var(\varepsilon_i) = \sigma^2$,
- $Cov(\varepsilon_i, \varepsilon_j) = 0 \quad \forall_{i \neq j}$
- $\varepsilon_i$ se distribuyen normal.

## En resumen

$$\varepsilon_i \sim NIID(0, \sigma^2)$$

# Introducción

## ¿Qué ocurre si se violan los supuestos?

- La violación de algunos supuestos es grave.
- El no cumplimiento de la normalidad invalida todos los procedimientos de inferencias.
- Si los errores no están descorrelacionados el cálculo de $y$ depende de los errores anteriores.
- Si la varianza no es constante, entonces la variabilidad de $y$ depende de la variabilidad de $\varepsilon$.

# Introducción

## ¿Qué ocurre si se violan los supuestos?

- La violación de algunos supuestos es grave.
- El no cumplimiento de la normalidad invalida todos los procedimientos de inferencias.
- Si los errores no estan descorrelacionados el cálculo de $y$ depende de los errores anteriores.
- Si la varianza no es constante, entonces la variabilidad de $y$ depende de la variabilidad de $\varepsilon$.

# Introducción

## ¿Qué ocurre si se violan los supuestos?

- La violación de algunos supuestos es grave.
- El no cumplimiento de la normalidad invalida todos los procedimientos de inferencias.
- Si los errores no estan descorrelacionados el cálculo de $y$ depende de los errores anteriores.
- Si la varianza no es constante, entonces la variabilidad de $y$ depende de la variabilidad de $\varepsilon$.

# Introducción

## ¿Qué ocurre si se violan los supuestos?

- La violación de algunos supuestos es grave.
- El no cumplimiento de la normalidad invalida todos los procedimientos de inferencias.
- Si los errores no estan descorrelacionados el cálculo de $y$ depende de los errores anteriores.
- Si la varianza no es constante, entonces la variabilidad de $y$ depende de la variabilidad de $\varepsilon$.

# Introducción

## ¿Qué ocurre si se violan los supuestos?

- La violación de algunos supuestos es grave.
- El no cumplimiento de la normalidad invalida todos los procedimientos de inferencias.
- Si los errores no estan descorrelacionados el cálculo de $y$ depende de los errores anteriores.
- Si la varianza no es constante, entonces la variabilidad de $y$ depende de la variabilidad de $\varepsilon$.

# Introducción

**¿Cómo evaluar el cumplimiento de los supuestos?**

Para evaluar los supuestos se realiza un análisis de los residuos, el cuál comprende un conjunto de técnicas tanto gráficas como pruebas estadísticas que permiten evaluar el cumplimiento de los supuestos.

# Introducción

## ¿Cómo evaluar el cumplimiento de los supuestos?

Para evaluar los supuestos se realiza un análisis de los residuos, el cuál comprende un conjunto de técnicas tanto gráficas como pruebas estadísticas que permiten evaluar el cumplimiento de los supuestos.

# Residuos

**Definición (Residuos)**

*Considere el modelo $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$, los residuales se definen como las n diferencias*

$$e_i = Y_i - \hat{Y}_i \qquad i = 1, 2, ..., n \tag{1}$$

*donde*

- *$Y_i$ es una observación*
- *$\hat{Y}_i$ es el correspondiente valor ajustado obtenido al usar la ecuación de regresión ajustada.*

# Residuos

## Definición (Residuos)

*Considere el modelo* $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$*, los residuales se definen como las n diferencias*

$$e_i = Y_i - \hat{Y}_i \qquad i = 1, 2, ..., n \tag{1}$$

*donde*

- *$Y_i$ es una observación*
- *$\hat{Y}_i$ es el correspondiente valor ajustado obtenido al usar la ecuación de regresión ajustada.*

# Residuos

## Definición (Residuos)

*Considere el modelo* $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$*, los residuales se definen como las* $n$ *diferencias*

$$e_i = Y_i - \hat{Y}_i \qquad i = 1, 2, ..., n \tag{1}$$

*donde*

- $Y_i$ *es una observación*
- $\hat{Y}_i$ *es el correspondiente valor ajustado obtenido al usar la ecuación de regresión ajustada.*

# Residuos

## ¿Por qué los residuos?

- Un residuo es la desviación entre los datos y el ajuste.
- Es una medida de la variabilidad de la variable respuesta que no explica el modelo.
- Los residuos se pueden ver como los valores observados o realizados de los errores si el modelo es correcto.

# Residuos

## ¿Por qué los residuos?

- Un residuo es la desviación entre los datos y el ajuste.
- Es una medida de la variabilidad de la variable respuesta que no explica el modelo.
- Los residuos se pueden ver como los valores observados o realizados de los errores si el modelo es correcto.

# Residuos

## ¿Por qué los residuos?

- Un residuo es la desviación entre los datos y el ajuste.
- Es una medida de la variabilidad de la variable respuesta que no explica el modelo.
- Los residuos se pueden ver como los valores observados o realizados de los errores si el modelo es correcto.

# Residuos

## ¿Por qué los residuos?

- Un residuo es la desviación entre los datos y el ajuste.
- Es una medida de la variabilidad de la variable respuesta que no explica el modelo.
- Los residuos se pueden ver como los valores observados o realizados de los errores si el modelo es correcto.

# Residuos

## Propiedades de los Residuos

- Tienen media igual a cero.
- La varianza está dada por

$$V(e_i) = \sigma^2 \left[ 1 - \left( \frac{1}{n} + \frac{(x_i - \bar{x})^2}{S_{xx}} \right) \right]$$

Lo cual implica que la varianza de los residuales no es constante.

# Residuos

## Propiedades de los Residuos

- Tienen media igual a cero.
- La varianza está dada por

$$V(e_i) = \sigma^2 \left[1 - \left(\frac{1}{n} + \frac{(x_i - \bar{x})^2}{S_{xx}}\right)\right]$$

Lo cual implica que la varianza de los residuales no es constante.

# Residuos

## Propiedades de los Residuos

- Tienen media igual a cero.
- La varianza está dada por

$$V(e_i) = \sigma^2 \left[ 1 - \left( \frac{1}{n} + \frac{(x_i - \bar{x})^2}{S_{xx}} \right) \right]$$

Lo cual implica que la varianza de los residuales no es constante.

# Tipos de Residuos

## Residuos Estandarizados

Se obtienen al dividir los residuos entre su respectiva desviación estándar,

$$d_i^* = \frac{e_i}{\sqrt{V(e_i)}} = \frac{e_i}{\sqrt{\sigma^2\left[1 - \left(\frac{1}{n} + \frac{(x_i - \bar{x})^2}{S_{xx}}\right)\right]}} \qquad (2)$$

Como $\sigma^2$ es deconocido se usa $CM_E$

$$d_i = \frac{e_i}{\sqrt{CM_E\left[1 - \left(\frac{1}{n} + \frac{(x_i - \bar{x})^2}{S_{xx}}\right)\right]}} \qquad (3)$$

# Tipos de Residuos

## Residuos Estandarizados

Se obtienen al dividir los residuos entre su respectiva desviación estándar,

$$d_i^* = \frac{e_i}{\sqrt{V(e_i)}} = \frac{e_i}{\sqrt{\sigma^2\left[1-\left(\frac{1}{n}+\frac{(x_i-\bar{x})^2}{S_{xx}}\right)\right]}} \tag{2}$$

Como $\sigma^2$ es deconocido se usa $CM_E$

$$d_i = \frac{e_i}{\sqrt{CM_E\left[1-\left(\frac{1}{n}+\frac{(x_i-\bar{x})^2}{S_{xx}}\right)\right]}} \tag{3}$$

# Tipos de Residuos

## Problema con los Residuos Estandarizados

En el cálculo de $d_i$ hay una relación de dependencia entre el numerador y el denominador.

## Residuos Estudentizados

Se obtienen al calcular el $CM_E$ una vez eliminada la observación correspondiente al residuo que se esta calculando

$$r_i = \frac{e_i}{\sqrt{CM_{E_i}\left[1 - \left(\frac{1}{n} + \frac{(x_i - \bar{x})^2}{S_{xx}}\right)\right]}} \qquad (4)$$

# Tipos de Residuos

## Problema con los Residuos Estandarizados

En el cálculo de $d_i$ hay una relación de dependencia entre el numerador y el denominador.

## Residuos Estudentizados

Se obtienen al calcular el $CM_E$ una vez eliminada la observación correspondiente al residuo que se esta calculando

$$r_i = \frac{e_i}{\sqrt{CM_{E_i}\left[1 - \left(\frac{1}{n} + \frac{(x_i - \bar{x})^2}{S_{xx}}\right)\right]}} \quad (4)$$

# Tipos de Residuos

## Problema con los Residuos Estandarizados

En el cálculo de $d_i$ hay una relación de dependencia entre el numerador y el denominador.

## Residuos Estudentizados

Se obtienen al calcular el $CM_E$ una vez eliminada la observación correspondiente al residuo que se esta calculando

$$r_i = \frac{e_i}{\sqrt{CM_{E_i}\left[1 - \left(\frac{1}{n} + \frac{(x_i - \bar{x})^2}{S_{xx}}\right)\right]}} \qquad (4)$$

# Gráficos de los Residuos

## Gráficos de los Residuos

- Las técnicas gráficas son muy efectivas para detectar un comportamiento anormal de los residuos
- Si el modelo es correcto y los supuestos se satisfacen, los residuales deberían aparecer en cualquier gráfico como una variación aleatoria alrededor del cero

# Gráficos de los Residuos

## Gráficos de los Residuos

- Las técnicas gráficas son muy efectivas para detectar un comportamiento anormal de los residuos
- Si el modelo es correcto y los supuestos se satisfacen, los residuales deberían aparecer en cualquier gráfico como una variación aleatoria alrededor del cero

# Gráficos de los Residuos

## Gráficos de los Residuos

- Las técnicas gráficas son muy efectivas para detectar un comportamiento anormal de los residuos
- Si el modelo es correcto y los supuestos se satisfacen, los residuales deberían aparecer en cualquier gráfico como una variación aleatoria alrededor del cero

# Gráficos de los Residuos

## Gráficos de los Residuos

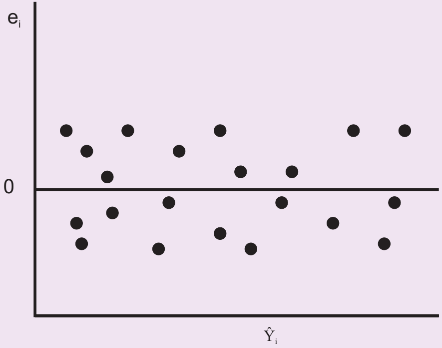



Figura: Comportamiento correcto

# Gráficos de los Residuos

## Gráficos de los Residuos

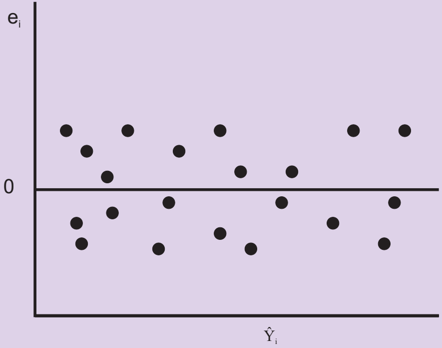
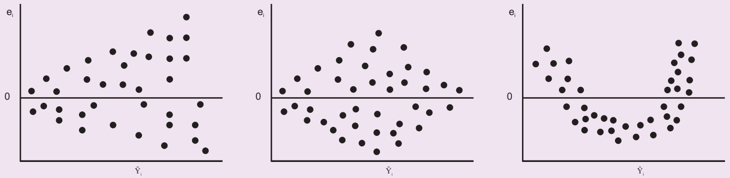



Figura: Comportamiento correcto

# Gráficos de los Residuos

## Gráficos de los Residuos

Cualquier patrón convincente de los residuales sugiere alguna inadecuación en el modelo o en los supuestos




(a) (b) (c)

Figura: Patrones Extaños

# Gráficos de los Residuos

## Gráficos de los Residuos

Cualquier patrón convincente de los residuales sugiere alguna inadecuación en el modelo o en los supuestos

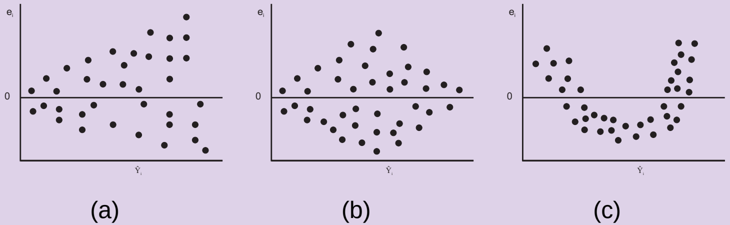



Figura: Patrones Extaños

# Gráficos de los Residuos

## Importancia de los Gráficos

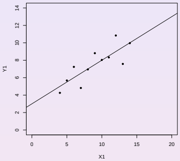
Anscombe (1973) presentó cuatro conjuntos de datos que dan los siguientes resultados

# Gráficos de los Residuos

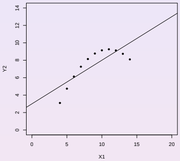
## Importancia de los Gráficos

Anscombe (1973) presentó cuatro conjuntos de datos que dan los siguientes resultados

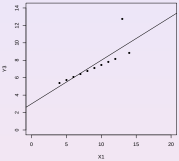
# Gráficos de los Residuos



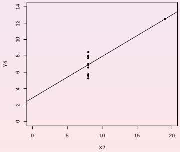

(a)




(b)

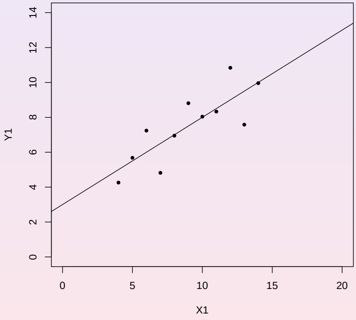


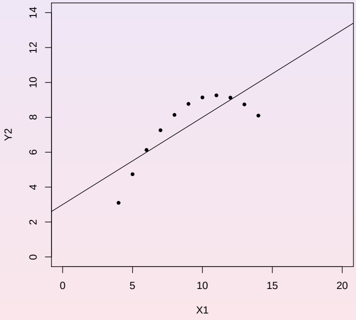

(c)



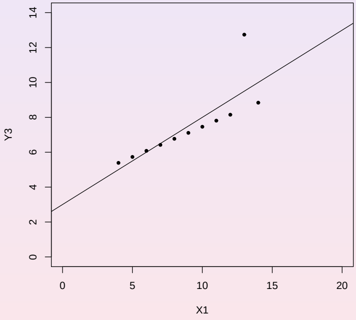

(d)

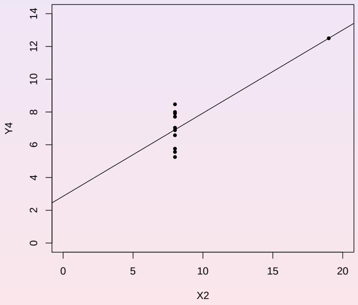
# Gráficos de los Residuos



# Gráficos de los Residuos



# Gráficos de los Residuos

## Gráficos más usados son...

- Histograma,
- gráficos de probabilidad normal,
- gráficos de los residuos versus los valores ajustados,
- gráficos de los residuos versus la variable independiente,
- gráfico de $(e_i)$ versus $(e_{i-1})$.

# Gráficos de los Residuos

## Gráficos más usados son...

- Histograma,
- gráficos de probabilidad normal,
- gráficos de los residuos versus los valores ajustados,
- gráficos de los residuos versus la variable independiente,
- gráfico de $(e_i)$ versus $(e_{i-1})$.

# Gráficos de los Residuos

## Gráficos más usados son...

- Histograma,
- gráficos de probabilidad normal,
- gráficos de los residuos versus los valores ajustados,
- gráficos de los residuos versus la variable independiente,
- gráfico de $(e_i)$ versus $(e_{i-1})$.

# Gráficos de los Residuos

## Gráficos más usados son...

- Histograma,
- gráficos de probabilidad normal,
- gráficos de los residuos versus los valores ajustados,
- gráficos de los residuos versus la variable independiente,
- gráfico de $(e_i)$ versus $(e_{i-1})$.

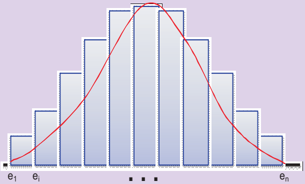
# Gráficos de los Residuos

## Gráficos más usados son...

- Histograma,
- gráficos de probabilidad normal,
- gráficos de los residuos versus los valores ajustados,
- gráficos de los residuos versus la variable independiente,
- gráfico de $(e_i)$ versus $(e_{i-1})$.

# Histograma

## Histograma

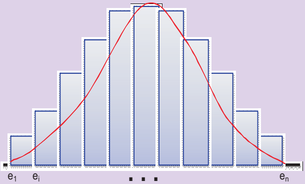



Se usan para:

- Observar el cumplimiento de normalidad y
- Observar la simetría de los datos

# Histograma

## Histograma

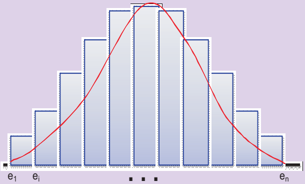



Se usan para:

- Observar el cumplimiento de normalidad y
- Observar la simetría de los datos

# Histograma

## Histograma




Se usan para:

- Observar el cumplimiento de normalidad y
- Observar la simetría de los datos

# Gráficos de Probabilidad Normal

## Gráficos de Probabilidad Normal

- Permiten visualizar el supuesto de normalidad y
- determinar la simetría de los datos
- Si la distribución de los residuales coincide con la normal, los puntos se concentrarán en torno a una línea recta.

# Gráficos de Probabilidad Normal

## Gráficos de Probabilidad Normal

- Permiten visualizar el supuesto de normalidad y
- determinar la simetría de los datos
- Si la distribución de los residuales coincide con la normal, los puntos se concentrarán en torno a una línea recta.

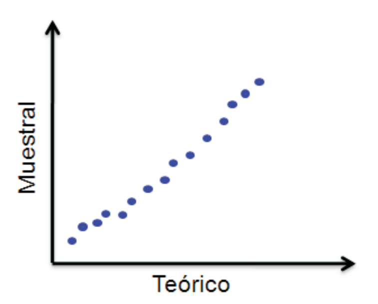
# Gráficos de Probabilidad Normal

## Gráficos de Probabilidad Normal

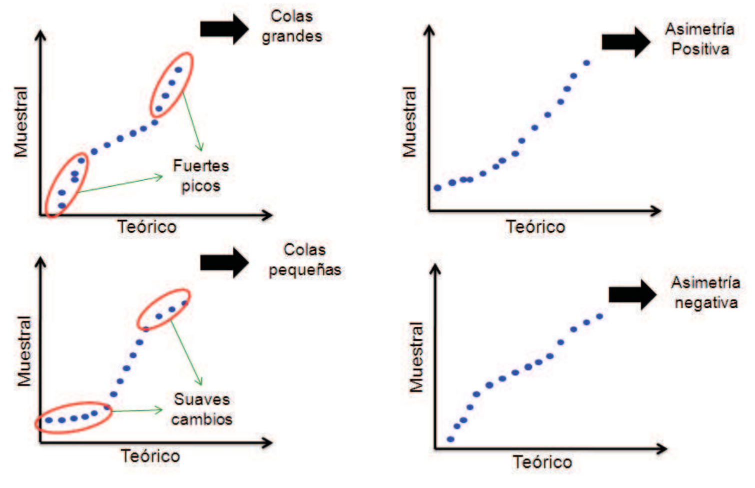
- Permiten visualizar el supuesto de normalidad y
- determinar la simetría de los datos
- Si la distribución de los residuales coincide con la normal, los puntos se concentrarán en torno a una línea recta.

# Gráficos de Probabilidad Normal

## Gráfico Q-Q



# Gráficos de Probabilidad Normal



# Gráficos de $(e_i)$ versus $(\hat{Y}_i)$

## Gráfico de $(e_i)$ versus $(\hat{Y}_i)$

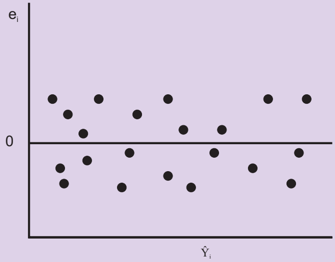



Figura: Correcto

Se usan para:

- Varianzas desiguales
- Datos atípicos
- Modelo inadecuado

# Gráficos de $(e_i)$ versus $(\hat{Y}_i)$

## Gráfico de $(e_i)$ versus $(\hat{Y}_i)$

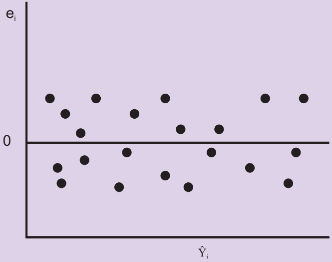



Figura: Correcto

Se usan para:

- Varianzas desiguales
- Datos atípicos
- Modelo inadecuado

# Gráficos de $(e_i)$ versus $(\hat{Y}_i)$

## Gráfico de $(e_i)$ versus $(\hat{Y}_i)$

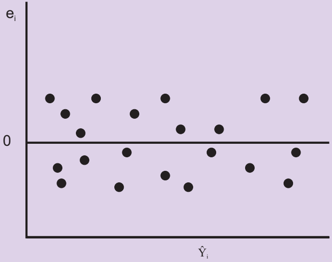



Figura: Correcto

Se usan para:

- Varianzas desiguales
- Datos atípicos
- Modelo inadecuado

# Gráficos de $(e_i)$ versus $(\hat{Y}_i)$

## Gráfico de $(e_i)$ versus $(\hat{Y}_i)$

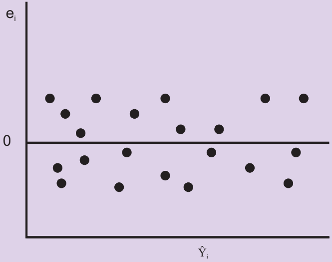


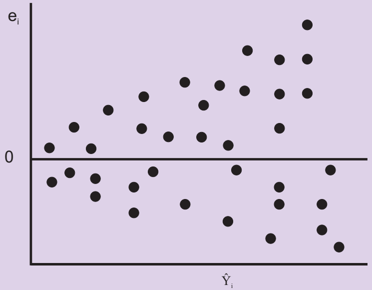

Figura: Correcto

Se usan para:

- Varianzas desiguales
- Datos atípicos
- Modelo inadecuado

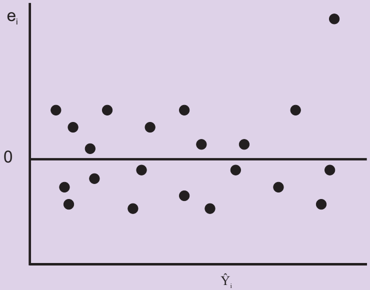
# Gráficos de los Residuos versus los valores ajustados

## Detectar Varianzas desiguales



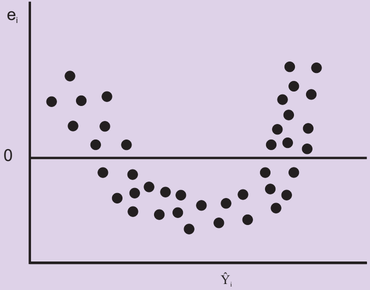
# Gráficos de los Residuos versus los valores ajustados

## Detectar datos atípicos



# Gráficos de los Residuos versus los valores ajustados

## Detectar forma funcional inadecuada



# Gráfico de $e_i$ versus $X_i$.

## Gráfico de $e_i$ versus $X_i$.

Tiene similar interpretación al gráfico de los residuales versus el valor ajustado, con la diferencia de que este permite deducir si la existencia de heterocedasticidad o la falta de linealidad en el modelo son debidas a la variable explicativa representada.

# Gráfico de $e_{i+1}$ versus $e_i$.

## Gráfico de $e_{i+1}$ versus $e_i$.

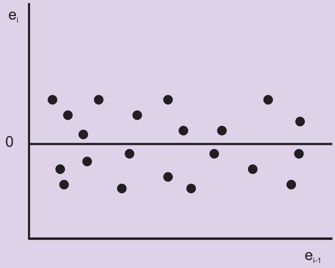



Figura: Correcto

Permite visualizar la correlación entre los errores.

# Gráfico de $e_{i+1}$ versus $e_i$.

## Gráfico de $e_{i+1}$ versus $e_i$.

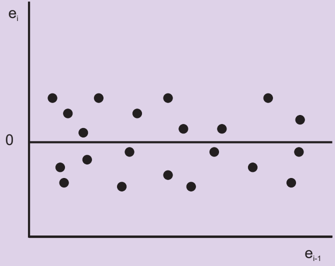



Figura: Correcto

Permite visualizar la correlación entre los errores.