

Regresión Lineal Simple con R Comander

Rosa María Fernández Alcalá

Curso de Especialización:
"HERRAMIENTAS DE SOFTWARE LIBRE EN EL ÁMBITO
MATEMÁTICO-ESTADÍSTICO: GEOGEBRA Y R"
Abril 2012

Objetivos

- Cuantificar el grado de relación lineal entre dos variables estadísticas.
- Utilizar el modelo de regresión lineal simple para interpretar la relación de una variable independiente con una variable respuesta.
- Desarrollar inferencias respecto de los parámetros del modelo.
- Realizar predicciones en base al modelo de regresión ajustado.
- Realizar una diagnosis básica del modelo de regresión lineal simple.

Objetivos

- Cuantificar el grado de relación lineal entre dos variables estadísticas.
- Utilizar el modelo de regresión lineal simple para interpretar la relación de una variable independiente con una variable respuesta.
- Desarrollar inferencias respecto de los parámetros del modelo.
- Realizar predicciones en base al modelo de regresión ajustado.
- Realizar una diagnosis básica del modelo de regresión lineal simple.

Objetivos

- Cuantificar el grado de relación lineal entre dos variables estadísticas.
- Utilizar el modelo de regresión lineal simple para interpretar la relación de una variable independiente con una variable respuesta.
- Desarrollar inferencias respecto de los parámetros del modelo.
- Realizar predicciones en base al modelo de regresión ajustado.
- Realizar una diagnosis básica del modelo de regresión lineal simple.

Objetivos

- Cuantificar el grado de relación lineal entre dos variables estadísticas.
- Utilizar el modelo de regresión lineal simple para interpretar la relación de una variable independiente con una variable respuesta.
- Desarrollar inferencias respecto de los parámetros del modelo.
- Realizar predicciones en base al modelo de regresión ajustado.
- Realizar una diagnosis básica del modelo de regresión lineal simple.

Objetivos

- Cuantificar el grado de relación lineal entre dos variables estadísticas.
- Utilizar el modelo de regresión lineal simple para interpretar la relación de una variable independiente con una variable respuesta.
- Desarrollar inferencias respecto de los parámetros del modelo.
- Realizar predicciones en base al modelo de regresión ajustado.
- Realizar una diagnosis básica del modelo de regresión lineal simple.

Contenidos

- 1 Introducción
- 2 Correlación
 - Coeficiente de correlación lineal
- 3 El modelo RLS
 - Formulación del modelo RLS
 - Estimación de los parámetros del modelo RLS
 - Inferencia estadística sobre los parámetros del modelo RLS
 - Coeficiente de determinación
- 4 Predicción
 - Predicción para un valor individual
 - Predicción para el valor medio
- 5 Diagnóstico
 - Hipótesis de normalidad
 - Hipótesis de Homocedasticidad
 - Hipótesis de Independencia

Bibliografía

Ross, S.M. (2007).

“Introducción a la Estadística”. Reverté.

Trívez Bielsa, F.J. (2004).

“Introducción a la Econometría”. Ediciones Pirámide, Madrid.

Contenidos

1 Introducción

2 Correlación

- Coeficiente de correlación lineal

3 El modelo RLS

- Formulación del modelo RLS
- Estimación de los parámetros del modelo RLS
- Inferencia estadística sobre los parámetros del modelo RLS
- Coeficiente de determinación

4 Predicción

- Predicción para un valor individual
- Predicción para el valor medio

5 Diagnos

- Hipótesis de normalidad
- Hipótesis de Homocedasticidad
- Hipótesis de Independencia

1. Introducción

- El análisis de la regresión es una de las herramientas más utilizada para la descripción y evaluación de la relación entre:
 - Variable Y : **endógena**, **explicada**, **dependiente** o **respuesta**.
 - Variables X_1, \dots, X_k : **exógenas**, **explicativas** o **independientes**.
- Dependiendo del número de variables exógenas, distinguimos:
 - **Regresión simple**, si $k = 1$.

Ejemplo (Regresión simple)

Relación entre el consumo de agua en un municipio (Y) y el número de habitantes (X).

- **Regresión múltiple**, si $k > 1$.

Ejemplo (Regresión múltiple)

Relación existente entre los gastos de consumo (Y) por un lado, y el ingreso familiar (X_1), los activos financieros de la familia (X_2) y el nº de miembros que la forman (X_3) por otro.

1. Introducción

- El análisis de la regresión es una de las herramientas más utilizada para la descripción y evaluación de la relación entre:
 - Variable Y : **endógena**, **explicada**, **dependiente** o **respuesta**.
 - Variables X_1, \dots, X_k : **exógenas**, **explicativas** o **independientes**.
- Dependiendo del número de variables exógenas, distinguimos:
 - **Regresión simple**, si $k = 1$.

Ejemplo (Regresión simple)

Relación entre el consumo de agua en un municipio (Y) y el número de habitantes (X).

- **Regresión múltiple**, si $k > 1$.

Ejemplo (Regresión múltiple)

Relación existente entre los gastos de consumo (Y) por un lado, y el ingreso familiar (X_1), los activos financieros de la familia (X_2) y el nº de miembros que la forman (X_3) por otro.

1. Introducción

- El análisis de la regresión es una de las herramientas más utilizada para la descripción y evaluación de la relación entre:
 - Variable Y : **endógena**, **explicada**, **dependiente** o **respuesta**.
 - Variables X_1, \dots, X_k : **exógenas**, **explicativas** o **independientes**.
- Dependiendo del número de variables exógenas, distinguimos:
 - **Regresión simple**, si $k = 1$.

Ejemplo (Regresión simple)

Relación entre el consumo de agua en un municipio (Y) y el número de habitantes (X).

- **Regresión múltiple**, si $k > 1$.

Ejemplo (Regresión múltiple)

Relación existente entre los gastos de consumo (Y) por un lado, y el ingreso familiar (X_1), los activos financieros de la familia (X_2) y el nº de miembros que la forman (X_3) por otro.

1. Introducción

- El análisis de la regresión es una de las herramientas más utilizada para la descripción y evaluación de la relación entre:
 - Variable Y : **endógena**, **explicada**, **dependiente** o **respuesta**.
 - Variables X_1, \dots, X_k : **exógenas**, **explicativas** o **independientes**.
- Dependiendo del número de variables exógenas, distinguimos:
 - **Regresión simple**, si $k = 1$.

Ejemplo (Regresión simple)

Relación entre el consumo de agua en un municipio (Y) y el número de habitantes (X).

- **Regresión múltiple**, si $k > 1$.

Ejemplo (Regresión múltiple)

Relación existente entre los gastos de consumo (Y) por un lado, y el ingreso familiar (X_1), los activos financieros de la familia (X_2) y el nº de miembros que la forman (X_3) por otro.

1. Introducción

- El estudio de estas relaciones supone un doble objetivo:
 - Analizar los efectos que tienen cada una de las variables independientes sobre la variable dependiente.
 - Pronosticar el valor de Y para un conjunto determinados de variables explicativas X_1, X_2, \dots, X_k .
- Los métodos de Regresión estudian la **construcción de un modelo matemático** (una función) que explique la posible relación de **dependencia estadística o estocástica** existente entre una variable Y respecto a las variables X_1, \dots, X_k :

$$Y = g(X_1, X_2, \dots, X_k; U) \text{ con } U \text{ perturbación aleatoria}$$

Estos modelos se denominan **modelos de regresión**.

1. Introducción

- El estudio de estas relaciones supone un doble objetivo:
 - Analizar los efectos que tienen cada una de las variables independientes sobre la variable dependiente.
 - Pronosticar el valor de Y para un conjunto determinados de variables explicativas X_1, X_2, \dots, X_k .
- Los métodos de Regresión estudian la **construcción de un modelo matemático** (una función) que explique la posible relación de **dependencia estadística o estocástica** existente entre una variable Y respecto a las variables X_1, \dots, X_k :

$$Y = g(X_1, X_2, \dots, X_k; U) \text{ con } U \text{ perturbación aleatoria}$$

Estos modelos se denominan **modelos de regresión**.

1. Introducción

- El estudio de estas relaciones supone un doble objetivo:
 - Analizar los efectos que tienen cada una de las variables independientes sobre la variable dependiente.
 - Pronosticar el valor de Y para un conjunto determinados de variables explicativas X_1, X_2, \dots, X_k .
- Los métodos de Regresión estudian la **construcción de un modelo matemático** (una función) que explique la posible relación de **dependencia estadística o estocástica** existente entre una variable Y respecto a las variables X_1, \dots, X_k :

$$Y = g(X_1, X_2, \dots, X_k; U) \text{ con } U \text{ perturbación aleatoria}$$

Estos modelos se denominan **modelos de regresión**.

1. Introducción

- El estudio de estas relaciones supone un doble objetivo:
 - Analizar los efectos que tienen cada una de las variables independientes sobre la variable dependiente.
 - Pronosticar el valor de Y para un conjunto determinados de variables explicativas X_1, X_2, \dots, X_k .
- Los métodos de Regresión estudian la **construcción de un modelo matemático** (una función) que explique la posible relación de **dependencia estadística o estocástica** existente entre una variable Y respecto a las variables X_1, \dots, X_k :

$$Y = g(X_1, X_2, \dots, X_k; U) \text{ con } U \text{ perturbación aleatoria}$$

Estos modelos se denominan **modelos de regresión**.

1. Introducción

- Si escogemos $g(X_1, X_2, \dots, X_k; U)$ una función de tipo lineal
 \Rightarrow **modelo de Regresión lineal:**

$$Y = \beta_0 + \beta_1 X_1 + \dots + \beta_k X_k + U$$

- En el caso de considerar una única variable exógena $X \Rightarrow$
modelo de Regresión lineal simple:

$$Y = \beta_0 + \beta_1 X + U$$

¿Es adecuada la función lineal para explicar la relación existente entre las variables?

1. Introducción

- Si escogemos $g(X_1, X_2, \dots, X_k; U)$ una función de tipo lineal
 \Rightarrow **modelo de Regresión lineal**:

$$Y = \beta_0 + \beta_1 X_1 + \dots + \beta_k X_k + U$$

- En el caso de considerar una única variable exógena $X \Rightarrow$
modelo de Regresión lineal simple:

$$Y = \beta_0 + \beta_1 X + U$$

¿Es adecuada la función lineal para explicar la
relación existente entre las variables?

1. Introducción

- Si escogemos $g(X_1, X_2, \dots, X_k; U)$ una función de tipo lineal
 \Rightarrow **modelo de Regresión lineal**:

$$Y = \beta_0 + \beta_1 X_1 + \dots + \beta_k X_k + U$$

- En el caso de considerar una única variable exógena $X \Rightarrow$
modelo de Regresión lineal simple:

$$Y = \beta_0 + \beta_1 X + U$$

¿Es adecuada la función lineal para explicar la
relación existente entre las variables?

Contenidos

- 1 Introducción
- 2 Correlación
 - Coeficiente de correlación lineal
- 3 El modelo RLS
 - Formulación del modelo RLS
 - Estimación de los parámetros del modelo RLS
 - Inferencia estadística sobre los parámetros del modelo RLS
 - Coeficiente de determinación
- 4 Predicción
 - Predicción para un valor individual
 - Predicción para el valor medio
- 5 Diagnos
 - Hipótesis de normalidad
 - Hipótesis de Homocedasticidad
 - Hipótesis de Independencia

2. Correlación

- Para seleccionar el modelo funcional más adecuado, debemos analizar el tipo de asociación existente entre las variables

métodos gráficos. La nube de puntos o diagrama de dispersión representa gráficamente la relación entre las variables.

métodos analíticos. El coeficiente de correlación lineal de Pearson permite evaluar la existencia de asociación lineal entre las variables.

2. Correlación

- Para seleccionar el modelo funcional más adecuado, debemos analizar el tipo de asociación existente entre las variables

métodos gráficos. La nube de puntos o diagrama de dispersión representa gráficamente la relación entre las variables.

métodos analíticos. El coeficiente de correlación lineal de Pearson permite evaluar la existencia de asociación lineal entre las variables.

2. Correlación

- Para seleccionar el modelo funcional más adecuado, debemos analizar el tipo de asociación existente entre las variables

métodos gráficos. La **nube de puntos** o **diagrama de dispersión** representa gráficamente la relación entre las variables.

métodos analíticos. El **coeficiente de correlación lineal de Pearson** permite evaluar la existencia de asociación lineal entre las variables.

métodos gráficos

Diagrama de dispersión

El diagrama de dispersión o nube de puntos es la **representación de los pares de valores observados** para dos variables cuantitativas en unos ejes cartesianos.

- Cada par de valores observados se representa por un punto en el plano.
- La disposición de los puntos en el plano nos da una primera idea del comportamiento conjunto de las variables.

métodos gráficos

Diagrama de dispersión

El diagrama de dispersión o nube de puntos es la **representación de los pares de valores observados** para dos variables cuantitativas en unos ejes cartesianos.

- Cada par de valores observados se representa por un punto en el plano.
- La disposición de los puntos en el plano nos da una primera idea del comportamiento conjunto de las variables.

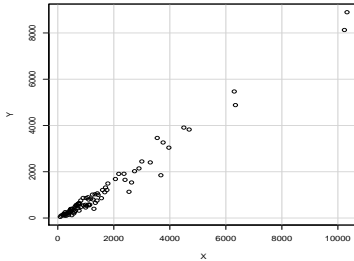
métodos gráficos

Diagrama de dispersión

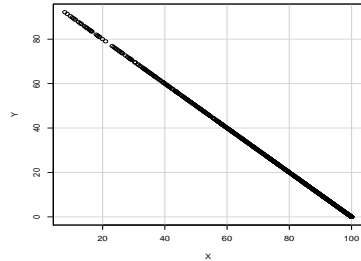
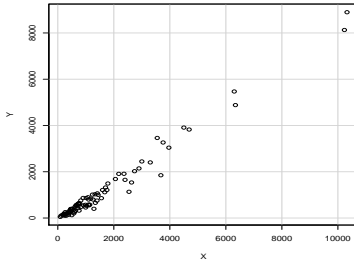
El diagrama de dispersión o nube de puntos es la **representación de los pares de valores observados** para dos variables cuantitativas en unos ejes cartesianos.

- Cada par de valores observados se representa por un punto en el plano.
- La disposición de los puntos en el plano nos da una primera idea del comportamiento conjunto de las variables.

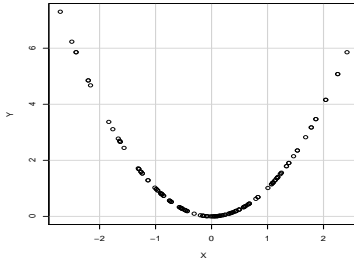
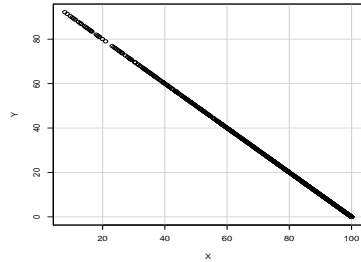
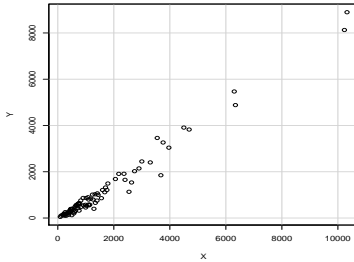
métodos gráficos



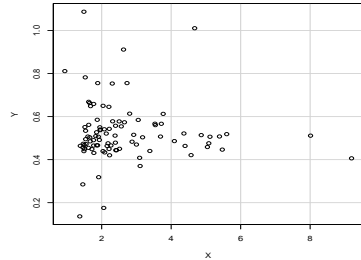
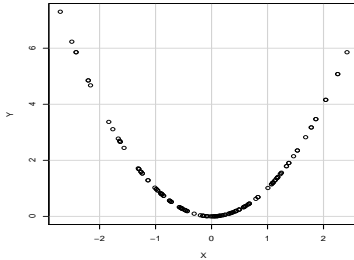
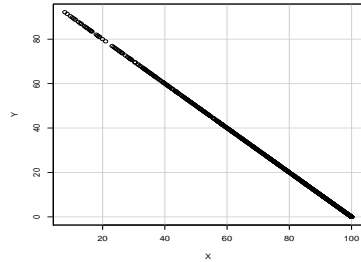
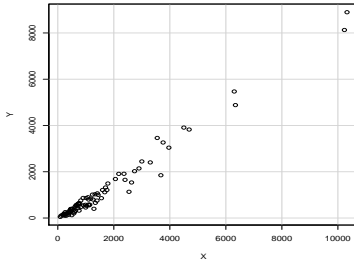
métodos gráficos



métodos gráficos



métodos gráficos



Contenidos

- 1 Introducción
- 2 Correlación
 - Coeficiente de correlación lineal
- 3 El modelo RLS
 - Formulación del modelo RLS
 - Estimación de los parámetros del modelo RLS
 - Inferencia estadística sobre los parámetros del modelo RLS
 - Coeficiente de determinación
- 4 Predicción
 - Predicción para un valor individual
 - Predicción para el valor medio
- 5 Diagnos
 - Hipótesis de normalidad
 - Hipótesis de Homocedasticidad
 - Hipótesis de Independencia

métodos analíticos

Definición

El coeficiente de correlación lineal de Pearson (r_{xy}) tiene por objeto medir el grado de asociación lineal entre dos variables, y viene dado por:

$$r = \frac{S_{xy}}{S_x S_y}$$

donde S_{xy} es la covarianza muestral entre X e Y , y S_x y S_y , las desviaciones típicas muestrales de X e Y , respectivamente.

métodos analíticos

Definición

El coeficiente de correlación lineal de Pearson (r_{xy}) tiene por objeto medir el grado de asociación lineal entre dos variables, y viene dado por:

$$r = \frac{S_{xy}}{S_x S_y}$$

donde S_{xy} es la covarianza muestral entre X e Y , y S_x y S_y , las desviaciones típicas muestrales de X e Y , respectivamente.

Interpretación: $-1 \leq r \leq 1$, entonces:

- Si $r = -1$, hay una asociación lineal negativa perfecta
- Si $r = 1$, hay una asociación lineal positiva perfecta
- Si $r = 0$, no hay asociación lineal entre las variables.

métodos analíticos

Además de calcular el coeficiente de correlación lineal de forma descriptiva, conviene determinar si existe o no una **relación estadísticamente significativa** entre las variables, es decir, evidenciar si los coeficientes de correlación poblacionales son significativamente distintos de cero o no.

Hipótesis nula $H_0 : r = 0$

Hipótesis alternativa $H_1 : r \neq 0$

métodos analíticos

Además de calcular el coeficiente de correlación lineal de forma descriptiva, conviene determinar si existe o no una **relación estadísticamente significativa** entre las variables, es decir, evidenciar si los coeficientes de correlación poblacionales son significativamente distintos de cero o no.

Hipótesis nula $H_0 : r = 0$

Hipótesis alternativa $H_1 : r \neq 0$

métodos analíticos

Además de calcular el coeficiente de correlación lineal de forma descriptiva, conviene determinar si existe o no una **relación estadísticamente significativa** entre las variables, es decir, evidenciar si los coeficientes de correlación poblacionales son significativamente distintos de cero o no.

Hipótesis nula $H_0 : r = 0$
Hipótesis alternativa $H_1 : r \neq 0$

Estadístico de contraste	$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$
región de rechazo	$ t > t_{n-2, 1-\alpha/2}$
p-valor	$p = 2P[t_{n-2} > t]$

métodos analíticos

Además de calcular el coeficiente de correlación lineal de forma descriptiva, conviene determinar si existe o no una **relación estadísticamente significativa** entre las variables, es decir, evidenciar si los coeficientes de correlación poblacionales son significativamente distintos de cero o no.

Hipótesis nula $H_0 : r = 0$
Hipótesis alternativa $H_1 : r \neq 0$

Estadístico de contraste	$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$
región de rechazo	$ t > t_{n-2, 1-\alpha/2}$
p-valor	$p = 2P[t_{n-2} > t]$

- Si aceptamos H_0 , decimos que el coeficiente de correlación lineal es cero y, por tanto, no existe una relación estadísticamente significativa entre las variables consideradas.

Ejemplo: Fichero DatosAndalucia.rda

Incluye las siguientes variables sobre los municipios de Andalucía en el año 2007

- Código INE del municipio (**Codigo.INE**)
- Nombre del municipio (**Municipio**)
- Edad media (**Edad.media.2007**)
- Nombre de la Provincia a la que pertenece (**Provincia**)
- número de líneas ADSL por cada 100 habitantes (**tasa.lineas.ADSL.2007**)
- Tasa de paro (**tasa.paro.2007**)

¿Existe relación lineal entre las variables?

¿Cómo de intensa es la relación lineal existente?

Ejemplo: Fichero DatosAndalucia.rda

Incluye las siguientes variables sobre los municipios de Andalucía en el año 2007

- Código INE del municipio (**Codigo.INE**)
- Nombre del municipio (**Municipio**)
- Edad media (**Edad.media.2007**)
- Nombre de la Provincia a la que pertenece (**Provincia**)
- número de líneas ADSL por cada 100 habitantes (**tasa.lineas.ADSL.2007**)
- Tasa de paro (**tasa.paro.2007**)

¿Existe relación lineal entre las variables?

¿Cómo de intensa es la relación lineal existente?

Ejemplo: Fichero DatosAndalucia.rda

Incluye las siguientes variables sobre los municipios de Andalucía en el año 2007

- Código INE del municipio (**Codigo.INE**)
- Nombre del municipio (**Municipio**)
- Edad media (**Edad.media.2007**)
- Nombre de la Provincia a la que pertenece (**Provincia**)
- número de líneas ADSL por cada 100 habitantes (**tasa.lineas.ADSL.2007**)
- Tasa de paro (**tasa.paro.2007**)

¿Existe relación lineal entre las variables?

¿Cómo de intensa es la relación lineal existente?

Ejemplo: Fichero DatosAndalucia.rda

Incluye las siguientes variables sobre los municipios de Andalucía en el año 2007

- Código INE del municipio (**Codigo.INE**)
- Nombre del municipio (**Municipio**)
- Edad media (**Edad.media.2007**)
- Nombre de la Provincia a la que pertenece (**Provincia**)
- número de líneas ADSL por cada 100 habitantes (**tasa.lineas.ADSL.2007**)
- Tasa de paro (**tasa.paro.2007**)

¿Existe relación lineal entre las variables?

¿Cómo de intensa es la relación lineal existente?

Ejemplo: Diagrama de dispersión

Gráficas → Diagrama de dispersión

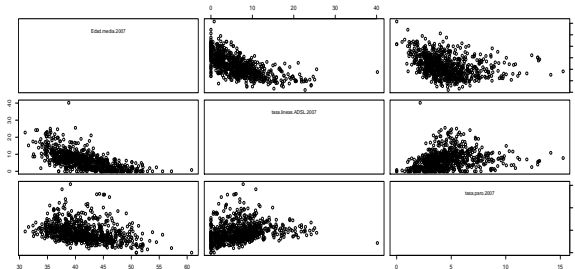
Gráficas → Matriz de diagramas de dispersión

- Posible relación lineal entre Edad media y Tasa líneas ADSL
- Posible relación no lineal entre Edad media y Tasa de paro
- Posible incorrelación entre Tasa de paro y Tasa líneas ADSL

Ejemplo: Diagrama de dispersión

Gráficas → Diagrama de dispersión

Gráficas → Matriz de diagramas de dispersión

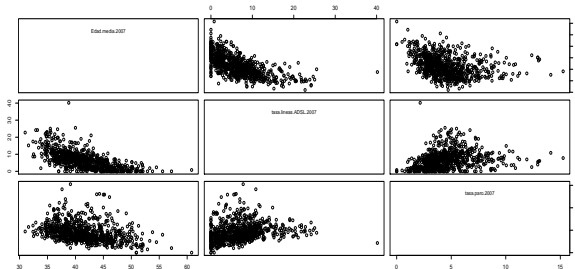


- Posible relación lineal entre Edad media y Tasa líneas ADSL
- Posible relación no lineal entre Edad media y Tasa de paro
- Posible incorrelación entre Tasa de paro y Tasa líneas ADSL

Ejemplo: Diagrama de dispersión

Gráficas → Diagrama de dispersión

Gráficas → Matriz de diagramas de dispersión

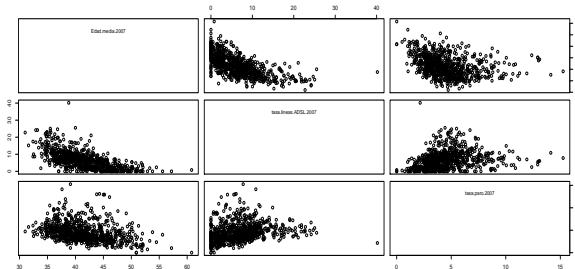


- Posible relación lineal entre Edad media y Tasa líneas ADSL
- Posible relación no lineal entre Edad media y Tasa de paro
- Posible incorrelación entre Tasa de paro y Tasa líneas ADSL

Ejemplo: Diagrama de dispersión

Gráficas → Diagrama de dispersión

Gráficas → Matriz de diagramas de dispersión

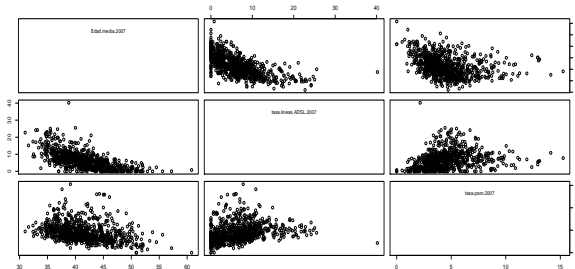


- Posible relación lineal entre Edad media y Tasa líneas ADSL
- Posible relación no lineal entre Edad media y Tasa de paro
- Posible incorrelación entre Tasa de paro y Tasa líneas ADSL

Ejemplo: Diagrama de dispersión

Gráficas → Diagrama de dispersión

Gráficas → Matriz de diagramas de dispersión



- Posible relación lineal entre Edad media y Tasa líneas ADSL
- Posible relación no lineal entre Edad media y Tasa de paro
- Posible incorrelación entre Tasa de paro y Tasa líneas ADSL

Ejemplo: Coeficiente de correlación lineal de Pearson

Estadísticos → Resúmenes → Matriz de correlaciones

- En la diagonal, obviamente aparecen unos
- La matriz es simétrica, ya que el coeficiente de correlación lineal también lo es
- Se observa relación directa entre la Tasa de paro y la Tasa líneas ADSL e inversa entre la Edad media y las otras dos variables
- El coeficiente entre la variable Tasa de paro con las otras dos indica relación lineal débil entre las variables. **¿Incorreladas?**

Ejemplo: Coeficiente de correlación lineal de Pearson

Estadísticos → Resúmenes → Matriz de correlaciones

	Edad media	Tasa líneas ADSL	Tasa paro
Edad media	1	-0.64	-0.32
Tasa líneas ADSL	-0.64	1	0.24
Tasa paro	-0.32	0.24	1

- En la diagonal, obviamente aparecen unos
- La matriz es simétrica, ya que el coeficiente de correlación lineal también lo es
- Se observa relación directa entre la Tasa de paro y la Tasa líneas ADSL e inversa entre la Edad media y las otras dos variables
- El coeficiente entre la variable Tasa de paro con las otras dos indica relación lineal débil entre las variables. **¿Incorreladas?**

Ejemplo: Coeficiente de correlación lineal de Pearson

Estadísticos → Resúmenes → Matriz de correlaciones

	Edad media	Tasa líneas ADSL	Tasa paro
Edad media	1	-0.64	-0.32
Tasa líneas ADSL	-0.64	1	0.24
Tasa paro	-0.32	0.24	1

- En la diagonal, obviamente aparecen unos
- La matriz es simétrica, ya que el coeficiente de correlación lineal también lo es
- Se observa relación directa entre la Tasa de paro y la Tasa líneas ADSL e inversa entre la Edad media y las otras dos variables
- El coeficiente entre la variable Tasa de paro con las otras dos indica relación lineal débil entre las variables. **¿Incorreladas?**

Ejemplo: Coeficiente de correlación lineal de Pearson

Estadísticos → Resúmenes → Matriz de correlaciones

	Edad media	Tasa líneas ADSL	Tasa paro
Edad media	1	-0.64	-0.32
Tasa líneas ADSL	-0.64	1	0.24
Tasa paro	-0.32	0.24	1

- En la diagonal, obviamente aparecen unos
- La matriz es simétrica, ya que el coeficiente de correlación lineal también lo es
- Se observa relación directa entre la Tasa de paro y la Tasa líneas ADSL e inversa entre la Edad media y las otras dos variables
- El coeficiente entre la variable Tasa de paro con las otras dos indica relación lineal débil entre las variables. **¿Incorreladas?**

Ejemplo: Coeficiente de correlación lineal de Pearson

Estadísticos → Resúmenes → Matriz de correlaciones

	Edad media	Tasa líneas ADSL	Tasa paro
Edad media	1	-0.64	-0.32
Tasa líneas ADSL	-0.64	1	0.24
Tasa paro	-0.32	0.24	1

- En la diagonal, obviamente aparecen unos
- La matriz es simétrica, ya que el coeficiente de correlación lineal también lo es
- Se observa relación directa entre la Tasa de paro y la Tasa líneas ADSL e inversa entre la Edad media y las otras dos variables
- El coeficiente entre la variable Tasa de paro con las otras dos indica relación lineal débil entre las variables. **¿Incorreladas?**

Ejemplo: Coeficiente de correlación lineal de Pearson

Estadísticos → Resúmenes → Matriz de correlaciones

	Edad media	Tasa líneas ADSL	Tasa paro
Edad media	1	-0.64	-0.32
Tasa líneas ADSL	-0.64	1	0.24
Tasa paro	-0.32	0.24	1

- En la diagonal, obviamente aparecen unos
- La matriz es simétrica, ya que el coeficiente de correlación lineal también lo es
- Se observa relación directa entre la Tasa de paro y la Tasa líneas ADSL e inversa entre la Edad media y las otras dos variables
- El coeficiente entre la variable Tasa de paro con las otras dos indica relación lineal débil entre las variables. **¿Incorreladas?**

Ejemplo: Test de correlación

Estadísticos → Resúmenes → Test de correlación

- El p-valor es muy próximo a cero → se rechaza la Hipótesis nula ($r = 0$) en favor de la alternativa ($r \neq 0$)
- El I.C. no contiene al cero, como cabía esperar

Ejemplo: Test de correlación

Estadísticos → Resúmenes → Test de correlación

Resultados del Test: Tasa de paro vs. Tasa líneas ADSL

Estad. contraste	grados libertad	p-valor
$t = 6.9677$	$df = 768$	$p\text{-value} = 6.934e-12$

- El p-valor es muy próximo a cero → se rechaza la Hipótesis nula ($r = 0$) en favor de la alternativa ($r \neq 0$)
- El I.C. no contiene al cero, como cabía esperar

Ejemplo: Test de correlación

Estadísticos → Resúmenes → Test de correlación

Resultados del Test: Tasa de paro vs. Tasa líneas ADSL

Estad. contraste	grados libertad	p-valor
$t = 6.9677$	$df = 768$	$p\text{-value} = 6.934e-12$

- El p-valor es muy próximo a cero → se rechaza la Hipótesis nula ($r = 0$) en favor de la alternativa ($r \neq 0$)
- El I.C. no contiene al cero, como cabía esperar

Ejemplo: Test de correlación

Estadísticos → Resúmenes → Test de correlación

Resultados del Test: Tasa de paro vs. Tasa líneas ADSL

Estad. contraste	grados libertad	p-valor
$t = 6.9677$	$df = 768$	$p\text{-value} = 6.934e-12$

- El p-valor es muy próximo a cero → se rechaza la Hipótesis nula ($r = 0$) en favor de la alternativa ($r \neq 0$)

Intervalo de confianza al 95 %: $[0.1762211, 0.3091636]$

- El I.C. no contiene al cero, como cabía esperar

Ejemplo: Test de correlación

Estadísticos → Resúmenes → Test de correlación

Resultados del Test: Tasa de paro vs. Tasa líneas ADSL

Estad. contraste	grados libertad	p-valor
$t = 6.9677$	$df = 768$	$p\text{-value} = 6.934e-12$

- El p-valor es muy próximo a cero → se rechaza la Hipótesis nula ($r = 0$) en favor de la alternativa ($r \neq 0$)

Intervalo de confianza al 95 %: $[0.1762211, 0.3091636]$

- El I.C. no contiene al cero, como cabía esperar

Ejemplo: Test de correlación

Estadísticos → Resúmenes → Test de correlación

Resultados del Test: Tasa de paro vs. Tasa líneas ADSL

Estad. contraste	grados libertad	p-valor
$t = 6.9677$	$df = 768$	$p\text{-value} = 6.934e-12$

- El p-valor es muy próximo a cero → se rechaza la Hipótesis nula ($r = 0$) en favor de la alternativa ($r \neq 0$)

Intervalo de confianza al 95 %: $[0.1762211, 0.3091636]$

- El I.C. no contiene al cero, como cabía esperar

Las variables **no** son **incorreladas**.

Desde el punto de vista estadístico, existe **relación lineal significativa** (aunque **no intensa**) entre las variables

Contenidos

- 1 Introducción
- 2 Correlación
 - Coeficiente de correlación lineal
- 3 El modelo RLS
 - Formulación del modelo RLS
 - Estimación de los parámetros del modelo RLS
 - Inferencia estadística sobre los parámetros del modelo RLS
 - Coeficiente de determinación
- 4 Predicción
 - Predicción para un valor individual
 - Predicción para el valor medio
- 5 Diagnos
 - Hipótesis de normalidad
 - Hipótesis de Homocedasticidad
 - Hipótesis de Independencia

Contenidos

- 1 Introducción
- 2 Correlación
 - Coeficiente de correlación lineal
- 3 **El modelo RLS**
 - **Formulación del modelo RLS**
 - Estimación de los parámetros del modelo RLS
 - Inferencia estadística sobre los parámetros del modelo RLS
 - Coeficiente de determinación
- 4 Predicción
 - Predicción para un valor individual
 - Predicción para el valor medio
- 5 Diagnos
 - Hipótesis de normalidad
 - Hipótesis de Homocedasticidad
 - Hipótesis de Independencia

3.1. Formulación del modelo RLS

Modelo teórico

$$Y = \beta_0 + \beta_1 X + U$$

3.1. Formulación del modelo RLS

Modelo teórico

$$Y = \beta_0 + \beta_1 X + U$$

A partir de n observaciones, se tiene:

$$y_i = \beta_0 + \beta_1 x_i + u_i, \quad i = 1, 2, \dots, n$$

donde

- El término u_i se denomina **error** o **residuo** y contiene todos los factores que afectan a y_i que no son x_i
 - β_0 término independiente o constante
 - β_1 pendiente del modelo
- } Coeficientes de regresión
(parámetros desconocidos)

3.1. Formulación del modelo RLS

Objetivo

Encontrar los estimadores de los parámetros β_0 y β_1 , que denotamos por $\hat{\beta}_0$ y $\hat{\beta}_1$, que dan lugar a la **recta de regresión**

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i, \quad i = 1, 2, \dots, n$$



Hipótesis sobre el modelo $y_i = \beta_0 + \beta_1 x_i + u_i$

- 1 Media cero: $E[u_i] = 0$, para todo $i = 1, 2, \dots, n$.
- 2 Homocedasticidad: $Var[u_i] = \sigma_u^2$ para todo $i = 1, 2, \dots, n$.
- 3 Normalidad: $u_i \rightsquigarrow N(0, \sigma_u)$
- 4 Independencia: u_i y u_j independientes para cualquier $i \neq j$ o, en general, hay *ausencia de autocorrelación* entre los términos aleatorios: $Cov(u_i, u_j) = 0$, para cualquier $i \neq j$.

3.1. Formulación del modelo RLS

Objetivo

Encontrar los estimadores de los parámetros β_0 y β_1 , que denotamos por $\hat{\beta}_0$ y $\hat{\beta}_1$, que dan lugar a la **recta de regresión**

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i, \quad i = 1, 2, \dots, n$$



Hipótesis sobre el modelo $y_i = \beta_0 + \beta_1 x_i + u_i$

- 1 **Media cero:** $E[u_i] = 0$, para todo $i = 1, 2, \dots, n$.
- 2 **Homocedasticidad:** $Var[u_i] = \sigma_u^2$ para todo $i = 1, 2, \dots, n$.
- 3 **Normalidad:** $u_i \rightsquigarrow N(0, \sigma_u)$
- 4 **Independencia:** u_i y u_j independientes para cualquier $i \neq j$ o, en general, hay *ausencia de autocorrelación* entre los términos aleatorios: $Cov(u_i, u_j) = 0$, para cualquier $i \neq j$.

3.1. Formulación del modelo RLS

Objetivo

Encontrar los estimadores de los parámetros β_0 y β_1 , que denotamos por $\hat{\beta}_0$ y $\hat{\beta}_1$, que dan lugar a la **recta de regresión**

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i, \quad i = 1, 2, \dots, n$$



Hipótesis sobre el modelo $y_i = \beta_0 + \beta_1 x_i + u_i$

- 1 **Media cero:** $E[u_i] = 0$, para todo $i = 1, 2, \dots, n$.
- 2 **Homocedasticidad:** $Var[u_i] = \sigma_u^2$ para todo $i = 1, 2, \dots, n$.
- 3 **Normalidad:** $u_i \sim N(0, \sigma_u)$
- 4 **Independencia:** u_i y u_j independientes para cualquier $i \neq j$ o, en general, hay *ausencia de autocorrelación* entre los términos aleatorios: $Cov(u_i, u_j) = 0$, para cualquier $i \neq j$.

Contenidos

- 1 Introducción
- 2 Correlación
 - Coeficiente de correlación lineal
- 3 El modelo RLS
 - Formulación del modelo RLS
 - **Estimación de los parámetros del modelo RLS**
 - Inferencia estadística sobre los parámetros del modelo RLS
 - Coeficiente de determinación
- 4 Predicción
 - Predicción para un valor individual
 - Predicción para el valor medio
- 5 Diagnos
 - Hipótesis de normalidad
 - Hipótesis de Homocedasticidad
 - Hipótesis de Independencia

Estimadores mínimos cuadráticos

Siguiendo el criterio de los mínimos cuadrados, basado en la idea de minimizar los residuos \hat{u}_i , se obtienen los estimadores:

Estimadores mínimos cuadráticos (EMC)

$$\hat{\beta}_0 = \bar{y} - \frac{S_{xy}}{S_x^2} \bar{x} \qquad \hat{\beta}_1 = \frac{S_{xy}}{S_x^2}$$

donde

$$S_{xy} = \frac{1}{n} \sum_{i=1}^n x_i y_i - \bar{x} \bar{y}$$

$$S_x^2 = \frac{1}{n} \sum_{i=1}^n x_i^2 - \bar{x}^2$$

Estimador de σ_u^2

Para realizar inferencias (► I.C. , ► C.H.) respecto a los parámetros β_0 y β_1 , es preciso disponer de un estimador para σ_u^2 .

Varianza residual o cuadrado medio del error (CME)

$$\hat{\sigma}_u^2 = S_{\hat{u}}^2 = \frac{SCE}{n-2} = \frac{1}{n-2} \sum_{i=1}^n \hat{u}_i^2 = \frac{1}{n-2} \sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$$

- A $\hat{\sigma}_u$ se le denomina **error estándar de regresión** y tiene las mismas unidades que la variable respuesta.
- $\hat{\sigma}_u$ puede utilizarse como indicador del **grado de ajuste del modelo de regresión**, aunque al tener la mismas unidades de medida que Y no nos va a permitir comparar la bondad del ajuste de dos modelos con variable endógena diferente.

Estimador de σ_u^2

Para realizar inferencias (► I.C. , ► C.H.) respecto a los parámetros β_0 y β_1 , es preciso disponer de un estimador para σ_u^2 .

Varianza residual o cuadrado medio del error (CME)

$$\hat{\sigma}_u^2 = S_{\hat{u}}^2 = \frac{SCE}{n-2} = \frac{1}{n-2} \sum_{i=1}^n \hat{u}_i^2 = \frac{1}{n-2} \sum_{i=1}^n \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i \right)^2$$

- A $\hat{\sigma}_u$ se le denomina **error estándar de regresión** y tiene las mismas unidades que la variable respuesta.
- $\hat{\sigma}_u$ puede utilizarse como indicador del **grado de ajuste del modelo de regresión**, aunque al tener la mismas unidades de medida que Y no nos va a permitir comparar la bondad del ajuste de dos modelos con variable endógena diferente.

Estimador de σ_u^2

Para realizar inferencias (▶ I.C. , ▶ C.H.) respecto a los parámetros β_0 y β_1 , es preciso disponer de un estimador para σ_u^2 .

Varianza residual o cuadrado medio del error (CME)

$$\hat{\sigma}_u^2 = S_{\hat{u}}^2 = \frac{SCE}{n-2} = \frac{1}{n-2} \sum_{i=1}^n \hat{u}_i^2 = \frac{1}{n-2} \sum_{i=1}^n \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i \right)^2$$

- A $\hat{\sigma}_u$ se le denomina **error estándar de regresión** y tiene las mismas unidades que la variable respuesta.
- $\hat{\sigma}_u$ puede utilizarse como indicador del **grado de ajuste del modelo de regresión**, aunque al tener la mismas unidades de medida que Y no nos va a permitir comparar la bondad del ajuste de dos modelos con variable endógena diferente.

Estimador de σ_u^2

Para realizar inferencias (▶ I.C. , ▶ C.H.) respecto a los parámetros β_0 y β_1 , es preciso disponer de un estimador para σ_u^2 .

Varianza residual o cuadrado medio del error (CME)

$$\hat{\sigma}_u^2 = S_{\hat{u}}^2 = \frac{SCE}{n-2} = \frac{1}{n-2} \sum_{i=1}^n \hat{u}_i^2 = \frac{1}{n-2} \sum_{i=1}^n \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i \right)^2$$

- A $\hat{\sigma}_u$ se le denomina **error estándar de regresión** y tiene las mismas unidades que la variable respuesta.
- $\hat{\sigma}_u$ puede utilizarse como indicador del **grado de ajuste del modelo de regresión**, aunque al tener la mismas unidades de medida que Y no nos va a permitir comparar la bondad del ajuste de dos modelos con variable endógena diferente.

Contenidos

- 1 Introducción
- 2 Correlación
 - Coeficiente de correlación lineal
- 3 El modelo RLS
 - Formulación del modelo RLS
 - Estimación de los parámetros del modelo RLS
 - **Inferencia estadística sobre los parámetros del modelo RLS**
 - Coeficiente de determinación
- 4 Predicción
 - Predicción para un valor individual
 - Predicción para el valor medio
- 5 Diagnos
 - Hipótesis de normalidad
 - Hipótesis de Homocedasticidad
 - Hipótesis de Independencia

C.H. para los coeficientes de regresión

Nos interesamos por *verificar* o *contrastar* Hipótesis acerca del valor de un parámetro determinado.

Problema de interés

Tras ajustar una recta por el método de los mínimos cuadrados, observamos que la estimación de la pendiente es $\hat{\beta}_1 = 0,001$.

- ¿El valor estimado representa la autentica relación existente entre las variables o, por el contrario, ese valor es cero y el obtenido se debe sencillamente a la presencia de perturbaciones en el modelo?
- La respuesta a esta cuestión pasa, ineludiblemente, por el planteamiento del contraste:

Hipótesis nula $H_0 : \beta_1 = 0$

Hipótesis alternativa $H_1 : \beta_1 \neq 0$

- Si aceptamos H_0 , decimos que dicho coeficiente no es significativo en el modelo (lo cual indicaría ausencia de relación entre X e Y).

C.H. para los coeficientes de regresión

Nos interesamos por *verificar* o *contrastar* Hipótesis acerca del valor de un parámetro determinado.

Problema de interés

Tras ajustar una recta por el método de los mínimos cuadrados, observamos que la estimación de la pendiente es $\hat{\beta}_1 = 0,001$.

- ¿El valor estimado representa la autentica relación existente entre las variables o, por el contrario, ese valor es cero y el obtenido se debe sencillamente a la presencia de perturbaciones en el modelo?
- La respuesta a esta cuestión pasa, ineludiblemente, por el planteamiento del contraste:

Hipótesis nula $H_0 : \beta_1 = 0$

Hipótesis alternativa $H_1 : \beta_1 \neq 0$

- Si aceptamos H_0 , decimos que dicho coeficiente no es significativo en el modelo (lo cual indicaría ausencia de relación entre X e Y).

C.H. para los coeficientes de regresión

Hipótesis nula	Estadístico bajo H_0	
$H_0 : \beta_i = \beta_i^0$	$t = \frac{\hat{\beta}_i - \beta_i^0}{S_{\hat{\beta}_i}}$	
Hipótesis alternativa	región crítica	p-valor
$H_1 : \beta_i \neq \beta_i^0$ $H_1 : \beta_i > \beta_i^0$ $H_1 : \beta_i < \beta_i^0$	$ t > t_{n-2, 1-\alpha/2}$ $t > t_{n-2, 1-\alpha}$ $t < -t_{n-2, 1-\alpha}$	$p = 2P[t_{n-2} > t]$ $p = P[t_{n-2} > t]$ $p = P[t_{n-2} < t]$

C.H. para β_0

$$H_0 : \beta_0 = 0$$

$$H_1 : \beta_0 \neq 0$$

C.H. para β_1

$$H_0 : \beta_1 = 0$$

$$H_1 : \beta_1 \neq 0$$

C.H. para los coeficientes de regresión

Hipótesis nula	Estadístico bajo H_0	
$H_0 : \beta_i = \beta_i^0$	$t = \frac{\hat{\beta}_i - \beta_i^0}{S_{\hat{\beta}_i}}$	
Hipótesis alternativa	región crítica	p-valor
$H_1 : \beta_i \neq \beta_i^0$ $H_1 : \beta_i > \beta_i^0$ $H_1 : \beta_i < \beta_i^0$	$ t > t_{n-2, 1-\alpha/2}$ $t > t_{n-2, 1-\alpha}$ $t < -t_{n-2, 1-\alpha}$	$p = 2P[t_{n-2} > t]$ $p = P[t_{n-2} > t]$ $p = P[t_{n-2} < t]$



C.H. para β_0

$$H_0 : \beta_0 = 0$$

$$H_1 : \beta_0 \neq 0$$

C.H. para β_1

$$H_0 : \beta_1 = 0$$

$$H_1 : \beta_1 \neq 0$$

I.C. para los coeficientes de regresión

I.C. para β_i al n.c. $(1 - \alpha) \%$

$$\left[\hat{\beta}_i - t_{n-2, 1-\alpha/2} S_{\hat{\beta}_i}, \hat{\beta}_i + t_{n-2, 1-\alpha/2} S_{\hat{\beta}_i} \right]$$

con $S_{\hat{\beta}_i}$ el estimador de la desviación típica de la distribución de $\hat{\beta}_i$.

I.C. para β_0 al n.c. $(1 - \alpha) \%$

$$\left[\hat{\beta}_0 \pm t_{n-2, 1-\alpha/2} S_{\hat{\beta}_0} \right]$$

donde $S_{\hat{\beta}_0} = \sqrt{S_{\hat{u}}^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{n S_x^2} \right)}$.

I.C. para β_1 al n.c. $(1 - \alpha) \%$

$$\left[\hat{\beta}_1 \pm t_{n-2, 1-\alpha/2} S_{\hat{\beta}_1} \right]$$

donde $S_{\hat{\beta}_1} = \sqrt{S_{\hat{u}}^2 / n S_x^2}$.

• $S_{\hat{\beta}_0}^2$

I.C. para los coeficientes de regresión

I.C. para β_i al n.c. $(1 - \alpha) \%$

$$\left[\hat{\beta}_i - t_{n-2, 1-\alpha/2} S_{\hat{\beta}_i}, \hat{\beta}_i + t_{n-2, 1-\alpha/2} S_{\hat{\beta}_i} \right]$$

con $S_{\hat{\beta}_i}$ el estimador de la desviación típica de la distribución de $\hat{\beta}_i$.



I.C. para β_0 al n.c. $(1 - \alpha) \%$

$$\left[\hat{\beta}_0 \pm t_{n-2, 1-\alpha/2} S_{\hat{\beta}_0} \right]$$

donde $S_{\hat{\beta}_0} = \sqrt{S_{\hat{u}}^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{n S_x^2} \right)}$.



I.C. para β_1 al n.c. $(1 - \alpha) \%$

$$\left[\hat{\beta}_1 \pm t_{n-2, 1-\alpha/2} S_{\hat{\beta}_1} \right]$$

donde $S_{\hat{\beta}_1} = \sqrt{S_{\hat{u}}^2 / n S_x^2}$.

→ $S_{\hat{u}}^2$

Contenidos

- 1 Introducción
- 2 Correlación
 - Coeficiente de correlación lineal
- 3 El modelo RLS
 - Formulación del modelo RLS
 - Estimación de los parámetros del modelo RLS
 - Inferencia estadística sobre los parámetros del modelo RLS
 - Coeficiente de determinación
- 4 Predicción
 - Predicción para un valor individual
 - Predicción para el valor medio
- 5 Diagnos
 - Hipótesis de normalidad
 - Hipótesis de Homocedasticidad
 - Hipótesis de Independencia

- Tras ajustar un modelo de Regresión a unos datos debemos de preguntarnos si el modelo ajustado es o no útil.
- La evaluación global de una recta de Regresión puede hacerse mediante la **varianza residual** $S_{\hat{u}}^2$: La línea de Regresión sería poco representativa cuando la varianza residual sea grande (las diferencias entre los valores ajustados y observados, los errores, son grandes).
- Como $S_{\hat{u}}^2$ depende de las unidades de medida de la variable dependiente, no es una medida útil para comparar rectas de Regresión de variables distintas.
- Como medida más adecuada del ajuste, se utiliza **coeficiente de determinación**, que expresa la proporción o grado de variabilidad de la variable dependiente explicada por el modelo lineal ajustado:

$$R^2 = \frac{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}{S_y^2} = \frac{SCR}{SCT} = 1 - \frac{SCE}{SCT}$$

- Tras ajustar un modelo de Regresión a unos datos debemos de preguntarnos si el modelo ajustado es o no útil.
- La evaluación global de una recta de Regresión puede hacerse mediante la **varianza residual** $S_{\hat{u}}^2$: La línea de Regresión sería poco representativa cuando la varianza residual sea grande (las diferencias entre los valores ajustados y observados, los errores, son grandes).
- Como $S_{\hat{u}}^2$ depende de las unidades de medida de la variable dependiente, no es una medida útil para comparar rectas de Regresión de variables distintas.
- Como medida más adecuada del ajuste, se utiliza **coeficiente de determinación**, que expresa la proporción o grado de variabilidad de la variable dependiente explicada por el modelo lineal ajustado:

$$R^2 = \frac{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}{S_y^2} = \frac{SCR}{SCT} = 1 - \frac{SCE}{SCT}$$

- Tras ajustar un modelo de Regresión a unos datos debemos de preguntarnos si el modelo ajustado es o no útil.
- La evaluación global de una recta de Regresión puede hacerse mediante la **varianza residual** $S_{\hat{u}}^2$: La línea de Regresión sería poco representativa cuando la varianza residual sea grande (las diferencias entre los valores ajustados y observados, los errores, son grandes).
- Como $S_{\hat{u}}^2$ depende de las unidades de medida de la variable dependiente, no es una medida útil para comparar rectas de Regresión de variables distintas.
- Como medida más adecuada del ajuste, se utiliza **coeficiente de determinación**, que expresa la proporción o grado de variabilidad de la variable dependiente explicada por el modelo lineal ajustado:

$$R^2 = \frac{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}{S_y^2} = \frac{SCR}{SCT} = 1 - \frac{SCE}{SCT}$$

- Tras ajustar un modelo de Regresión a unos datos debemos de preguntarnos si el modelo ajustado es o no útil.
- La evaluación global de una recta de Regresión puede hacerse mediante la **varianza residual** $S_{\hat{u}}^2$: La línea de Regresión sería poco representativa cuando la varianza residual sea grande (las diferencias entre los valores ajustados y observados, los errores, son grandes).
- Como $S_{\hat{u}}^2$ depende de las unidades de medida de la variable dependiente, no es una medida útil para comparar rectas de Regresión de variables distintas.
- Como medida más adecuada del ajuste, se utiliza **coeficiente de determinación**, que expresa la proporción o grado de variabilidad de la variable dependiente explicada por el modelo lineal ajustado:

$$R^2 = \frac{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}{S_y^2} = \frac{SCR}{SCT} = 1 - \frac{SCE}{SCT}$$

Propiedades

- 1 En el caso de la Regresión lineal:

$$R^2 = \hat{\beta}_1^2 \frac{S_x^2}{S_y^2} = \frac{S_{xy}^2}{S_x^2 S_y^2}$$

- 2 Interpretación: $0 \leq R^2 \leq 1$, entonces:

- Si la recta ajustada pasa por todos los puntos observados $\Rightarrow SCE = 0 \Rightarrow R^2 = 1$, es decir, el modelo de Regresión explica el 100 % de la variabilidad de Y .
- Si la variable independiente (X) no explica ninguna variación de la variable dependiente (Y) $\Rightarrow SCR = 0 \Rightarrow R^2 = 0$.
- En general, si el ajuste es suficientemente bueno, R^2 será próximo a 1. **La implicación contraria no es cierta.**

- 3 **ATENCIÓN!!!!** R^2 no mide la validez del modelo de Regresión propuesto sino la proporción de la variación total que se explica mediante la ecuación de Regresión estimada.

Propiedades

- ① En el caso de la Regresión lineal:

$$R^2 = \hat{\beta}_1^2 \frac{S_x^2}{S_y^2} = \frac{S_{xy}^2}{S_x^2 S_y^2}$$

- ② Interpretación: $0 \leq R^2 \leq 1$, entonces:

- Si la recta ajustada pasa por todos los puntos observados $\Rightarrow SCE = 0 \Rightarrow R^2 = 1$, es decir, el modelo de Regresión explica el 100 % de la variabilidad de Y .
- Si la variable independiente (X) no explica ninguna variación de la variable dependiente (Y) $\Rightarrow SCR = 0 \Rightarrow R^2 = 0$.
- En general, si el ajuste es suficientemente bueno, R^2 será próximo a 1. **La implicación contraria no es cierta.**

- ③ **ATENCIÓN!!!!** R^2 no mide la validez del modelo de Regresión propuesto sino la proporción de la variación total que se explica mediante la ecuación de Regresión estimada.

Propiedades

- ① En el caso de la Regresión lineal:

$$R^2 = \hat{\beta}_1^2 \frac{S_x^2}{S_y^2} = \frac{S_{xy}^2}{S_x^2 S_y^2}$$

- ② Interpretación: $0 \leq R^2 \leq 1$, entonces:

- Si la recta ajustada pasa por todos los puntos observados $\Rightarrow SCE = 0 \Rightarrow R^2 = 1$, es decir, el modelo de Regresión explica el 100 % de la variabilidad de Y .
- Si la variable independiente (X) no explica ninguna variación de la variable dependiente (Y) $\Rightarrow SCR = 0 \Rightarrow R^2 = 0$.
- En general, si el ajuste es suficientemente bueno, R^2 será próximo a 1. **La implicación contraria no es cierta.**

- ③ **ATENCIÓN!!!!** R^2 no mide la validez del modelo de Regresión propuesto sino la proporción de la variación total que se explica mediante la ecuación de Regresión estimada.

Ejemplo: Fichero DatosAndalucia.rda

Ajustar un modelo lineal de Regresión que represente la tasa de líneas ADSL en función de la edad media del municipio

Recordemos que $r = -0,643 \Rightarrow$ existe una tendencia lineal bastante intensa a que los municipios de más edad tengan menos líneas ADSL por habitante

¿Qué sentido tiene ajustar la recta de Regresión a estas variables?

- Analizar en qué medida afectaría la implantación de líneas ADSL ante un cambio en la edad media poblacional
- Predecir algún dato desconocido sobre la tasa de líneas ADSL en un municipio
- Estudiar las diferencias entre lo observado y lo esperado, y tratar de explicar estas diferencias

Ejemplo: Fichero DatosAndalucia.rda

Ajustar un modelo lineal de Regresión que represente la tasa de líneas ADSL en función de la edad media del municipio

Recordemos que $r = -0,643 \Rightarrow$ existe una tendencia lineal bastante intensa a que los municipios de más edad tengan menos líneas ADSL por habitante

¿Qué sentido tiene ajustar la recta de Regresión a estas variables?

- Analizar en qué medida afectaría la implantación de líneas ADSL ante un cambio en la edad media poblacional
- Predecir algún dato desconocido sobre la tasa de líneas ADSL en un municipio
- Estudiar las diferencias entre lo observado y lo esperado, y tratar de explicar estas diferencias

Ejemplo: Fichero DatosAndalucia.rda

Ajustar un modelo lineal de Regresión que represente la tasa de líneas ADSL en función de la edad media del municipio

Recordemos que $r = -0,643 \Rightarrow$ existe una tendencia lineal bastante intensa a que los municipios de más edad tengan menos líneas ADSL por habitante

¿Qué sentido tiene ajustar la recta de Regresión a estas variables?

- Analizar en qué medida afectaría la implantación de líneas ADSL ante un cambio en la edad media poblacional
- Predecir algún dato desconocido sobre la tasa de líneas ADSL en un municipio
- Estudiar las diferencias entre lo observado y lo esperado, y tratar de explicar estas diferencias

Ejemplo: Fichero DatosAndalucia.rda

Ajustar un modelo lineal de Regresión que represente la tasa de líneas ADSL en función de la edad media del municipio

Recordemos que $r = -0,643 \Rightarrow$ existe una tendencia lineal bastante intensa a que los municipios de más edad tengan menos líneas ADSL por habitante

¿Qué sentido tiene ajustar la recta de Regresión a estas variables?

- Analizar en qué medida afectaría la implantación de líneas ADSL ante un cambio en la edad media poblacional
- Predecir algún dato desconocido sobre la tasa de líneas ADSL en un municipio
- Estudiar las diferencias entre lo observado y lo esperado, y tratar de explicar estas diferencias

Ejemplo: Fichero DatosAndalucia.rda

Ajustar un modelo lineal de Regresión que represente la tasa de líneas ADSL en función de la edad media del municipio

Recordemos que $r = -0,643 \Rightarrow$ existe una tendencia lineal bastante intensa a que los municipios de más edad tengan menos líneas ADSL por habitante

¿Qué sentido tiene ajustar la recta de Regresión a estas variables?

- Analizar en qué medida afectaría la implantación de líneas ADSL ante un cambio en la edad media poblacional
- Predecir algún dato desconocido sobre la tasa de líneas ADSL en un municipio
- Estudiar las diferencias entre lo observado y lo esperado, y tratar de explicar estas diferencias

Ejemplo: Fichero DatosAndalucia.rda

Ajustar un modelo lineal de Regresión que represente la tasa de líneas ADSL en función de la edad media del municipio

Recordemos que $r = -0,643 \Rightarrow$ existe una tendencia lineal bastante intensa a que los municipios de más edad tengan menos líneas ADSL por habitante

¿Qué sentido tiene ajustar la recta de Regresión a estas variables?

- Analizar en qué medida afectaría la implantación de líneas ADSL ante un cambio en la edad media poblacional
- Predecir algún dato desconocido sobre la tasa de líneas ADSL en un municipio
- Estudiar las diferencias entre lo observado y lo esperado, y tratar de explicar estas diferencias

Ejemplo: Ajuste de la recta de regresión

Estadísticos → Ajuste de modelos → Regresión lineal

- Variable explicada (o dependiente) - y : Tasa líneas ADSL
- Variable explicativa (o independiente) - x : Edad media

Ejemplo: Ajuste de la recta de regresión

Estadísticos → Ajuste de modelos → Regresión lineal

- **Variable explicada (o dependiente) - y** : Tasa líneas ADSL
- **Variable explicativa (o independiente) - x** : Edad media

Ejemplo: Ajuste de la recta de regresión

Estadísticos → Ajuste de modelos → Regresión lineal

- **Variable explicada (o dependiente) - y** : Tasa líneas ADSL
- **Variable explicativa (o independiente) - x**: Edad media

```
Residuals:
    Min       1Q   Median       3Q      Max
-10.758   -2.538   -0.204    1.821   31.388

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  38.35290    1.36416   28.11  <2e-16 ***
Edad.media.2007 -0.76192    0.03271  -23.29  <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.035 on 768 degrees of freedom
Multiple R-squared:  0.4139, Adjusted R-squared:  0.4132
F-statistic: 542.5 on 1 and 768 DF,  p-value: < 2.2e-16
```

Resumen Estadístico de los residuos u_i

Ejemplo: Ajuste de la recta de regresión

Estadísticos → Ajuste de modelos → Regresión lineal

- **Variable explicada (o dependiente) - y** : Tasa líneas ADSL
- **Variable explicativa (o independiente) - x**: Edad media

Coeficientes del modelo RLS

```
Residuals:
    Min       1Q   Median       3Q      Max
-10.758  -2.538  -0.204   1.821   31.388

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  38.35290    1.36416   28.11  <2e-16 ***
Edad.media.2007 -0.76192    0.03271  -23.29  <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.035 on 768 degrees of freedom
Multiple R-squared:  0.4139, Adjusted R-squared:  0.4132
F-statistic: 542.5 on 1 and 768 DF,  p-value: < 2.2e-16
```

$$\bullet \beta_0 = 38.35, \beta_1 = -0.76192$$



$$\hat{y} = 38.35 - 0.76192x$$

$$\bullet S_{\hat{\beta}_0} = 1.364, S_{\hat{\beta}_1} = 0.033$$

$$\bullet \left\{ \begin{array}{l} H_0 : \beta_0 = 0 \\ H_1 : \beta_0 \neq 0 \end{array} \right\} \Rightarrow t = 28.11$$

$$\bullet \left\{ \begin{array}{l} H_0 : \beta_1 = 0 \\ H_1 : \beta_1 \neq 0 \end{array} \right\} \Rightarrow t = -23.29$$

La recta ajustada aparece especificada a través de sus coeficientes.

Ejemplo: Ajuste de la recta de regresión

Estadísticos → Ajuste de modelos → Regresión lineal

- **Variable explicada (o dependiente) - y** : Tasa líneas ADSL
- **Variable explicativa (o independiente) - x**: Edad media

Coeficientes del modelo RLS

```
Residuals:
    Min       1Q   Median       3Q      Max
-10.758  -2.538  -0.204   1.821   31.388
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  38.35290    1.36416   28.11  <2e-16 ***
Edad.media.2007 -0.76192    0.03271  -23.29  <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 4.035 on 768 degrees of freedom
Multiple R-squared:  0.4139, Adjusted R-squared:  0.4132
F-statistic: 542.5 on 1 and 768 DF,  p-value: < 2.2e-16
```

$$\bullet \beta_0 = 38.35, \beta_1 = -0.76192$$



$$\hat{y} = 38.35 - 0.76192x$$

$$\bullet S_{\hat{\beta}_0} = 1.364, S_{\hat{\beta}_1} = 0.033$$

$$\bullet \left\{ \begin{array}{l} H_0 : \beta_0 = 0 \\ H_1 : \beta_0 \neq 0 \end{array} \right\} \Rightarrow \begin{array}{l} t = 28.11 \\ p\text{-valor} = 2 * 10^{-16} \end{array}$$

$$\bullet \left\{ \begin{array}{l} H_0 : \beta_1 = 0 \\ H_1 : \beta_1 \neq 0 \end{array} \right\} \Rightarrow \begin{array}{l} t = -23.29 \\ p\text{-valor} = 2 * 10^{-16} \end{array}$$

La recta ajustada aparece especificada a través de sus coeficientes.

Ejemplo: Ajuste de la recta de regresión

Estadísticos → Ajuste de modelos → Regresión lineal

- **Variable explicada (o dependiente) - y** : Tasa líneas ADSL
- **Variable explicativa (o independiente) - x**: Edad media

Coeficientes del modelo RLS

```
Residuals:
    Min       1Q   Median       3Q      Max
-10.758  -2.538  -0.204   1.821   31.388
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  38.35290    1.36416   28.11  <2e-16 ***
Edad.media.2007 -0.76192    0.03271  -23.29  <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 4.035 on 768 degrees of freedom
Multiple R-squared:  0.4139, Adjusted R-squared:  0.4132
F-statistic: 542.5 on 1 and 768 DF,  p-value: < 2.2e-16
```

$$\bullet \beta_0 = 38.35, \beta_1 = -0.76192$$



$$\hat{y} = 38.35 - 0.76192x$$

$$\bullet S_{\hat{\beta}_0} = 1.364, S_{\hat{\beta}_1} = 0.033$$

$$\bullet \left\{ \begin{array}{l} H_0 : \beta_0 = 0 \\ H_1 : \beta_0 \neq 0 \end{array} \right\} \Rightarrow \begin{array}{l} t = 28.11 \\ p\text{-valor} = 2 * 10^{-16} \end{array}$$

$$\bullet \left\{ \begin{array}{l} H_0 : \beta_1 = 0 \\ H_1 : \beta_1 \neq 0 \end{array} \right\} \Rightarrow \begin{array}{l} t = -23.29 \\ p\text{-valor} = 2 * 10^{-16} \end{array}$$

La recta ajustada aparece especificada a través de sus coeficientes.

Ejemplo: Ajuste de la recta de regresión

Estadísticos → Ajuste de modelos → Regresión lineal

- **Variable explicada (o dependiente) - y** : Tasa líneas ADSL
- **Variable explicativa (o independiente) - x**: Edad media

Coeficientes del modelo RLS

```
Residuals:
    Min       1Q   Median       3Q      Max
-10.758  -2.538  -0.204   1.821   31.388
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  38.35290    1.36416   28.11  <2e-16 ***
Edad.media.2007 -0.76192    0.03271  -23.29  <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 4.035 on 768 degrees of freedom
Multiple R-squared:  0.4139, Adjusted R-squared:  0.4132
F-statistic: 542.5 on 1 and 768 DF,  p-value: < 2.2e-16
```

$$\bullet \beta_0 = 38.35, \beta_1 = -0.76192$$



$$\hat{y} = 38.35 - 0.76192x$$

$$\bullet S_{\hat{\beta}_0} = 1.364, S_{\hat{\beta}_1} = 0.033$$

$$\bullet \left\{ \begin{array}{l} H_0 : \beta_0 = 0 \\ H_1 : \beta_0 \neq 0 \end{array} \right\} \Rightarrow \begin{array}{l} t = 28.11 \\ p\text{-valor} = 2 * 10^{-16} \end{array}$$

$$\bullet \left\{ \begin{array}{l} H_0 : \beta_1 = 0 \\ H_1 : \beta_1 \neq 0 \end{array} \right\} \Rightarrow \begin{array}{l} t = -23.29 \\ p\text{-valor} = 2 * 10^{-16} \end{array}$$

La recta ajustada aparece especificada a través de sus coeficientes.

Ejemplo: Ajuste de la recta de regresión

Estadísticos → Ajuste de modelos → Regresión lineal

- **Variable explicada (o dependiente) - y** : Tasa líneas ADSL
- **Variable explicativa (o independiente) - x**: Edad media

Coeficientes del modelo RLS

```
Residuals:
    Min       1Q   Median       3Q      Max
-10.758  -2.538  -0.204   1.821   31.388

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  38.35290    1.36416   28.11  <2e-16 ***
Edad.media.2007 -0.76192    0.03271  -23.29  <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.035 on 768 degrees of freedom
Multiple R-squared:  0.4139, Adjusted R-squared:  0.4132
F-statistic: 542.5 on 1 and 768 DF,  p-value: < 2.2e-16
```

$$\bullet \beta_0 = 38.35, \beta_1 = -0.76192$$



$$\hat{y} = 38.35 - 0.76192x$$

$$\bullet S_{\hat{\beta}_0} = 1.364, S_{\hat{\beta}_1} = 0.033$$

$$\bullet \left\{ \begin{array}{l} H_0 : \beta_0 = 0 \\ H_1 : \beta_0 \neq 0 \end{array} \right\} \Rightarrow \begin{array}{l} t = 28.11 \\ p\text{-valor} = 2 * 10^{-16} \end{array}$$

$$\bullet \left\{ \begin{array}{l} H_0 : \beta_1 = 0 \\ H_1 : \beta_1 \neq 0 \end{array} \right\} \Rightarrow \begin{array}{l} t = -23.29 \\ p\text{-valor} = 2 * 10^{-16} \end{array}$$

La recta ajustada aparece especificada a través de sus coeficientes.

Ejemplo: Ajuste de la recta de regresión

Estadísticos → Ajuste de modelos → Regresión lineal

- **Variable explicada (o dependiente) - y** : Tasa líneas ADSL
- **Variable explicativa (o independiente) - x**: Edad media

Coeficientes del modelo RLS

```
Residuals:
    Min       1Q   Median       3Q      Max
-10.758  -2.538  -0.204   1.821   31.388
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  38.35290    1.36416   28.11  <2e-16 ***
Edad.media.2007 -0.76192    0.03271  -23.29  <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 4.035 on 768 degrees of freedom
Multiple R-squared:  0.4139, Adjusted R-squared:  0.4132
F-statistic: 542.5 on 1 and 768 DF,  p-value: < 2.2e-16
```

$$\bullet \beta_0 = 38.35, \beta_1 = -0.76192$$



$$\hat{y} = 38.35 - 0.76192x$$

$$\bullet S_{\hat{\beta}_0} = 1.364, S_{\hat{\beta}_1} = 0.033$$

$$\bullet \left\{ \begin{array}{l} H_0 : \beta_0 = 0 \\ H_1 : \beta_0 \neq 0 \end{array} \right\} \Rightarrow \begin{array}{l} t = 28.11 \\ p\text{-valor} = 2 * 10^{-16} \end{array}$$

$$\bullet \left\{ \begin{array}{l} H_0 : \beta_1 = 0 \\ H_1 : \beta_1 \neq 0 \end{array} \right\} \Rightarrow \begin{array}{l} t = -23.29 \\ p\text{-valor} = 2 * 10^{-16} \end{array}$$

La recta ajustada aparece especificada a través de sus coeficientes.

Ejemplo: Coeficiente de determinación

¿Es adecuado el modelo ajustado?

- El error estandar del ajuste es $S_{\hat{u}} = 4.035$
- Coeficiente de determinación: $R^2 = 0,4139$.

El 41,39 % de toda la variabilidad de la tasa de líneas ADSL puede ser explicado por la edad media del municipio.

Ejemplo: Coeficiente de determinación

¿Es adecuado el modelo ajustado?

```
Residuals:
    Min       1Q   Median       3Q      Max
-10.758  -2.538  -0.204   1.821   31.388

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  38.35290    1.36416   28.11  <2e-16 ***
Edad.media.2007 -0.76192    0.03271  -23.29  <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.035 on 768 degrees of freedom
Multiple R-squared:  0.4139, Adjusted R-squared:  0.4132
F-statistic: 542.5 on 1 and 768 DF, p-value: < 2.2e-16
```

- El error estandar del ajuste es $S_{\hat{u}} = 4.035$
- Coeficiente de determinación: $R^2 = 0,4139$.

El 41,39 % de toda la variabilidad de la tasa de líneas ADSL puede ser explicado por la edad media del municipio.

Ejemplo: Coeficiente de determinación

¿Es adecuado el modelo ajustado?

```
Residuals:
    Min       1Q   Median       3Q      Max
-10.758  -2.538  -0.204   1.821   31.388

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  38.35290    1.36416   28.11  <2e-16 ***
Edad.media.2007 -0.76192    0.03271  -23.29  <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.035 on 768 degrees of freedom
Multiple R-squared:  0.4139, Adjusted R-squared:  0.4132
F-statistic: 542.5 on 1 and 768 DF,  p-value: < 2.2e-16
```

- El error estandar del ajuste es $S_{\hat{u}} = 4.035$
- Coeficiente de determinación: $R^2 = 0,4139$.

El 41,39 % de toda la variabilidad de la tasa de líneas ADSL puede ser explicado por la edad media del municipio.

Ejemplo: Coeficiente de determinación

¿Es adecuado el modelo ajustado?

```
Residuals:
    Min       1Q   Median       3Q      Max
-10.758  -2.538  -0.204   1.821   31.388

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  38.35290    1.36416   28.11  <2e-16 ***
Edad.media.2007 -0.76192    0.03271  -23.29  <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.035 on 768 degrees of freedom
Multiple R-squared:  0.4139, Adjusted R-squared:  0.4132
F-statistic: 542.5 on 1 and 768 DF, p-value: < 2.2e-16
```

- El error estandar del ajuste es $S_{\hat{u}} = 4.035$
- Coeficiente de determinación: $R^2 = 0,4139$.

El 41,39 % de toda la variabilidad de la tasa de líneas ADSL puede ser explicado por la edad media del municipio.

Ejemplo: Coeficiente de determinación

¿Es adecuado el modelo ajustado?

```
Residuals:
    Min       1Q   Median       3Q      Max
-10.758  -2.538  -0.204   1.821   31.388

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  38.35290    1.36416   28.11  <2e-16 ***
Edad.media.2007 -0.76192    0.03271  -23.29  <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.035 on 768 degrees of freedom
Multiple R-squared:  0.4139, Adjusted R-squared:  0.4132
F-statistic: 542.5 on 1 and 768 DF, p-value: < 2.2e-16
```

- El error estandar del ajuste es $S_{\hat{u}} = 4.035$
- Coeficiente de determinación: $R^2 = 0,4139$.

El 41,39% de toda la variabilidad de la tasa de líneas ADSL puede ser explicado por la edad media del municipio.

Ejemplo: Intervalos de confianza para β_0 y β_1

Modelos → Intervalos de confianza

- Al 95 %, $I.C.(\beta_0) = [35.67, 41.03]$
- Al 95 %, $I.C.(\beta_1) = [-0.83, -0.70]$

Ejemplo: Intervalos de confianza para β_0 y β_1

Modelos → Intervalos de confianza

	Estimate	2,5 %	97,5 %
(Intercept)	38.3528953	35.6749790	41.0308115
Edad.media.2007	-0.7619193	-0.8261374	-0.6977012

- Al 95 %, $I.C.(\beta_0) = [35.67, 41.03]$
- Al 95 %, $I.C.(\beta_1) = [-0.83, -0.70]$

Ejemplo: Intervalos de confianza para β_0 y β_1

Modelos → Intervalos de confianza

	Estimate	2,5 %	97,5 %
(Intercept)	38.3528953	35.6749790	41.0308115
Edad.media.2007	-0.7619193	-0.8261374	-0.6977012

- Al 95 %, $I.C.(\beta_0) = [35.67, 41.03]$
- Al 95 %, $I.C.(\beta_1) = [-0.83, -0.70]$

Ejemplo: Intervalos de confianza para β_0 y β_1

Modelos → Intervalos de confianza

	Estimate	2,5 %	97,5 %
(Intercept)	38.3528953	35.6749790	41.0308115
Edad.media.2007	-0.7619193	-0.8261374	-0.6977012

- Al 95 %, $I.C.(\beta_0) = [35.67, 41.03]$
- Al 95 %, $I.C.(\beta_1) = [-0.83, -0.70]$

Contenidos

- 1 Introducción
- 2 Correlación
 - Coeficiente de correlación lineal
- 3 El modelo RLS
 - Formulación del modelo RLS
 - Estimación de los parámetros del modelo RLS
 - Inferencia estadística sobre los parámetros del modelo RLS
 - Coeficiente de determinación
- 4 **Predicción**
 - Predicción para un valor individual
 - Predicción para el valor medio
- 5 Diagnos
 - Hipótesis de normalidad
 - Hipótesis de Homocedasticidad
 - Hipótesis de Independencia

La recta de regresión ajustada se puede utilizar para determinar el valor que toma la variable y para un valor dado de la variable x

¿Cuál sería la tasa de líneas ADSL en los municipios andaluces con una edad media de 45 años?

¿Cuál sería la tasa media de líneas ADSL en los municipios andaluces con una edad media de 45 años?

La recta de regresión ajustada se puede utilizar para determinar el valor que toma la variable y para un valor dado de la variable x

¿Cuál sería la tasa de líneas ADSL en los municipios andaluces con una edad media de 45 años?

¿Cuál sería la tasa media de líneas ADSL en los municipios andaluces con una edad media de 45 años?

La recta de regresión ajustada se puede utilizar para determinar el valor que toma la variable y para un valor dado de la variable x

¿Cuál sería la tasa de líneas ADSL en los municipios andaluces con una edad media de 45 años?

¿Cuál sería la tasa media de líneas ADSL en los municipios andaluces con una edad media de 45 años?

Contenidos

- 1 Introducción
- 2 Correlación
 - Coeficiente de correlación lineal
- 3 El modelo RLS
 - Formulación del modelo RLS
 - Estimación de los parámetros del modelo RLS
 - Inferencia estadística sobre los parámetros del modelo RLS
 - Coeficiente de determinación
- 4 **Predicción**
 - **Predicción para un valor individual**
 - Predicción para el valor medio
- 5 Diagnos
 - Hipótesis de normalidad
 - Hipótesis de Homocedasticidad
 - Hipótesis de Independencia

Predicción puntual

La **predicción puntual mínimo cuadrática** \hat{y}_p de un valor individual de Y correspondiente al valor x_p es:

$$\hat{y}_p = \hat{\beta}_0 + \hat{\beta}_1 x_p$$

Intervalo de confianza

La predicción por intervalo o el **intervalo de confianza, al nivel de confianza $1 - \alpha$, para y_p** viene dado por

$$\hat{y}_p \pm t_{n-2, 1-\alpha/2} S_{\hat{u}} \sqrt{1 + \frac{1}{n} + \frac{(x_p - \bar{x})^2}{n S_x^2}}$$

donde $S_{\hat{u}}^2$ ya ha sido definida en ► varianza residual.

Predicción puntual

La **predicción puntual mínimo cuadrática** \hat{y}_p de un valor individual de Y correspondiente al valor x_p es:

$$\hat{y}_p = \hat{\beta}_0 + \hat{\beta}_1 x_p$$

Intervalo de confianza

La predicción por intervalo o el **intervalo de confianza, al nivel de confianza $1 - \alpha$, para y_p** viene dado por

$$\hat{y}_p \pm t_{n-2, 1-\alpha/2} S_{\hat{u}} \sqrt{1 + \frac{1}{n} + \frac{(x_p - \bar{x})^2}{nS_x^2}}$$

donde $S_{\hat{u}}^2$ ya ha sido definida en [varianza residual](#).

Contenidos

- 1 Introducción
- 2 Correlación
 - Coeficiente de correlación lineal
- 3 El modelo RLS
 - Formulación del modelo RLS
 - Estimación de los parámetros del modelo RLS
 - Inferencia estadística sobre los parámetros del modelo RLS
 - Coeficiente de determinación
- 4 Predicción**
 - Predicción para un valor individual
 - Predicción para el valor medio**
- 5 Diagnos
 - Hipótesis de normalidad
 - Hipótesis de Homocedasticidad
 - Hipótesis de Independencia

Predicción puntual

Se considera el **estimador del valor medio o esperado de Y**,
dado el valor x_p de X:

$$\widehat{E[y_p]} = \hat{\beta}_0 + \hat{\beta}_1 x_p$$

Intervalo de confianza

La predicción por intervalo o el **intervalo de confianza, al nivel de confianza $1 - \alpha$, para $E[y_p]$** viene dado por

$$\widehat{E[y_p]} \pm t_{n-2, 1-\alpha/2} S_{\hat{u}} \sqrt{\frac{1}{n} + \frac{(\bar{x} - x_p)^2}{n S_x^2}}$$

donde $S_{\hat{u}}^2$ ya ha sido definida en ► varianza residual.

Predicción puntual

Se considera el **estimador del valor medio o esperado de Y**,
dado el valor x_p de X:

$$\widehat{E[y_p]} = \hat{\beta}_0 + \hat{\beta}_1 x_p$$

Intervalo de confianza

La predicción por intervalo o el **intervalo de confianza, al nivel de confianza $1 - \alpha$, para $E[y_p]$** viene dado por

$$\widehat{E[y_p]} \pm t_{n-2, 1-\alpha/2} S_{\hat{u}} \sqrt{\frac{1}{n} + \frac{(\bar{x} - x_p)^2}{n S_x^2}}$$

donde $S_{\hat{u}}^2$ ya ha sido definida en ► varianza residual.

Ejemplo: Predicciones

Modelos → Prediction intervals (HH)

Permite calcular:

- Valor de la recta de Regresión (*point estimate only*)

Para $x = 45$ años, $\hat{y} = 4.165\%$

Para $x = 50$ años, $\hat{y} = 2.250236\%$

- Intervalo de predicción para el valor promedio
(*confidence interval for mean*)

Al 95%, para $x = 45$ años, $CI_{\mu} = [3.701712, 4.6134]$

- Intervalo de predicción para el valor individual
(*prediction interval for individual*)

Al 95%, para $x = 45$ años, $PI_{\mu} = [-3.363147, 11.69022]$

Ejemplo: Predicciones

Modelos → Prediction intervals (HH)

Permite calcular:

- Valor de la recta de Regresión (*point estimate only*)
Para $x = 25$ años, $\hat{y} = 4.165\%$
Para $x = 50$ años, $\hat{y} = 2.25236\%$
- Intervalo de predicción para el valor promedio
(*confidence interval for mean*)
Al 95%, para $x = 45$ años, $\hat{y}_{[.95]} \in [3.70112, 4.4134]$
- Intervalo de predicción para el valor individual
(*prediction interval for individual*)
Al 95%, para $x = 45$ años, $\hat{y}_{[.95]} \in [-3.36317, 11.0922]$

Ejemplo: Predicciones

Modelos → Prediction intervals (HH)

Permite calcular:

- Valor de la recta de Regresión (*point estimate only*)

Para $x = 45$ años , $\hat{y} = 4.066 \%$

Para $x = 50$ años , $\hat{y} = 0.2569296 \%$

- Intervalo de predicción para el valor promedio
(*confidence interval for mean*)

Al 95 %, para $x = 45$ años , $E[y_p] \in [3.701712, 4.43134]$

- Intervalo de predicción para el valor individual
(*prediction interval for individual*)

Al 95 %, para $x = 45$ años , $y_p \in [-3.863147, 11.9962]$

Ejemplo: Predicciones

Modelos → Prediction intervals (HH)

Permite calcular:

- Valor de la recta de Regresión (*point estimate only*)
Para $x = 45$ años , $\hat{y} = 4.066\%$
Para $x = 50$ años , $\hat{y} = 0.2569296\%$
- Intervalo de predicción para el valor promedio (*confidence interval for mean*)
Al 95 %, para $x = 45$ años , $E[y_p] \in [3.701712, 4.43134]$
- Intervalo de predicción para el valor individual (*prediction interval for individual*)
Al 95 %, para $x = 45$ años , $y_p \in [-3.863147, 11.9962]$

Ejemplo: Predicciones

Modelos → Prediction intervals (HH)

Permite calcular:

- Valor de la recta de Regresión (*point estimate only*)
Para $x = 45$ años , $\hat{y} = 4.066\%$
Para $x = 50$ años , $\hat{y} = 0.2569296\%$
- Intervalo de predicción para el valor promedio (*confidence interval for mean*)
Al 95 %, para $x = 45$ años , $E[y_p] \in [3.701712, 4.43134]$
- Intervalo de predicción para el valor individual (*prediction interval for individual*)
Al 95 %, para $x = 45$ años , $y_p \in [-3.863147, 11.9962]$

Ejemplo: Predicciones

Modelos → Prediction intervals (HH)

Permite calcular:

- Valor de la recta de Regresión (*point estimate only*)
Para $x = 45$ años , $\hat{y} = 4.066 \%$
Para $x = 50$ años , $\hat{y} = 0.2569296 \%$
- Intervalo de predicción para el valor promedio (*confidence interval for mean*)
Al 95 %, para $x = 45$ años , $E[y_p] \in [3.701712, 4.43134]$
- Intervalo de predicción para el valor individual (*prediction interval for individual*)
Al 95 %, para $x = 45$ años , $y_p \in [-3.863147, 11.9962]$

Ejemplo: Predicciones

Modelos → Prediction intervals (HH)

Permite calcular:

- Valor de la recta de Regresión (*point estimate only*)
Para $x = 45$ años , $\hat{y} = 4.066 \%$
Para $x = 50$ años , $\hat{y} = 0.2569296 \%$
- Intervalo de predicción para el valor promedio
(*confidence interval for mean*)
Al 95 %, **para** $x = 45$ años , $E[y_p] \in [3.701712, 4.43134]$
- Intervalo de predicción para el valor individual
(*prediction interval for individual*)
Al 95 %, **para** $x = 45$ años , $y_p \in [-3.863147, 11.9962]$

Ejemplo: Predicciones

Modelos → Prediction intervals (HH)

Permite calcular:

- Valor de la recta de Regresión (*point estimate only*)
Para $x = 45$ años , $\hat{y} = 4.066 \%$
Para $x = 50$ años , $\hat{y} = 0.2569296 \%$
- Intervalo de predicción para el valor promedio (*confidence interval for mean*)
Al 95 %, **para** $x = 45$ años , $E[y_p] \in [3.701712, 4.43134]$
- Intervalo de predicción para el valor individual (*prediction interval for individual*)
Al 95 %, **para** $x = 45$ años , $y_p \in [-3.863147, 11.9962]$

Ejemplo: Predicciones

Modelos → Prediction intervals (HH)

Permite calcular:

- Valor de la recta de Regresión (*point estimate only*)
Para $x = 45$ años , $\hat{y} = 4.066 \%$
Para $x = 50$ años , $\hat{y} = 0.2569296 \%$
- Intervalo de predicción para el valor promedio
(*confidence interval for mean*)
Al 95 %, **para** $x = 45$ años , $E[y_p] \in [3.701712, 4.43134]$
- Intervalo de predicción para el valor individual
(*prediction interval for individual*)
Al 95 %, **para** $x = 45$ años , $y_p \in [-3.863147, 11.9962]$

Modelos → Confidence intervals plot

Contenidos

- 1 Introducción
- 2 Correlación
 - Coeficiente de correlación lineal
- 3 El modelo RLS
 - Formulación del modelo RLS
 - Estimación de los parámetros del modelo RLS
 - Inferencia estadística sobre los parámetros del modelo RLS
 - Coeficiente de determinación
- 4 Predicción
 - Predicción para un valor individual
 - Predicción para el valor medio
- 5 Diagnos
 - Hipótesis de normalidad
 - Hipótesis de Homocedasticidad
 - Hipótesis de Independencia

Contenidos

- 1 Introducción
- 2 Correlación
 - Coeficiente de correlación lineal
- 3 El modelo RLS
 - Formulación del modelo RLS
 - Estimación de los parámetros del modelo RLS
 - Inferencia estadística sobre los parámetros del modelo RLS
 - Coeficiente de determinación
- 4 Predicción
 - Predicción para un valor individual
 - Predicción para el valor medio
- 5 Diagnos
 - Hipótesis de normalidad
 - Hipótesis de Homocedasticidad
 - Hipótesis de Independencia

Normalidad

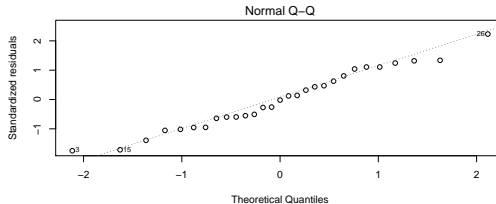
- La Hipótesis de normalidad de los errores aleatorios es necesaria para realizar inferencias respecto a los parámetros y para la construcción de los intervalos de predicción.
- Dicha Hipótesis puede verificarse mediante un gráfico de normalidad de los residuos^a.

Si fueran completamente normales, todos los puntos estarían sobre la recta. En la medida en que se alejen de ella, se alejan de la normalidad.

^atambién se puede aplicar un test de Bondad de Ajuste.

Normalidad

- La Hipótesis de normalidad de los errores aleatorios es necesaria para realizar inferencias respecto a los parámetros y para la construcción de los intervalos de predicción.
- Dicha Hipótesis puede verificarse mediante un gráfico de normalidad de los residuos^a.

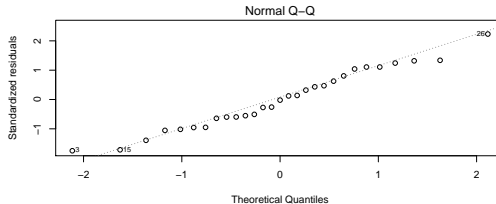


Si fueran completamente normales, todos los puntos estarían sobre la recta. En la medida en que se alejen de ella, se alejan de la normalidad.

^atambién se puede aplicar un test de Bondad de Ajuste.

Normalidad

- La Hipótesis de normalidad de los errores aleatorios es necesaria para realizar inferencias respecto a los parámetros y para la construcción de los intervalos de predicción.
- Dicha Hipótesis puede verificarse mediante un **gráfico de normalidad de los residuos^a**.



Si fueran completamente normales, todos los puntos estarían sobre la recta. En la medida en que se alejen de ella, se alejan de la normalidad.

^atambién se puede aplicar un test de Bondad de Ajuste.

Contenidos

- 1 Introducción
- 2 Correlación
 - Coeficiente de correlación lineal
- 3 El modelo RLS
 - Formulación del modelo RLS
 - Estimación de los parámetros del modelo RLS
 - Inferencia estadística sobre los parámetros del modelo RLS
 - Coeficiente de determinación
- 4 Predicción
 - Predicción para un valor individual
 - Predicción para el valor medio
- 5 Diagnos
 - Hipótesis de normalidad
 - **Hipótesis de Homocedasticidad**
 - Hipótesis de Independencia

Homocedasticidad

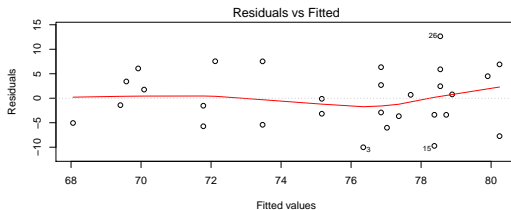
- La varianza de los residuos han de ser constantes.
 - Dicha Hipótesis puede verificarse mediante el gráfico de residuos frente a valores ajustados
-
- En el caso de homocedasticidad, la dispersión vertical de los puntos de la gráfica no debe variar según varíe el eje X. En caso contrario, se habla de heterocedasticidad.

Homocedasticidad

- La varianza de los residuos han de ser constantes.
- Dicha Hipótesis puede verificarse mediante el gráfico de residuos frente a valores ajustados
- En el caso de homocedasticidad, la dispersión vertical de los puntos de la gráfica no debe variar según varíe el eje X. En caso contrario, se habla de heterocedasticidad.

Homocedasticidad

- La varianza de los residuos han de ser constantes.
- Dicha Hipótesis puede verificarse mediante el **gráfico de residuos frente a valores ajustados**



- En el caso de homocedasticidad, la dispersión vertical de los puntos de la gráfica no debe variar según varíe el eje X. En caso contrario, se habla de heterocedasticidad.

Contenidos

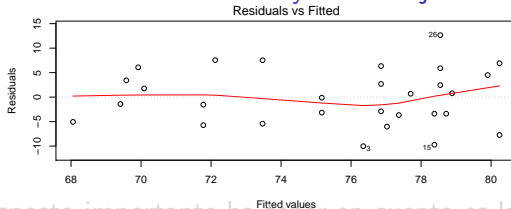
- 1 Introducción
- 2 Correlación
 - Coeficiente de correlación lineal
- 3 El modelo RLS
 - Formulación del modelo RLS
 - Estimación de los parámetros del modelo RLS
 - Inferencia estadística sobre los parámetros del modelo RLS
 - Coeficiente de determinación
- 4 Predicción
 - Predicción para un valor individual
 - Predicción para el valor medio
- 5 Diagnos
 - Hipótesis de normalidad
 - Hipótesis de Homocedasticidad
 - **Hipótesis de Independencia**

Independencia

- Los errores han de ser independientes unos de otros, es decir, la magnitud de un error no influye en absoluto en la magnitud de otros errores.
- Si los errores son independientes, no debe observarse ningún patrón en el **gráfico de residuos frente a valores ajustados**, es decir, ningún efecto en el mismo que haga pensar en algún tipo de relación entre residuos y valores ajustados.
- Otro aspecto importante ha tener en cuenta es la identificación de **observaciones atípicas**.

Independencia

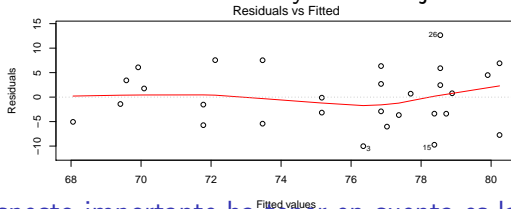
- Los errores han de ser independientes unos de otros, es decir, la magnitud de un error no influye en absoluto en la magnitud de otros errores.
- Si los errores son independientes, no debe observarse ningún patrón en el **gráfico de residuos frente a valores ajustados**, es decir, ningún efecto en el mismo que haga pensar en algún tipo de relación entre residuos y valores ajustados.



- Otro aspecto importante ha tener en cuenta es la identificación de **observaciones atípicas**.

Independencia

- Los errores han de ser independientes unos de otros, es decir, la magnitud de un error no influye en absoluto en la magnitud de otros errores.
- Si los errores son independientes, no debe observarse ningún patrón en el **gráfico de residuos frente a valores ajustados**, es decir, ningún efecto en el mismo que haga pensar en algún tipo de relación entre residuos y valores ajustados.



- Otro aspecto importante ha tener en cuenta es la identificación de **observaciones atípicas**.

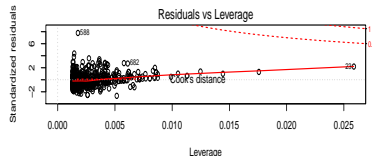
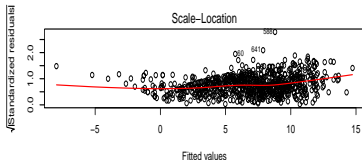
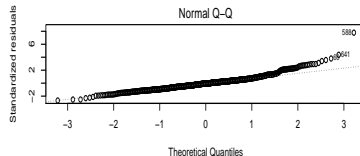
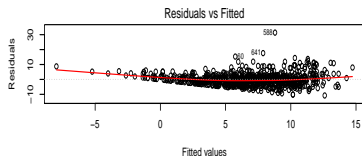
Ejemplo: Diagnos

Modelos → Gráficas → Gráficas básicas de diagnóstico

Ejemplo: Diagnóstico

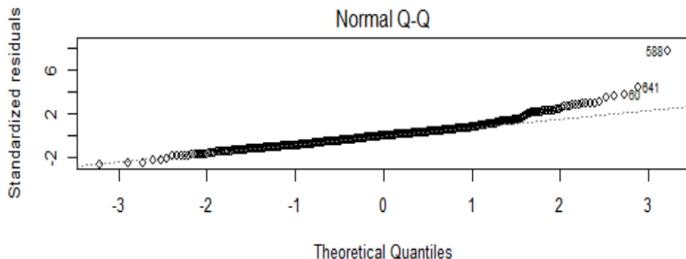
Modelos → Gráficas → Gráficas básicas de diagnóstico

lm(tasa.lineas.ADSL.2007 ~ Edad.media.2007)



Ejemplo: Diagnóstico

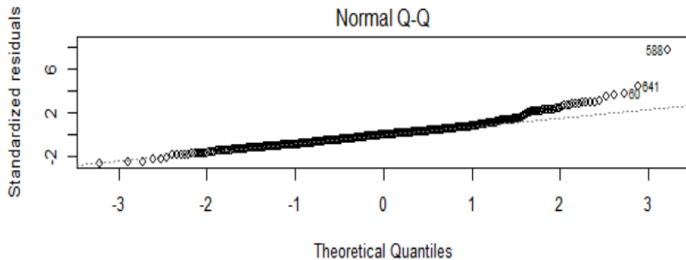
Modelos → Gráficas → Gráficas básicas de diagnóstico



- No se observa gran desviación de los puntos respecto de la recta, salvo en el tramo final \Rightarrow Hip. normalidad

Ejemplo: Diagnóstico

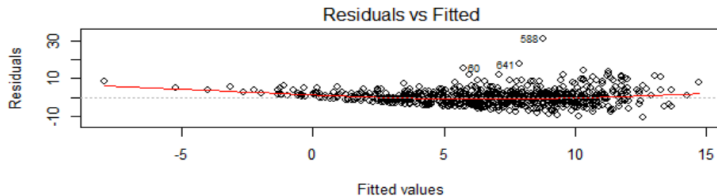
Modelos → Gráficas → Gráficas básicas de diagnóstico



- No se observa gran desviación de los puntos respecto de la recta, salvo en el tramo final \Rightarrow **Hip. normalidad**

Ejemplo: Diagnóstico

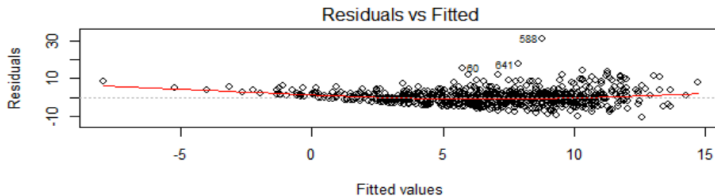
Modelos → Gráficas → Gráficas básicas de diagnóstico



- Como la media de los residuos es cero, la curva de medias (línea roja) ha de coincidir con el eje X.
- No se observa ningún patrón extraño no lineal ⇒ **Hip. Independencia**
- Se observa mayor dispersión en la parte dcha ⇒ Varianza de los residuos no constante ⇒ **Heterocedasticidad**
- Se señalan 3 datos como valores atípicos

Ejemplo: Diagnóstico

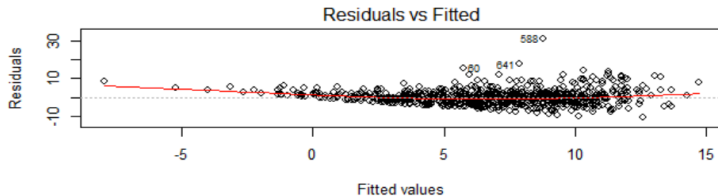
Modelos → Gráficas → Gráficas básicas de diagnóstico



- Como la media de los residuos es cero, la curva de medias (línea roja) ha de coincidir con el eje X.
- No se observa ningún patrón extraño no lineal ⇒ **Hip. Independencia**
- Se observa mayor dispersión en la parte dcha ⇒ Varianza de los residuos no constante ⇒ **Heterocedasticidad**
- Se señalan 3 datos como valores atípicos

Ejemplo: Diagnóstico

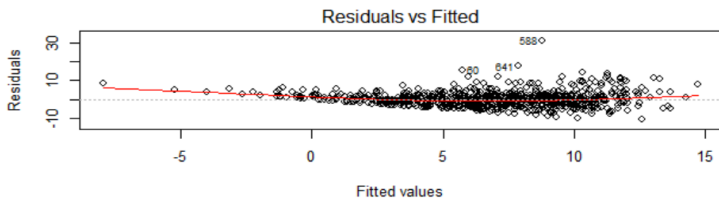
Modelos → Gráficas → Gráficas básicas de diagnóstico



- Como la media de los residuos es cero, la curva de medias (línea roja) ha de coincidir con el eje X.
- No se observa ningún patrón extraño no lineal ⇒ **Hip. Independencia**
- Se observa mayor dispersión en la parte dcha ⇒ Varianza de los residuos no constante ⇒ **Heterocedasticidad**
- Se señalan 3 datos como valores atípicos

Ejemplo: Diagnóstico

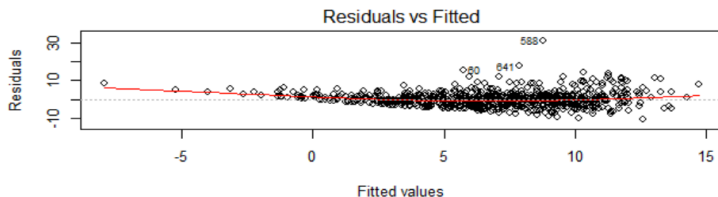
Modelos → Gráficas → Gráficas básicas de diagnóstico



- Como la media de los residuos es cero, la curva de medias (línea roja) ha de coincidir con el eje X.
- No se observa ningún patrón extraño no lineal ⇒ **Hip. Independencia**
- Se observa mayor dispersión en la parte dcha ⇒ Varianza de los residuos no constante ⇒ **Heterocedasticidad**
- Se señalan 3 datos como valores atípicos

Ejemplo: Diagnóstico

Modelos → Gráficas → Gráficas básicas de diagnóstico



- Como la media de los residuos es cero, la curva de medias (línea roja) ha de coincidir con el eje X.
- No se observa ningún patrón extraño no lineal ⇒ **Hip. Independencia**
- Se observa mayor dispersión en la parte dcha ⇒ Varianza de los residuos no constante ⇒ **Heterocedasticidad**
- Se señalan 3 datos como valores atípicos