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

Métodos Estadísticos I

Regresión con R

Prof. Douglas Rivas

7 de julio de 2010

lm()

¿Que hace?

Determina las estimaciones de los parámetros de un modelo de regresión lineal.

Sintaxis

La sintaxis más simple de dicha función es:
lm(formula,data)

lm()

¿Que hace?

Determina las estimaciones de los parámetros de un modelo de regresión lineal.

Sintaxis

La sintaxis más simple de dicha función es:
lm(formula,data)

lm()

¿Que hace?

Determina las estimaciones de los parámetros de un modelo de regresión lineal.

Sintaxis

La sintaxis más simple de dicha función es:
lm(formula,data)

lm()

- *formula* es un objeto que representa el modelo planteado y se representa en la forma *respuesta regresoras* donde respuesta es la variable dependiente o respuesta y regresoras es el conjunto de variables independientes que en el caso de regresión múltiple van separadas por el signo +. Por ejemplo si y es la variable respuesta y x_1 y x_2 son las variables independientes entonces la formula es $y \sim x_1 + x_2$.
- *data* es el conjunto de datos que se están estudiando.

lm()

- *formula* es un objeto que representa el modelo planteado y se representa en la forma *respuesta regresoras* donde respuesta es la variable dependiente o respuesta y regresoras es el conjunto de variables independientes que en el caso de regresión múltiple van separadas por el signo +. Por ejemplo si y es la variable respuesta y x_1 y x_2 son las variables independientes entonces la formula es $y \sim x_1 + x_2$.
- *data* es el conjunto de datos que se están estudiando.

lm()

Ejemplo

Este es un ejemplo tomado de Montgomery(2002):Un embotellador de bebidas gaseosas analiza las rutas de servicio de las máquinas expendedoras en su sistema de distribución. Le interesa predecir el tiempo necesario para que el representante de ruta atienda las máquinas expendedoras en una tienda. Esta actividad de servicio consiste en abastecer la máquina con productos embotellados, y algo de mantenimiento o limpieza. El ingeniero industrial responsable del estudio ha sugerido que las dos variables más importantes que afectan el tiempo de entrega y son la cantidad de cajas de producto abastecido, x_1, y la distancia caminada por el representante, x_2. El ingeniero ha reunido 25 observaciones de tiempo de entrega que se ven en la tabla más adelante.

lm()

Ejemplo

Este es un ejemplo tomado de Montgomery(2002):Un embotellador de bebidas gaseosas analiza las rutas de servicio de las máquinas expendedoras en su sistema de distribución. Le interesa predecir el tiempo necesario para que el representante de ruta atienda las máquinas expendedoras en una tienda. Esta actividad de servicio consiste en abastecer la máquina con productos embotellados, y algo de mantenimiento o limpieza. El ingeniero industrial responsable del estudio ha sugerido que las dos variables más importantes que afectan el tiempo de entrega y son la cantidad de cajas de producto abastecido, x_1, y la distancia caminada por el representante, x_2. El ingeniero ha reunido 25 observaciones de tiempo de entrega que se ven en la tabla más adelante.

lm()

Modelo del ejemplo

Se ajustará el modelo de regresión lineal simple siguiente

$$y = \beta_0 + \beta x_1 + \varepsilon$$

lm()

Modelo del ejemplo

Se ajustará el modelo de regresión lineal simple siguiente

$$y = \beta_0 + \beta x_1 + \varepsilon$$

lm()

Datos del ejemplo

Cuadro: Datos de tiempo de entrega

Observación	y	x_1	x_2
1	16,68	7	560
2	11,50	3	220
3	12,03	3	340
4	14,88	4	80
⋮	⋮	⋮	⋮
24	19,83	8	635
25	10,75	4	150

lm()

Datos del ejemplo

Cuadro: Datos de tiempo de entrega

Observación	y	x_1	x_2
1	16,68	7	560
2	11,50	3	220
3	12,03	3	340
4	14,88	4	80
⋮	⋮	⋮	⋮
24	19,83	8	635
25	10,75	4	150

lm()

Cargar la base de datos

```
> Datos <-read.table("tiempodeentrega.txt",header = TRUE)
```

Cargar la base de datos

```
> attach(Datos)
```

lm()

Cargar la base de datos

$> Datos < -read.table("tiempodeentrega.txt", header = TRUE)$

Cargar la base de datos

$> attach(Datos)$

lm()

Cargar la base de datos

$> Datos < -read.table(''tiempodeentrega.txt'', header = TRUE)$

Cargar la base de datos

$> attach(Datos)$

lm()

Estimación de los Parámetros:

$> MRL1 <- lm(y\ x1, data = Datos)$

Resultados obtenidos

	Estimate	Std. Error	t value	Pr(> |t|)
(Intercept)	3.3208	1.3711	2.42	0.0237
x1	2.1762	0.1240	17.55	0.0000

Cuadro: Estimación y Significancia de los Parámetros

lm()

Estimación de los Parámetros:

$> MRL1 < -lm(y\ x1, data = Datos)$

Resultados obtenidos

	Estimate	Std. Error	t value	Pr(> |t|)
(Intercept)	3.3208	1.3711	2.42	0.0237
x1	2.1762	0.1240	17.55	0.0000

Cuadro: Estimación y Significancia de los Parámetros

lm()

Estimación de los Parámetros:

$> MRL1 < -lm(y\ x1, data = Datos)$

Resultados obtenidos

	Estimate	Std. Error	t value	Pr(> |t|)
(Intercept)	3.3208	1.3711	2.42	0.0237
x1	2.1762	0.1240	17.55	0.0000

Cuadro: Estimación y Significancia de los Parámetros

summary()

¿Qué hace?

Esta función despliega información más abundante sobre el análisis de regresión que la impresa directamente por el objeto de la función **lm()**.

Información que aporta

- Algunos Estadísticos descriptivos sobre los residuos,
- las estimaciones de los parámetros y de la desviación estándar, los estadísticos para medir la significancia de cada parámetro.
- Otros estadísticos que permiten evaluar la bondad del ajuste del modelo.

summary()

¿Qué hace?

Esta función despliega información más abundante sobre el análisis de regresión que la impresa directamente por el objeto de la función **lm()**.

Información que aporta

- Algunos Estadísticos descriptivos sobre los residuos,
- las estimaciones de los parámetros y de la desviación estándar, los estadísticos para medir la significancia de cada parámetro.
- Otros estadísticos que permiten evaluar la bondad del ajuste del modelo.

summary()

¿Qué hace?

Esta función despliega información más abundante sobre el análisis de regresión que la impresa directamente por el objeto de la función **lm()**.

Información que aporta

- Algunos Estadísticos descriptivos sobre los residuos,
- las estimaciones de los parámetros y de la desviación estándar, los estadísticos para medir la significancia de cada parámetro.
- Otros estadísticos que permiten evaluar la bondad del ajuste del modelo.

summary()

Continuando con el ejemplo

$> summary(MRL1)$

Resultados obtenidos

Call: lm(formula = y x1, data = Datos)

Residuals:

Min	1Q	Median	3Q	Max
-7.5811	-1.8739	-0.3493	2.1807	10.6342

Coefficients:

	Estimate	Std. Error	t value	Pr(> |t|)
(Intercept)	3.3208	1.3711	2.42	0.0237
x1	2.1762	0.1240	17.55	0.0000

—

Residual standard error: 4.181 on 23 degrees of freedom

Multiple R-squared: 0.9305, Adjusted R-squared: 0.9275

F-statistic: 307.8 on 1 and 23 DF, p-value: 8.22e-15

summary()

Continuando con el ejemplo

$> summary(MRL1)$

Resultados obtenidos

Call: lm(formula = y x1, data = Datos)

Residuals:

Min	1Q	Median	3Q	Max
-7.5811	-1.8739	-0.3493	2.1807	10.6342

Coefficients:

	Estimate	Std. Error	t value	Pr(> |t|)
(Intercept)	3.3208	1.3711	2.42	0.0237
x1	2.1762	0.1240	17.55	0.0000

—

Residual standard error: 4.181 on 23 degrees of freedom

Multiple R-squared: 0.9305, Adjusted R-squared: 0.9275

F-statistic: 307.8 on 1 and 23 DF, p-value: 8.22e-15

summary()

Continuando con el ejemplo

$> summary(MRL1)$

Resultados obtenidos

Call: lm(formula = y x1, data = Datos)

Residuals:

Min	1Q	Median	3Q	Max
-7.5811	-1.8739	-0.3493	2.1807	10.6342

Coefficients:

	Estimate	Std. Error	t value	Pr(> |t|)
(Intercept)	3.3208	1.3711	2.42	0.0237
x1	2.1762	0.1240	17.55	0.0000

—

Residual standard error: 4.181 on 23 degrees of freedom

Multiple R-squared: 0.9305, Adjusted R-squared: 0.9275

F-statistic: 307.8 on 1 and 23 DF, p-value: 8.22e-15

anova()

¿Qué hace?

Esta función proporciona el análisis de varianza que se usa para evaluar la significancia del modelo de regresión lineal.

El argumento de la función **anova()** es un objeto de **lm()**.

anova()

¿Qué hace?

Esta función proporciona el análisis de varianza que se usa para evaluar la significancia del modelo de regresión lineal.

El argumento de la función **anova()** es un objeto de **lm()**.

anova()

¿Qué hace?

Esta función proporciona el análisis de varianza que se usa para evaluar la significancia del modelo de regresión lineal.

El argumento de la función **anova()** es un objeto de **lm()**.

anova()

Continuando con el ejemplo

$> anova(MRL1)$

Resultados obtenidos

Cuadro: Análisis de Varianza

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
x1	1	5382.41	5382.41	307.85	0.0000
Residuals	23	402.13	17.48		

anova()

Continuando con el ejemplo

$> anova(MRL1)$

Resultados obtenidos

Cuadro: Análisis de Varianza

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
x1	1	5382.41	5382.41	307.85	0.0000
Residuals	23	402.13	17.48		

anova()

Continuando con el ejemplo

$> anova(MRL1)$

Resultados obtenidos

Cuadro: Análisis de Varianza

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
x1	1	5382.41	5382.41	307.85	0.0000
Residuals	23	402.13	17.48		

fitted()

¿Qué hace?

Devuelve los valores ajustados por el modelo de regresión lineal.

El argumento de la función **fitted()** es un objeto de **lm()**.

fitted()

¿Qué hace?

Devuelve los valores ajustados por el modelo de regresión lineal.

El argumento de la función **fitted()** es un objeto de **lm()**.

fitted()

¿Qué hace?

Devuelve los valores ajustados por el modelo de regresión lineal.

El argumento de la función **fitted()** es un objeto de **lm()**.

residuals()

¿Qué hace?

Devuelve los residuales del modelo de regresión lineal.

El argumento de la función **residuals()** es un objeto de **lm()**.

residuals()

¿Qué hace?

Devuelve los residuales del modelo de regresión lineal.

El argumento de la función **residuals()** es un objeto de **lm()**.

residuals()

¿Qué hace?

Devuelve los residuales del modelo de regresión lineal.

El argumento de la función **residuals()** es un objeto de **lm()**.

rstudent(), rstandard()

¿Qué hace?

Devuelven los residuales estudentizados y estandarizados respectivamente del modelo de regresión lineal.

El argumento de ambas funciones es un objeto de **lm()**.

rstudent(), rstandard()

¿Qué hace?

Devuelven los residuales estudentizados y estandarizados respectivamente del modelo de regresión lineal.

El argumento de ambas funciones es un objeto de **lm()**.

rstudent(), rstandard()

¿Qué hace?

Devuelven los residuales estudentizados y estandarizados respectivamente del modelo de regresión lineal.

El argumento de ambas funciones es un objeto de **lm()**.

Residuos

Continuando con el ejemplo

```
> fitted(MRL1)
> residuals(MRL1)
> rstudent(MRL1)
```

Resultados obtenidos

Observación	y_i	$\hat{y}_i$	e_i	r_i
1	16.68	18.5539	-1.8739	-0.4500
2	11.50	9.8493	1.6507	0.4017
3	12.03	9.8493	2.1807	0.5321
⋮	⋮	⋮	⋮	
24	19.83	20.7301	-0.9001	-0.2152
25	10.75	12.0254	-1.2754	-0.3084

Residuos

Continuando con el ejemplo

$> fitted(MRL1)$
$> residuals(MRL1)$
$> rstudent(MRL1)$

Resultados obtenidos

Observación	y_i	$\hat{y}_i$	e_i	r_i
1	16.68	18.5539	-1.8739	-0.4500
2	11.50	9.8493	1.6507	0.4017
3	12.03	9.8493	2.1807	0.5321
⋮	⋮	⋮	⋮	
24	19.83	20.7301	-0.9001	-0.2152
25	10.75	12.0254	-1.2754	-0.3084

Residuos

Continuando con el ejemplo

$> fitted(MRL1)$
$> residuals(MRL1)$
$> rstudent(MRL1)$

Resultados obtenidos

Observación	y_i	$\hat{y}_i$	e_i	r_i
1	16.68	18.5539	-1.8739	-0.4500
2	11.50	9.8493	1.6507	0.4017
3	12.03	9.8493	2.1807	0.5321
⋮	⋮	⋮	⋮	
24	19.83	20.7301	-0.9001	-0.2152
25	10.75	12.0254	-1.2754	-0.3084

Gráficos de los residuos

plot()

Se usa para obtener los gráficos de los residuos excepto el histograma y el gráfico Q-Q.

Sintaxis

plot(variablex,variabley,xlab=”Nombre del eje x”,ylab=”Nombre del eje y”,main=”Titulo del gráfico”)

hist()

se usa obtener el histograma de los residuales

Sintaxis

hist(variable,xlab=”Nombre del eje x”,ylab=”Nombre del eje y”,main=”Titulo del gráfico”)

Gráficos de los residuos

plot()

Se usa para obtener los gráficos de los residuos excepto el histograma y el gráfico Q-Q.

Sintaxis

plot(variablex,variabley,xlab="Nombre del eje x",ylab="Nombre del eje y",main="Titulo del gráfico")

hist()

se usa obtener el histograma de los residuales

Sintaxis

hist(variable,xlab="Nombre del eje x",ylab="Nombre del eje y",main="Titulo del gráfico")

Gráficos de los residuos

plot()

Se usa para obtener los gráficos de los residuos excepto el histograma y el gráfico Q-Q.

Sintaxis

plot(variablex,variabley,xlab=”Nombre del eje x”,ylab=”Nombre del eje y”,main=”Titulo del gráfico”)

hist()

se usa obtener el histograma de los residuales

Sintaxis

hist(variable,xlab=”Nombre del eje x”,ylab=”Nombre del eje y”,main=”Titulo del gráfico”)

Gráficos de los residuos

plot()

Se usa para obtener los gráficos de los residuos excepto el histograma y el gráfico Q-Q.

Sintaxis

plot(variablex,variabley,xlab=”Nombre del eje x”,ylab=”Nombre del eje y”,main=”Titulo del gráfico”)

hist()

se usa obtener el histograma de los residuales

Sintaxis

hist(variable,xlab=”Nombre del eje x”,ylab=”Nombre del eje y”,main=”Titulo del gráfico”)

Gráficos de los residuos

plot()

Se usa para obtener los gráficos de los residuos excepto el histograma y el gráfico Q-Q.

Sintaxis

plot(variablex,variabley,xlab=”Nombre del eje x”,ylab=”Nombre del eje y”,main=”Titulo del gráfico”)

hist()

se usa obtener el histograma de los residuales

Sintaxis

hist(variable,xlab=”Nombre del eje x”,ylab=”Nombre del eje y”,main=”Titulo del gráfico”)

Gráficos de los residuos

qqnorm()

Se usa para obtener el gráfico de probabilidad normal Q-Q.

Sintaxis

qqnorm(variable,xlab=”Nombre del eje x”,ylab=”Nombre del eje y”,main=”Titulo del gráfico”)

Gráficos de los residuos

qqnorm()

Se usa para obtener el gráfico de probabilidad normal Q-Q.

Sintaxis

qqnorm(variable,xlab="Nombre del eje x",ylab="Nombre del eje y",main="Titulo del gráfico")

Gráficos de los residuos

qqnorm()

Se usa para obtener el gráfico de probabilidad normal Q-Q.

Sintaxis

qqnorm(variable,xlab=”Nombre del eje x”,ylab=”Nombre del eje y”,main=”Titulo del gráfico”)

Gráficos de los Residuos

Continuando con el ejemplo

```
> hist(residuals(MRL1), main = "", xlab =
"Residuales", ylab = "Frecuencia")
> qqnorm(rstudent(MRL1), main = "", pch = 19, xlab =
"CuantilesTeoricos", ylab = "CuantilesMuestrales")
> plot(fitted(MRL1), residuals(MRL1), xlab =
expression(hat(y)[i]), ylab = expression(e[i]))
> plot(x1, rstudent(MRL1), xlab = "Cajas", ylab =
expression(r[i]))
> resi < −rstudent(MRL1)
> plot(resi[−25], resi[−1], xlab = expression(e[i]), ylab =
expression(e[i + 1]))
> resi < −residuals(MRL1)
> plot(resi, xlab = "Observacion", ylab = expression(e[i]))
```

Gráficos de los Residuos

Continuando con el ejemplo

$> hist(residuals(MRL1), main = "", xlab = "Residuales", ylab = "Frecuencia")$

$> qqnorm(rstudent(MRL1), main = "", pch = 19, xlab = "CuantilesTeoricos", ylab = "CuantilesMuestrales")$

$> plot(fitted(MRL1), residuals(MRL1), xlab = expression(hat(y)[i]), ylab = expression(e[i]))$

$> plot(x1, rstudent(MRL1), xlab = "Cajas", ylab = expression(r[i]))$

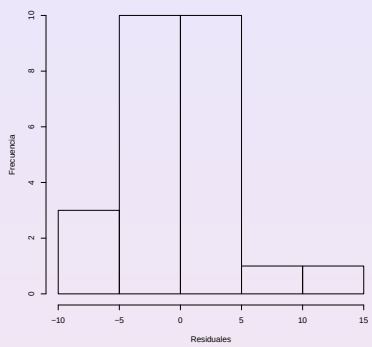
$> resi < -rstudent(MRL1)$

$> plot(resi[-25], resi[-1], xlab = expression(e[i]), ylab = expression(e[i+1]))$

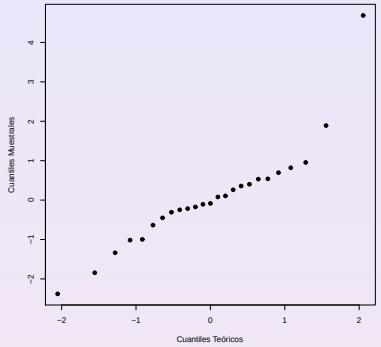
$> resi < -residuals(MRL1)$

$> plot(resi, xlab = "Observacion", ylab = expression(e[i]))$

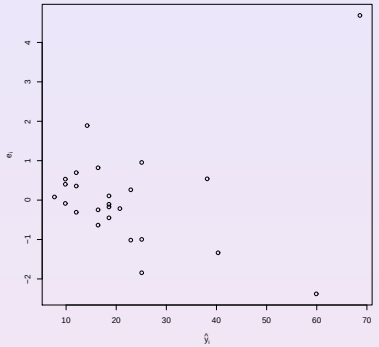
Gráficos de los Residuos



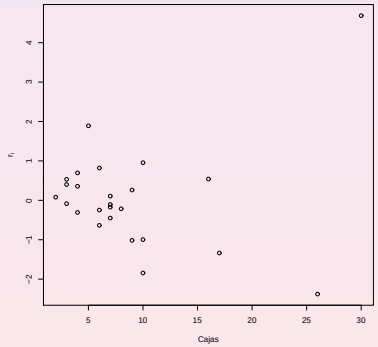
(a) Histograma de residuales



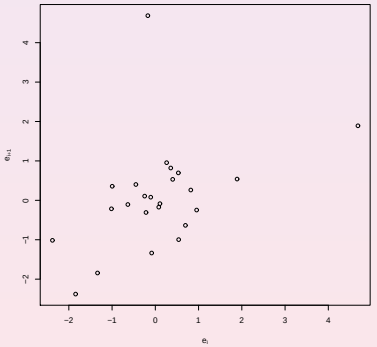
(b) Gráfico Q-Q



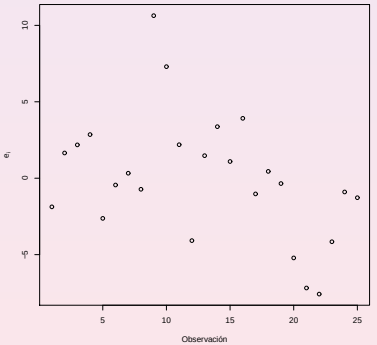
(c) Residuales vs Ajustados



(d) Residuales vs Variable independiente



(e) e_{i+1} vs e_i



(f) Residuales vs tiempo