



# ANÁLISIS DE REGRESIÓN LINEAL

---

**Con R**

**Curso de nivelación**

**Maestría en Estadística Aplicada**

**Facultad de Ciencias Económicas y Estadística**

**Junio 2019**

**Lic. Noelia Castellana**

**Lic. Mara Catalano**



# Índice

## **Introducción a R.STUDIO**

---

- Breve descripción general del programa

## **Regresión simple (ejercicio 1)**

---

- Creación del data set
- Medidas descriptivas
- Gráfico de dispersión
- Ajuste del modelo de regresión simple



# INTRODUCCIÓN-R.STUDIO

## ¿Qué es R.STUDIO?

- Es un entorno de desarrollo integrado (IDE) para el lenguaje de programación R
- Tiene una versión libre (licencia AGPL v3) y una versión comercial
- Se puede ejecutar sobre distintas plataformas (Windows, Mac, or Linux)
- También se puede ejecutar desde la web usando RStudio Server.



# INTRODUCCIÓN-R.STUDIO

## ¿Qué es R.STUDIO?

---

- Instalación:

1. Instalar R
2. Instalar R-Studio. Descargar desde el sitio oficial

<https://www.rstudio.com>

*última versión disponible:* RStudio Desktop 1.2.1335

# INTRODUCCIÓN-R.STUDIO

The screenshot displays the RStudio application window. The top menu bar includes File, Edit, Code, View, Plots, Session, Build, Debug, Tools, and Help. The toolbar below the menu contains icons for file operations and running code. The main editor area on the left shows a script file named 'Untitled1.R' with two lines of code: '1' and '2'. A blue box highlights the file explorer at the top left, showing 'MLM con RStudio.R' and 'Untitled1.R'. A text box with a blue border is overlaid on the editor area, containing the following text:

- Muestra los ficheros abiertos
- Se escriben las sentencias

The right-hand side of the interface features the 'Environment' pane, which lists the 'Global Environment' and 'Data' objects. The 'Data' pane shows a list of objects: 'aircraft' (709 obs. of 8 variables), 'buffalo' (62 obs. of 1 variable), 'cora' (num [1:11, 1:11] 1 0.825 0.764 0.659 0.635 ...), 'corb' (num [1:11, 1:11] 1 0.861 0.834 0.681 0.67 ...), 'cova' (num [1:11, 1:11] 106 106 101 92 102 ...), 'covb' (num [1:11, 1:11] 105.3 92 103.8 98.4 101.4 ...), and 'datos' (32 obs. of 7 variables). Below the 'Environment' pane is the 'Files' pane, which shows the 'User Library' of installed packages. The 'User Library' table lists the following packages:

Name	Description	Version
abind	Combine multi-dimensional arrays	1.4-0
aplpack	Another Plot PACKage: stem.leaf, bagplot, faces, spin3R, plotsummary, plothulls, and some slider functions	1.2.9
bdsmatrix	Routines for Block Diagonal Symmetric matrices	1.3-2
betareg	Beta Regression	3.0-5
bitops	Bitwise Operations	1.0-6
car	Companion to Applied Regression	2.0-20
CHAI	CHI-squared Automated Interaction Detection	0.1-2
coin	Conditional Inference Procedures in a Permutation Test Framework	1.0-23
colorspace	Color Space Manipulation	1.2-4
DAAG	Data Analysis and Graphics Data and Functions	1.22
dichromat	Color Schemes for Dichromats	2.0-0
digest	Create cryptographic hash digests of R objects	0.6.4
e1071	Misc Functions of the Department of Statistics (e1071), TU W/ien	1.6-3

The bottom pane is the 'Console', which displays the following text:

```
R is free software and comes with ABSOLUTELY NO WARRANTY.  
You are welcome to redistribute it under certain conditions.  
Type 'license()' or 'licence()' for distribution details.  
  
R is a collaborative project with many contributors.  
Type 'contributors()' for more information and  
'citation()' on how to cite R or R packages in publications.  
  
Type 'demo()' for some demos, 'help()' for on-line help, or  
'help.start()' for an HTML browser interface to help.  
Type 'q()' to quit R.  
  
[workspace loaded from ~/.RData]  
> |
```

# INTRODUCCIÓN-R.STUDIO

The screenshot displays the RStudio environment. The top menu bar includes File, Edit, Code, View, Plots, Session, Build, Debug, Tools, and Help. The toolbar contains icons for file operations, running code, and source control. The main editor window shows a script with two lines of code. The Environment pane on the right lists global objects, including 'aircraft', 'buffalo', 'cora', 'corb', 'cova', 'covb', and 'datos'. The History pane shows the execution history. The Console at the bottom displays the R startup message and the workspace loaded from ~/.RData.

Environment History

Global Environment

Data

- aircraft 709 obs. of 8 variables
- buffalo 62 obs. of 1 variable
- cora num [1:11, 1:11] 1 0.825 0.764 0.659 0.635 ...
- corb num [1:11, 1:11] 1 0.861 0.834 0.681 0.67 ...
- cova num [1:11, 1:11] 106 106 101 92 102 ...
- covb num [1:11, 1:11] 105.3 92 103.8 98.4 101.4 ...
- datos 32 obs. of 7 variables

Files Plots Packages Help Viewer

Install Update

User Library

Name	Description	Version
abind	Combine multi-dimensional arrays	1.4-0
aplpack	Another Plot PACKage: stem.leaf, bagplot, faces, spin3R, plotsummary, plothulls, and some slider functions	1.2.9
bdsmatrix	Routines for Block Diagonal Symmetric matrices	1.3-2
betareg	Beta Regression	3.0-5
bitops	Bitwise Operations	1.0-6
car	Companion to Applied Regression	2.0-20
CHAID	CHI-squared Automated Interaction Detection	0.1-2
coin	Conditional Inference Procedures in a Permutation Test Framework	1.0-23
colorspace	Color Space Manipulation	1.2-4
DAAG	Data Analysis and Graphics Data and Functions	1.22
dichromat	Color Schemes for Dichromats	2.0-0
digest	Create cryptographic hash digests of R objects	0.6.4
e1071	Misc Functions of the Department of Statistics (e1071), TU W/ien	1.6-3

Console ~ /

R is free software and comes with ABSOLUTELY NO WARRANTY. You are welcome to redistribute it under certain conditions. Type 'license()' or 'licence()' for distribution details.

R is a collaborative project with many contributors. Type 'contributors()' for more information and 'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or 'help.start()' for an HTML browser interface to help. Type 'q()' to quit R.

[workspace loaded from ~/.RData]

> |

# INTRODUCCIÓN-R.STUDIO

The screenshot displays the RStudio application window. The top menu bar includes File, Edit, Code, View, Plots, Session, Build, Debug, Tools, and Help. Below the menu is a toolbar with icons for file operations and running code. The main editor area on the left shows a script file named 'Untitled1.R' with two lines of code: '1' and '2'. The console at the bottom left shows the R startup message, including the R license and instructions on how to use the software. The environment pane on the right shows the 'Global Environment' with a list of data objects: 'aircraft' (709 obs. of 8 variables), 'buffalo' (62 obs. of 1 variable), 'cora' (num [1:11, 1:11] 1 0.825 0.764 0.659 0.635 ...), 'corb' (num [1:11, 1:11] 1 0.861 0.834 0.681 0.67 ...), 'cova' (num [1:11, 1:11] 106 106 101 92 102 ...), 'covb' (num [1:11, 1:11] 105.3 92 103.8 98.4 101.4 ...), and 'datos' (32 obs. of 7 variables). The 'Packages' pane on the right shows a list of installed and available packages, including 'abind', 'aplpack', 'bdsmatrix', 'betareg', 'bitops', 'car', 'CHAID', 'coin', 'colorspace', 'DAAG', 'dichromat', 'digest', and 'e1071'.

Console ~/ |

R is free software and comes with ABSOLUTELY NO WARRANTY. You are welcome to redistribute it under certain conditions. Type 'license()' or 'licence()' for distribution details.

R is a collaborative project with many contributors. Type 'contributors()' for more information and 'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or 'help.start()' for an HTML browser interface to help. Type 'q()' to quit R.

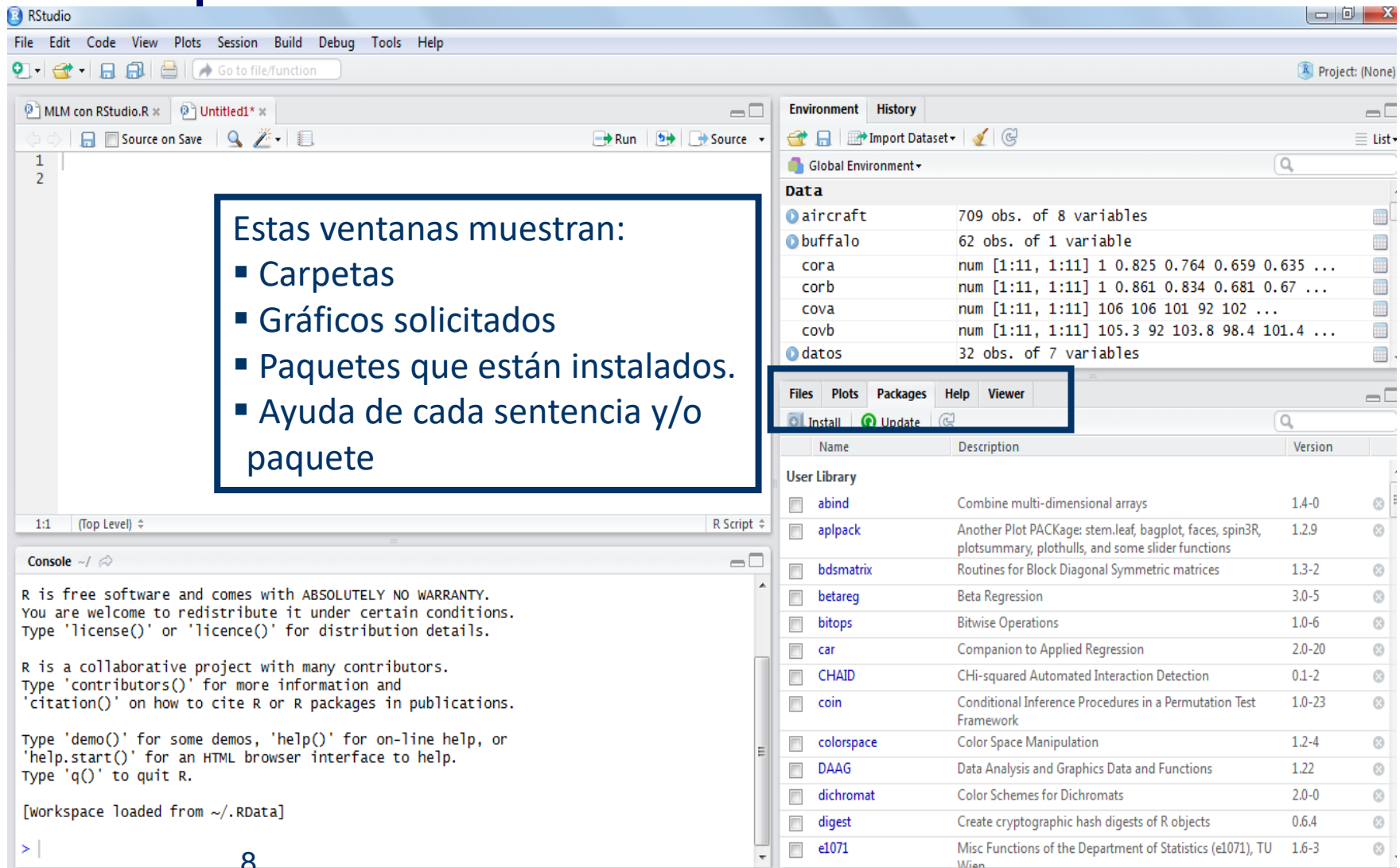
[workspace loaded from ~/.RData]

> |

■ Muestra las sentencias y resultados

Name	Description	Version
abind	Combine multi-dimensional arrays	1.4-0
aplpack	Another Plot PACKage: stem.leaf, bagplot, faces, spin3R, plotsummary, plothulls, and some slider functions	1.2.9
bdsmatrix	Routines for Block Diagonal Symmetric matrices	1.3-2
betareg	Beta Regression	3.0-5
bitops	Bitwise Operations	1.0-6
car	Companion to Applied Regression	2.0-20
CHAID	CHI-squared Automated Interaction Detection	0.1-2
coin	Conditional Inference Procedures in a Permutation Test Framework	1.0-23
colorspace	Color Space Manipulation	1.2-4
DAAG	Data Analysis and Graphics Data and Functions	1.22
dichromat	Color Schemes for Dichromats	2.0-0
digest	Create cryptographic hash digests of R objects	0.6.4
e1071	Misc Functions of the Department of Statistics (e1071), TU Witten	1.6-3

# INTRODUCCIÓN-R.STUDIO



The screenshot shows the RStudio environment. The top menu bar includes File, Edit, Code, View, Plots, Session, Build, Debug, Tools, and Help. The toolbar below the menu has icons for file operations and running code. The main editor window on the left contains a script with two lines of code. A text box is overlaid on the editor, listing features shown in the windows. The Environment pane on the right shows the Global Environment with a list of data objects. The Packages pane at the bottom right shows a list of installed and available packages. The Console pane at the bottom left shows the R startup message.

Estas ventanas muestran:

- Carpetas
- Gráficos solicitados
- Paquetes que están instalados.
- Ayuda de cada sentencia y/o paquete

Environment History

Global Environment

Data

Object	Obs	Vars
aircraft	709 obs.	of 8 variables
buffalo	62 obs.	of 1 variable
cora	num [1:11, 1:11]	1 0.825 0.764 0.659 0.635 ...
corb	num [1:11, 1:11]	1 0.861 0.834 0.681 0.67 ...
cova	num [1:11, 1:11]	106 106 101 92 102 ...
covb	num [1:11, 1:11]	105.3 92 103.8 98.4 101.4 ...
datos	32 obs.	of 7 variables

Files Plots Packages Help Viewer

Install Update

Name	Description	Version
abind	Combine multi-dimensional arrays	1.4-0
aplpack	Another Plot PACKage: stem.leaf, bagplot, faces, spin3R, plotsummary, plothulls, and some slider functions	1.2.9
bdsmatrix	Routines for Block Diagonal Symmetric matrices	1.3-2
betareg	Beta Regression	3.0-5
bitops	Bitwise Operations	1.0-6
car	Companion to Applied Regression	2.0-20
CHAID	CHI-squared Automated Interaction Detection	0.1-2
coin	Conditional Inference Procedures in a Permutation Test Framework	1.0-23
colorspace	Color Space Manipulation	1.2-4
DAAG	Data Analysis and Graphics Data and Functions	1.22
dichromat	Color Schemes for Dichromats	2.0-0
digest	Create cryptographic hash digests of R objects	0.6.4
e1071	Misc Functions of the Department of Statistics (e1071), TU W/ien	1.6-3

User Library

Console ~/

R is free software and comes with ABSOLUTELY NO WARRANTY. You are welcome to redistribute it under certain conditions. Type 'license()' or 'licence()' for distribution details.

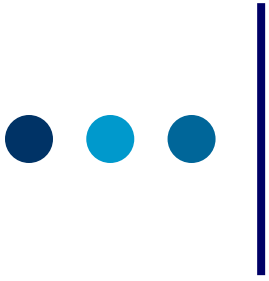
R is a collaborative project with many contributors. Type 'contributors()' for more information and 'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or 'help.start()' for an HTML browser interface to help. Type 'q()' to quit R.

[workspace loaded from ~/.RData]

> |





# REGRESIÓN SIMPLE

## Ejercicio 1



# REGRESIÓN SIMPLE

## Ejercicio 1

Los productores de caña de azúcar están interesados en la relación entre **la superficie de tierras cosechadas** (hectáreas) y la **producción total de caña de azúcar** (en toneladas) de esta superficie. Para dar respuesta a la inquietud de los productores se analizó la cosecha del año 2014 de **14 departamentos** productores de caña de azúcar del norte argentino y se relevaron los siguientes datos:



# REGRESIÓN SIMPLE

## Ejercicio 1

Departamento	Sup (Ha)	Producción
1	13638	940000
2	6151	460000
3	5828	440000
4	931	65000
5	12222	830000
6	5302	380000
7	11979	860000

Departamento	Sup (Ha)	Producción
8	8175	590000
9	13679	1020000
10	8296	585000
11	13396	1020000
12	3238	200000
13	16633	1130000
14	7244	570000



## Conjunto de datos

La tabla de datos se puede obtener mediante:

- ✓ El **ingreso** de datos mediante sentencia
- ✓ La **lectura** de ficheros de datos externos (texto, csv, Excel, etc.)

Para este ejemplo → se ingresan los datos mediante sentencia

# Conjunto de datos

```
Ejercicio_01.R x
# EJERCICIO 1: hectareas vs producción de caña azúcar
# creación del conjunto de datos EJER1 en el objeto data (data.frame)
dpto<-c(1,2,3,4,5,6,7,8,9,10,11,12,13,14)
ha<-c(13638,6151,5828,931,12222,5302,11979,8175,13679,8296,13396,3238,166
produccion<-c(940000,460000,440000,65000,830000,380000,860000,590000,1020
dpto<-as.factor(dpto)
datos1<-data.frame(dpto,ha,produccion)
datos1
```

## Objetos creados (vectores):






- ✓ **dpto**: contiene los números de los departamentos
- ✓ **ha**: contiene los valores de las superficies cosechadas (en hectáreas) de cada departamento
- ✓ **produccion**: contiene los valores de la producción de caña de azúcar (en toneladas) de cada departamento

# Conjunto de datos

```
Ejercicio_01.R x
# EJERCICIO 1: hectareas vs producción de caña azúcar
# creación del conjunto de datos EJER1 en el objeto data (data.frame)
dpto<-c(1,2,3,4,5,6,7,8,9,10,11,12,13,14)
ha<-c(13638,6151,5828,931,12222,5302,11979,8175,13679,8296,13396,3238,166
produccion<-c(940000,460000,440000,65000,830000,380000,860000,590000,1020
dpto<-as.factor(dpto)
datos1<-data.frame(dpto,ha,produccion)
datos1
```

- ✓ **as.factor(dpto)**: se considera al objeto **dpto** como un factor
- ✓ **datos1**: se crea este data frame con los objetos anteriores

# Conjunto de datos

Environment History	
    <span>List ▾</span>	
Global Environment ▾	
Data	
 datos1	14 obs. of 3 variables
Values	
dpto	Factor w/ 14 levels "1","2","3","4",...: 1 2 ...
ha	num [1:14] 13638 6151 5828 931 12222 ...
produccion	num [1:14] 940000 460000 440000 65000 830000...

Console ~/

```
> produccion<-c(940000,460000,440000,65000,830000,380000,860000,590000,1020000,585000,1020000,200000,1130000,570000)
> dpto<-as.factor(dpto)
> datos1<-data.frame(dpto,ha,produccion)
> datos1
```

	dpto	ha	produccion
1	1	13638	940000
2	2	6151	460000
3	3	5828	440000
4	4	931	65000
5	5	12222	830000
6	6	5302	380000
7	7	11979	860000
8	8	8175	590000
9	9	13679	1020000
10	10	8296	585000
11	11	13396	1020000
12	12	3238	200000
13	13	16633	1130000
14	14	7244	570000

&gt; |

# Medidas descriptivas

## 1.1 Calcular las medidas descriptivas para cada variable.

```
summary(datos1)
```



```
> summary(datos1)
```

	dpto	ha	produccion
1	:1	Min. : 931	Min. : 65000
2	:1	1st Qu.: 5909	1st Qu.: 445000
3	:1	Median : 8236	Median : 587500
4	:1	Mean : 9051	Mean : 649286
5	:1	3rd Qu.:13102	3rd Qu.: 920000
6	:1	Max. :16633	Max. :1130000
(Other):8			

```
colMeans(datos1[, -1])
```



```
> colMeans(datos1[, -1])
```

ha	produccion
9050.857	649285.714



# Gráfico de dispersión

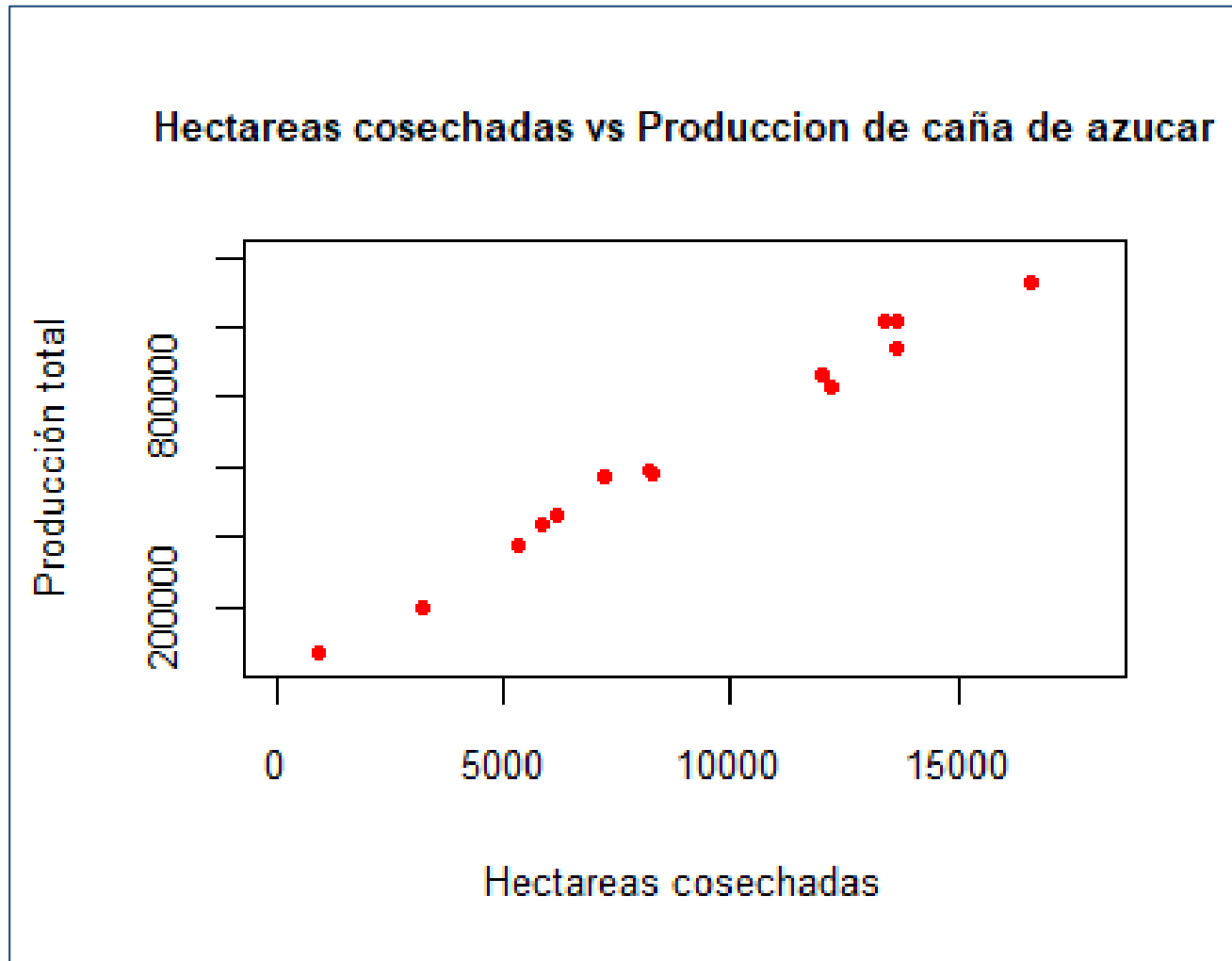
1.2 ¿Cómo es la relación existente entre las hectáreas cosechadas y las toneladas de producción de azúcar?

```
plot(produccion~ha,main="Hectareas cosechadas vs  
Produccion de caña de azucar", cex.main=0.8, ylab =  
"Producción total",xlab="Hectareas cosechadas",  
cex.lab=0.8, xlim=c(0,18000),ylim=c(50000,1200000),  
cex.axis=0.8,col="red",cex=0.75, pch=19)
```

# Gráfico de dispersión

- ✓ **plot (produccion~ha)** : gráfico de dispersión  $Y \sim X$
- ✓ **main="Hectareas ...."** : Título principal del gráfico
- ✓ **cex.main=0.8**: Tamaño del texto del título
- ✓ **ylab = "Producción total"**: Título del eje Y
- ✓ **xlab = "Hectareas cosechadas"**: Título del eje X
- ✓ **cex.lab=0.8**: Tamaño del texto de las etiquetas de los ejes
- ✓ **xlim=c(0,18000)**: Valores mínimos y máximos del eje X
- ✓ **ylim=c(0,1200000)**: Valores mínimos y máximos del eje Y
- ✓ **cex.axis=0.8**: Tamaño de los valores de los ejes
- ✓ **col="red"**: Color de los puntos
- ✓ **cex=0.95**: Tamaño del texto relativo al valor. Por defecto que es 1
- ✓ **pch=19 (círculo)**: Especifica el símbolo o caracter utilizado para representar los puntos del gráfico. Si se desea que los puntos sean representados por algún carácter específico se pone entre comillas, por ejemplo: **pch="&"**

# Gráfico de dispersión



# Modelo de regresión simple

## 1.3 Ajustar un modelo de regresión lineal simple y graficarlo junto a los datos

### Modelo planteado

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \quad i=1,2,\dots,14$$

$$\begin{matrix} \text{iid} \\ \varepsilon_i \sim N(0, \sigma^2) \end{matrix}$$

# Estimación del modelo

## Estimación del modelo (mínimos cuadrados)

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i \quad i = 1, 2, \dots, 14$$

$$\hat{\beta}_1 = b_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

21

$$\hat{\beta}_0 = b_0 = \bar{y} - b_1 \bar{x}$$

# Estimación del modelo

```
modelo1 = lm(produccion ~ ha, data=datos1)
```

## Paquete lm

- ✓ **Uso:** para ajustar modelos lineales.
- ✓ **Lm** (formula, data, method = "qr", etc )
  - **Formula: produccion ~ ha:** indica el ajuste de un modelo lineal con “produccion” como variable respuesta y “ha” como variable explicativa. Por defecto incluye ordenada al origen. En caso de querer excluirla se puede escribir:  $y \sim x - 1$  ó  $y \sim 0 + x$ .
  - **data=datos1:** nombre del conjunto de datos
- ✓ **modelo1:** objeto resultante

# Estimación del modelo

modelo1



**Coefficients:**

(Intercept)	ha
12324.901	70.376

$$\hat{y}_i = 12324,9 + 70,376 x_i \quad i = 1, 2, \dots, 14$$

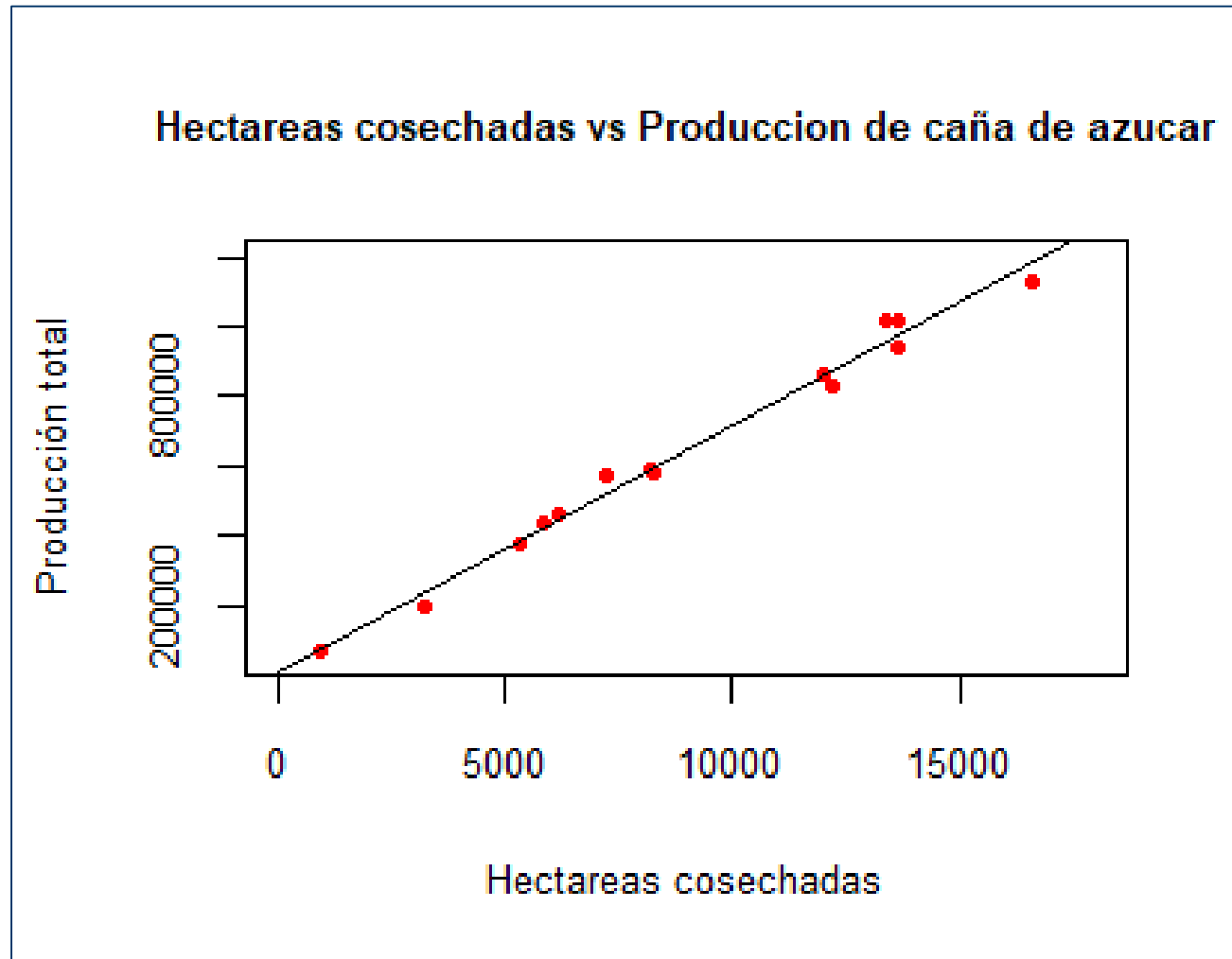
## Gráfico dispersión con recta estimada

```
plot(produccion~ha,main="Hectareas cosechadas vs  
Produccion de caña de azucar", cex.main=0.8, ylab =  
"Producción total",xlab="Hectareas cosechadas",  
cex.lab=0.8, xlim=c(0,18000),ylim=c(50000,1200000),  
cex.axis=0.8,col="red",cex=0.75, pch=19)  
abline(modelo1)
```

Al gráfico solicitado con la función **plot** se le agregue la **recta estimada** con el modelo 1



## Gráfico dispersión con recta estimada



# ● ● ● Interpretación coeficientes

## 1.4 Interpretar los coeficientes de regresión estimados en términos del problema

$$\hat{y}_i = 12324,9 + 70,376 x_i \quad i = 1, 2, \dots, 14$$

**b1=70,376:** *A medida que la superficie de tierras cosechadas **aumenta en 1 hectárea**, la **producción media** de caña de azúcar **aumenta en 70,376 toneladas**.*

**b0=12.324,9:** *No tiene sentido la interpretación cuando la cantidad de superficie cosechadas es **cero**.*



# Anova

## 1.5 Construir el cuadro Anova.

FV	SC	gl	CM	F
<b>Regresión (ajustada)</b>	$SCR_m = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2$	$p=1$	$CMR = \frac{SCR_m}{1}$	$F = \frac{CMR}{CME}$
<b>Error</b>	$SCE = \sum_{i=1}^n (y_i - \hat{y}_i)^2$	$n-p-1=n-2=12$	$CME = \frac{SCE}{n-2}$	
<b>Total (ajustado)</b>	$SCT_m = \sum_{i=1}^n (y_i - \bar{y})^2$	$n-1=13$		

# Anova

```
anova(modelo1)
```

## Analysis of Variance Table

Response: produccion

Variable respuesta

SCRm

(en regresión simple)

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
ha	1	1.3602e+12	1.3602e+12	980.25	7.095e-13 ***
Residuals	12	1.6651e+10	1.3876e+09		

---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

SCE

Estadística F



# Anova

FV	SC	gl	CM	F
<b>Regresión (ajustada)</b>	$SCR_m = 1360200000000$	1	$CMR = 1360200000000$	F = 980,25
<b>Error</b>	$SCE = 16651000000$	12	$CME = 1387583333$	
<b>Total (ajustado)</b>	$SCT_m = 1376851000000$	13		

# Test de Regresión

1.6 Probar la significación de la regresión utilizando la estadística F y la estadística t. Comparar los resultados.

**TEST DE REGRESIÓN:**  $H_0) \beta_1 = 0$   $H_1) \beta_1 \neq 0$

**Estadística t**

$$t = \frac{b_1}{\sqrt{\hat{V}(b_1)}} \sim t_{12}^{H_0}$$

**Regla de decisión**

Rechazo  $H_0$  si  $|t_{obs}| > t_{12; 0.025}$   
ó si p-value  $< 0.05$

**Estadística F**

$$F = \frac{CMR}{CME_{H_0}} \sim F_{1;12}$$

**Regla de decisión**

Rechazo  $H_0$  si  $F_{obs} > F_{1;12;0.05}$   
ó si p-value  $< 0.05$

# Test de Regresión

## Estadística F

$$F = \frac{CMR}{CME_{H_0}} \sim F_{1,12}$$

**F= 980,2 (del cuadro Anova)**

**$F_{1,12, 5\%} = 4,75$**

**¿Conclusión?**

# Test de Regresión

## summary(modelo1)

Call:  
lm(formula = produccion ~ ha, data = datos1)

### Residuals:

Min	1Q	Median	3Q	Max
-52885	-27293	-1552	16842	64922

### Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	12324.901	22649.679	0.544	0.596
ha	70.376	2.248	31.309	7.09e-13 ***

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 37250 on 12 degrees of freedom

Multiple R-squared: 0.9879, Adjusted R-squared: 0.9869

F-statistic: 980.2 on 1 and 12 DF, p-value: 7.095e-13

Sentencia a la que hacen referencia los resultados

Medidas descriptivas de los residuos

- Coeficientes estimados
- Error estándar de los coeficientes estimados
- Valor de la estadística t-student
- P-value

Referencias de los valores significativos

$\sqrt{CME}$

$R^2$

Valor de la estadística F, grados de libertad y probabilidad asociada



# Test de Regresión

## Estadística t

$$T = \frac{b_1}{\sqrt{\hat{V}(b_1)}} \sim t_{12}^{H_0}$$

### Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	12324.901	22649.679	0.544	0.596
ha	70.376	2.248	31.309	7.09e-13 ***

Corroborar  $t^2 = F$

¿Conclusión?

# Intervalo confianza- Betas

1.7 Construir un intervalo del 95% de confianza para  $\beta_1$  e interpretar.

`confint(modelo1)`



	2.5 %	97.5 %
(Intercept)	-37024.509	61674.31
ha	65.478	75.27

`t(confint(modelo1))`



	(Intercept)	ha
2.5 %	-37024.51	65.478
97.5 %	61674.31	75.273

IC  $\beta_1$ ;95% → (65,48 ; 75,27)

# Coeficiente $R^2$

## 1.8 ¿Qué valor toma el coeficiente de determinación $R^2$ ?

$$R^2 = \frac{SCR_m}{SCT_m} = 1 - \frac{SCE}{SCT_m} = \frac{130200000000}{(130200000000 + 1651000000)} = 0,9879$$

```
summary(modelo1) $r.squared
```

```
[1] 0.9879063
```

*O, como se vio anteriormente:*

```
summary(modelo1)
```

```
Multiple R-squared: 0.9879
```

**$R^2=0,9879$ :** El 98,79% de la variación de la producción de caña de azúcar se puede explicar por la relación lineal entre la superficie cosechada y la producción de caña de azúcar.

# IC respuesta media

1.9 ¿Cuál será la producción media esperada de azúcar para 5302 hectáreas cosechadas? Agregar un intervalo de confianza para su estimación.

```
predict(modelo1, list(ha=5302), interval="conf")
```

	fit	lwr	upr
1	385457.10	357038.65	413875.5

## Fit:

Valor estimado por el modelo para cada unidad (estimación puntual)

## Lwr:

Límite inferior del IC de confianza del 95% para la respuesta media

## Upr:

Límite superior del IC de confianza del 95% para la respuesta media

# IC respuesta media

Para todos los casos

```
predict(modelo1, interval="conf")
```

	fit	lwr	upr
1	972109.31	940880.83	1003337.8
2	445206.11	419278.99	471133.2
3	422474.74	395648.44	449301.0
4	77844.72	32546.45	123143.0
.....			
12	240201.56	204410.90	275992.2
13	1182884.67	1139879.80	1225889.5
14	522126.80	498699.82	545553.8

# Intervalo de predicción

1.10 Un grupo de productores de un determinado departamento están interesados en predecir cuál será la producción de azúcar al final de este año. Esperan cosechar 7244 hectáreas. Realizar la estimación puntual y por intervalo.

```
predict(modelo1, list(ha=7244), interval="pred")
```



	fit	lwr	upr
1	522126.8	437651.6	606602

# Intervalo de predicción

Para todos los casos

```
predict(modelo1, interval="pred")
```

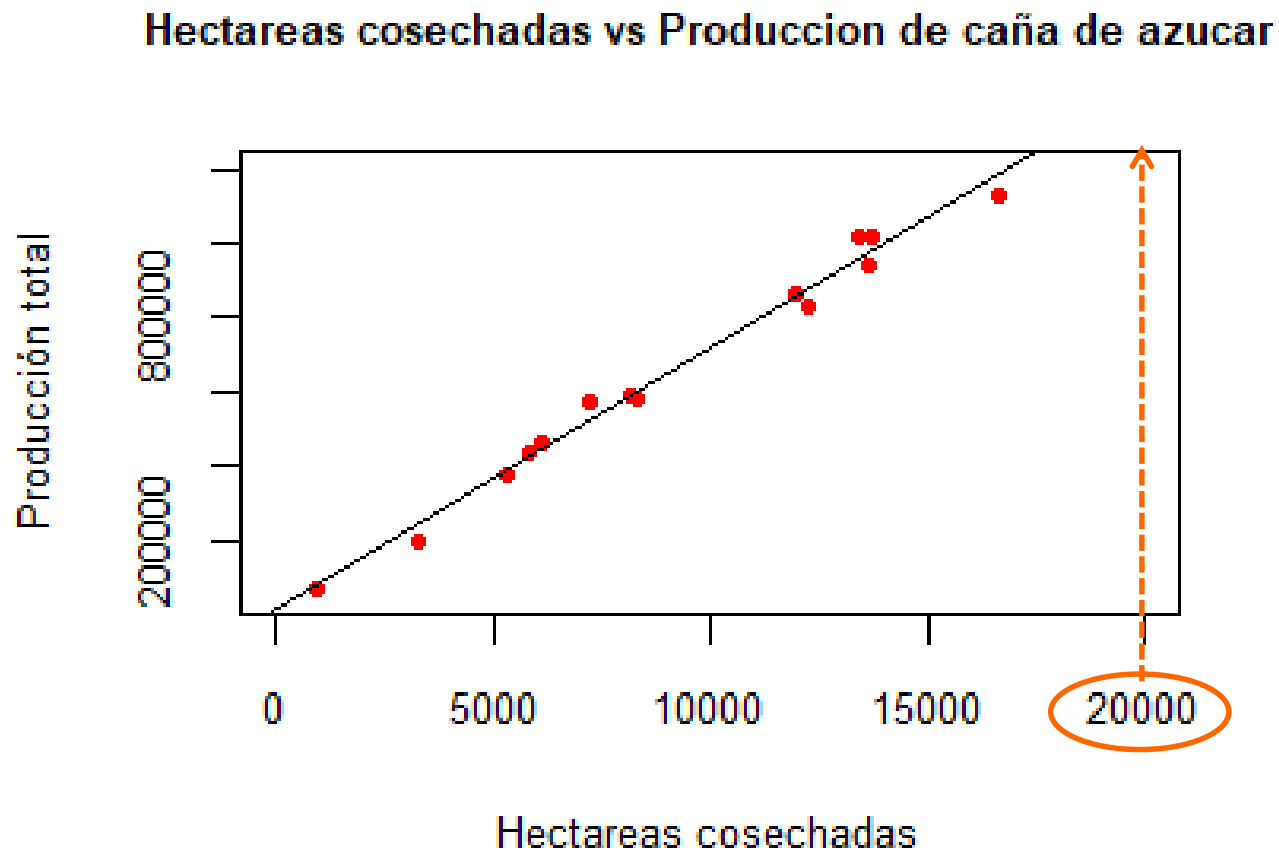
	fit	lwr	upr
1	972109.31	885146.92	1059071.7
2	445206.11	360003.67	530408.5
3	422474.74	336994.39	507955.1
4	77844.72	-15102.43	170791.9
5	872457.25	787023.31	957891.2
.....			
14	522126.80	437651.57	606602.0

**Warning message:**

**In predict.lm(modelo1, interval = "pred") :  
 predictions on current data refer to \_future\_  
 responses**

# Extrapolación

1.11 ¿Podemos utilizar este modelo para predecir la producción media esperada de azúcar para 20000 hectáreas cosechadas?





# Residuos

Obtener el valor del residuo para la observación 1.

```
residuos<-resid(modelo1)
residuos
```

<b>1</b>	<b>2</b>	<b>3</b>	<b>4</b>	<b>5</b>	<b>6</b>
-32109.310	14793.892	17525.258	-12844.720	-42457.255	-5457.100
<b>7</b>	<b>8</b>	<b>9</b>	<b>10</b>	<b>11</b>	<b>12</b>
4644.051	2353.385	45005.285	-11162.081	64921.621	-40201.563
<b>13</b>	<b>14</b>				
-52884.666	47873.203				

$$e_1 = y_1 - \hat{y}_1 = 940000 - (12324,9011 + 70,37574 \cdot 13638)$$

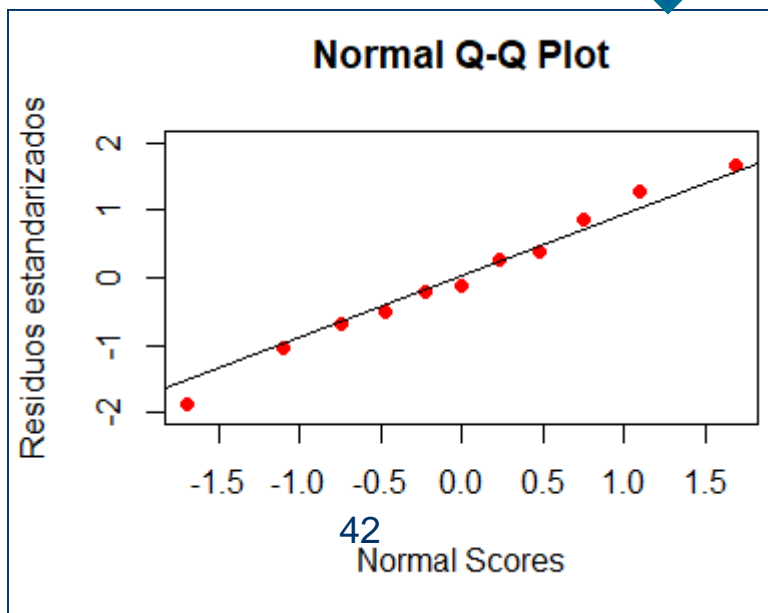
$$e_1 = y_1 - \hat{y}_1 = 940000 - 972109,243$$

$$e_1 = -32109,24$$

# Comprobación de los supuestos

Evaluar si los errores se distribuyen Normal

```
qqnorm(standresid, ylab="Residuos estandarizados",  
xlab="Normal Scores",ylim=c(-2,2),col="red",cex=0.95,  
pch=19)  
qqline(standresid)
```



Los puntos se ajusta a la recta

Sugiere que los errores tienen distribución Normal

# Comprobación de los supuestos

Evaluar si los errores se distribuyen Normal

`ad.test(residuos)`



Anderson-Darling normality test  
data: residuos A = 0.22207, p-value = 0.7875

$H_0$ ) Los errores tienen distribución normal

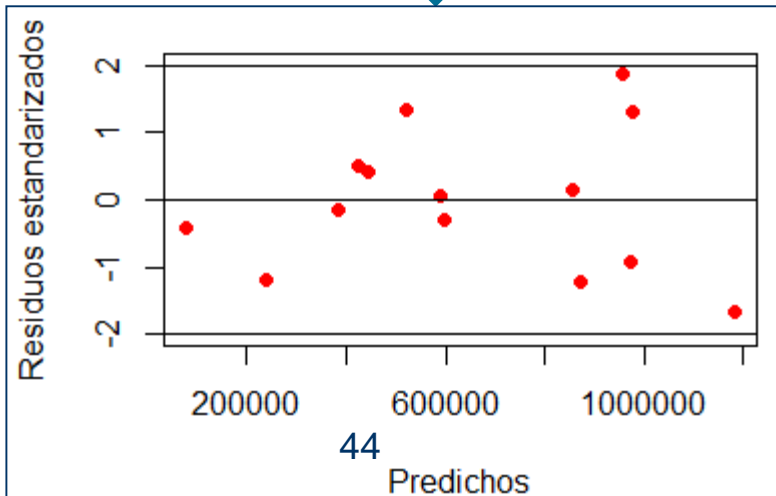
$H_1$ ) Los errores no tienen distribución normal

El valor  $p=0,7875 > 0,05$

# Comprobación de los supuestos

Evaluar si la media de los errores es igual a 0 y la variancia es constante

```
plot(standresid~predichos,ylab="Residuos estandarizados",  
xlab="Predichos",ylim=c(-2,2),col="red",cex=0.95, pch=19)  
abline(0,0)  
abline(-2,0)  
abline(2,0)
```



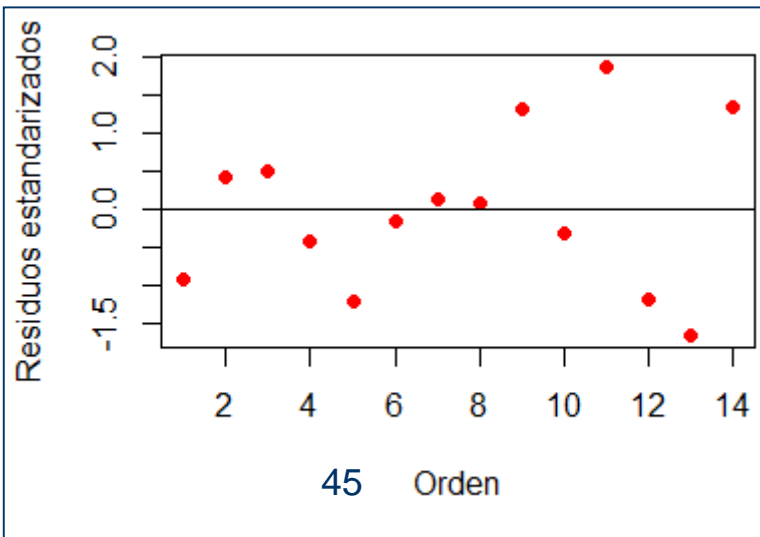
Los puntos caen dentro de una banda horizontal alrededor del cero

Sugiere que los errores tienen media cero y variancia constante

# Comprobación de los supuestos

Evaluar si los errores están correlacionados

```
plot(standresid,ylab="Residuos estandarizados",xlab =  
"Orden",col="red",cex=0.95, pch=19)  
abline(0,0)
```



Si los puntos no muestran un patrón,  
se presentan aleatoriamente

Sugiere que los errores están correlacionados

# Comprobación de los supuestos

Evaluar si los errores están correlacionados

```
dwtest(modelo1)
```



```
Durbin-Watson test  
data: modelo1 DW = 2.2587, p-value = 0.7454  
alternative hypothesis: true autocorrelation is greater than 0
```

$H_0$ ) Los errores no están correlacionados

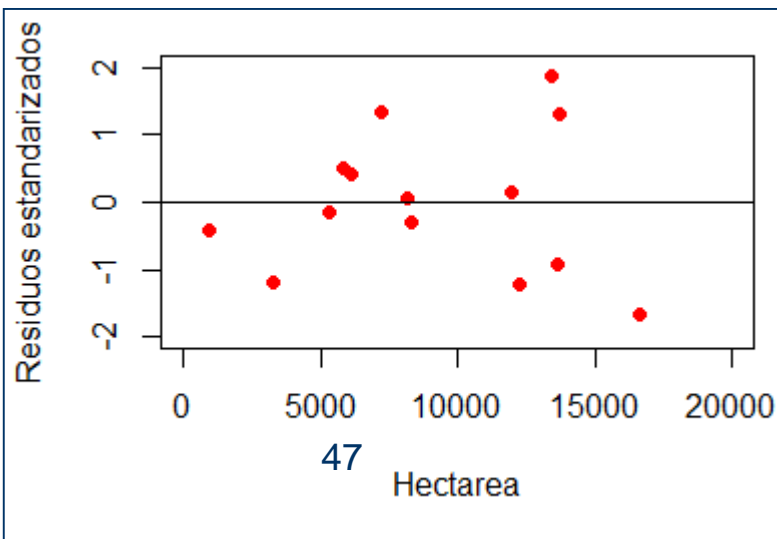
$H_1$ ) Los errores están correlacionados

El valor  $p=0,7454 > 0,05$

# Comprobación de los supuestos

Evaluar si la relación propuesta entre X e Y es adecuada

```
plot(standresid~ha, ylab="Residuos estandarizados",  
xlab="Hectarea",xlim=c(0,20000),ylim=c(-2,2),col="red",  
cex=0.95,pch=19)  
abline(0, 0)
```



Los puntos no muestran un patrón,  
se presentan aleatoriamente

Sugiere que la relación propuesta entre Y y X  
es correcta