REGRESION LINEAL SIMPLE

Selección de actividades resueltas

© Jose Fco. Martínez Boscá, Arnau Mir Torres, Lluís M. Pla Aragonés, Àngel J. Gil Estallo (Autors) & Ángel A. Juan (Editor)

© FUOC 2009

UOC Universitat Oberta de Catalunya

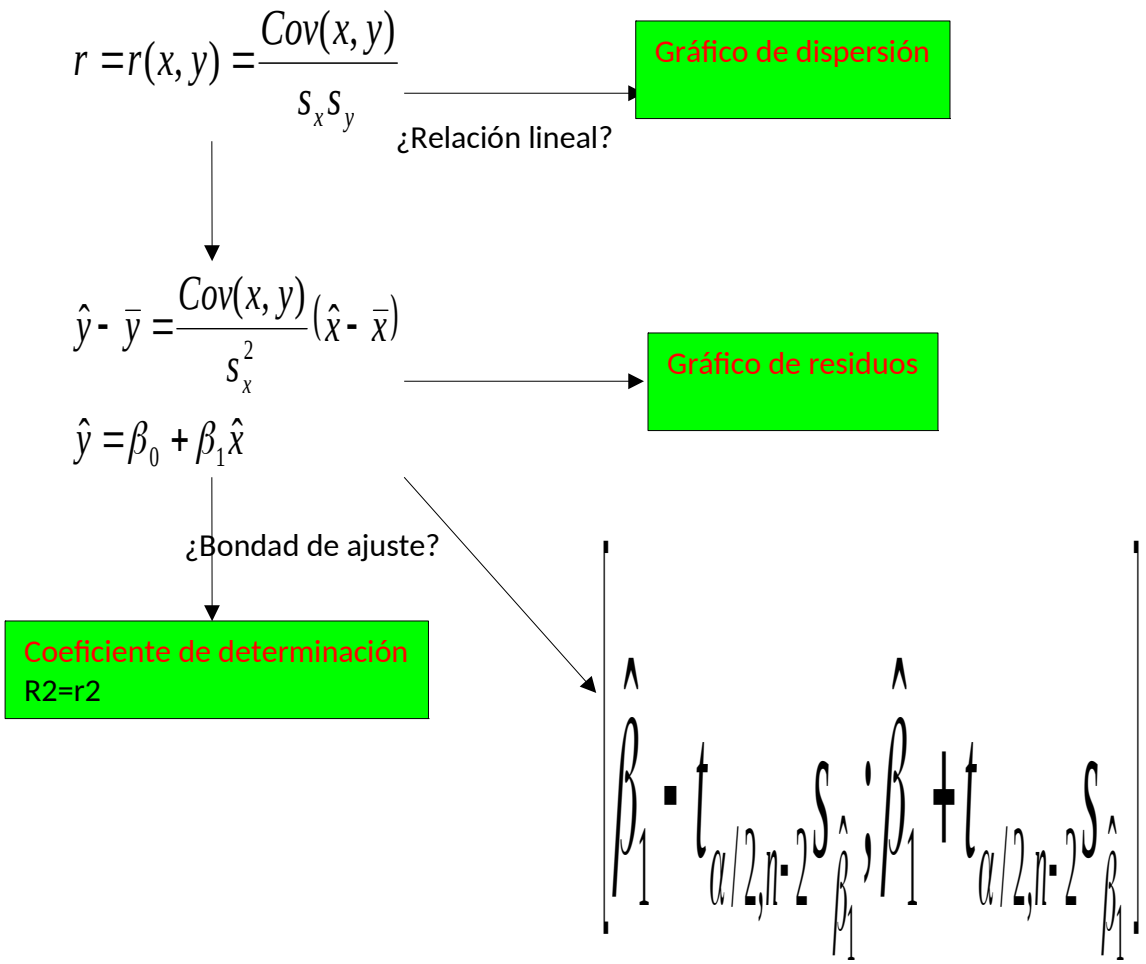
Introducción

En este *módulo*, se pretende estudiar las relaciones que se pueden presentar entre dos variables a diferencia de los análisis anteriores que se centraban en el estudio de una única variable. En concreto se estudiaran posibles relaciones lineales intentando encontrar una expresión que permitirá estimar una variable en función de la otra.
Por ejemplo, ¿en qué medida, un aumento de los gastos en software específico hace aumentar los beneficios de una empresa?, ¿cómo representamos que el aumento de prestaciones de los microprocesadores provocan un incremento en el uso de aplicaciones informáticas?, ¿cómo relacionar el número de accesos a internet con el nivel de estudios de los usuarios?...
Una primera aproximación al grado de relación entre dos variables en lo que llamaremos análisis de correlación. Para valorar esta relación de forma visual utilizaremos una representación gráfica llamada diagrama de dispersión y, finalmente, estudiaremos un modelo matemático para estimar el valor de una variable basándonos en el valor de otra, en lo que llamaremos análisis de regresión. El análisis de regresión incluirá adicionalmente la introducción de medidas que orienten sobe la bondad de esa relación lineal entre las variables estudiadas.

No se debe perder de vista que el objetivo principal del modelo de regresión es explicar el comportamiento de una variable (variable explicada) a partir de otra variable (variable explicativa).

Mapa conceptual

MODELO DE REGRESION SIMPLE



Actividades

Actividad 1: Estudio del rendimiento informático en función del número de buffers.

Modelo de regresión lineal. Software R.

En la tabla siguiente, se muestran las variables Y, rendimiento de un sistema informático, respecto a la variable regresora X_1, número de buffers:

X_1	5	10	15	20	25	5	10	15	20	25	5	10	15	20	25
Y	9.6	20.1	29.9	39.1	50.0	9.6	19.4	29.7	40.3	49.9	10.7	21.3	30.7	41.8	51.2

A partir de la tabla anterior, queremos ajustar la variable Y como función de la variable X_1.

a) Reproducir los resultados que se obtienen con la función lm() de R a la que añadiremos el gráfico de residuos frente de valores estimados.
b) Comentar los resultados siguientes:
b.1) Recta de regresión del rendimiento del sistema informático frente al número de buffers e interpretación de los coeficientes.
b.2) Gráfico de dispersión con el ajuste a la recta.
b.3) Contraste de hipótesis sobre la pendiente de la recta.
b.4) Coeficiente de determinación y coeficiente de regresión lineal.
b.5) Gráfico de residuos frente a valores estimados.

Solución

a) Para empezar introduciremos los datos. Lo haremos así:

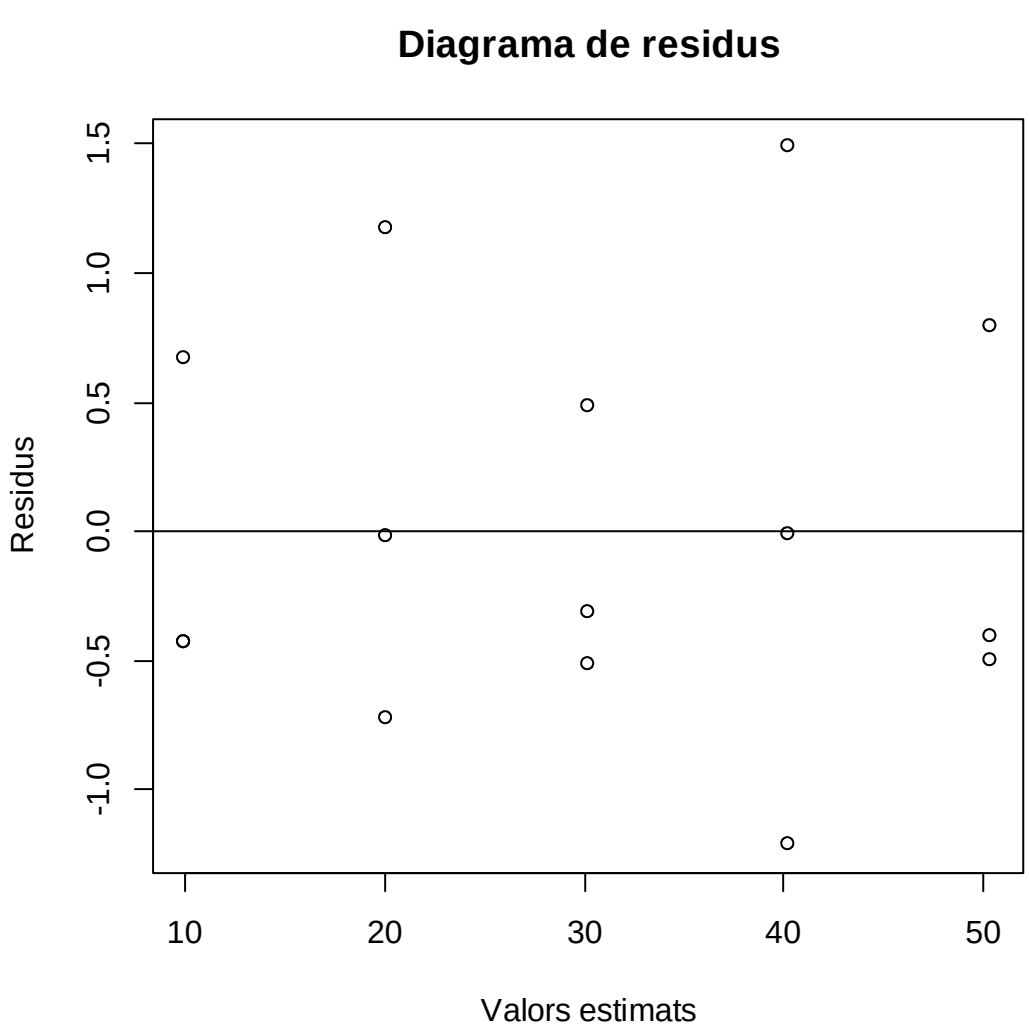
```
> x1 <- c(5,10,15,20,25,5,10,15,,20,25,5,10,15,20,25)
> y  <- c(9.6,20.1,29.9,39.1,50.0,9.6,19.4,29.7,40.3,49.9,10.7,21.3,30.7,41.8,51.2)
```

Ahora podemos calcular los parámetros utilizando la función lm() de R. Esta función estima modelos lineales, que es lo que queremos hacer.

```
> lm_11 <- lm(y~x1)
> summary(lm_11)

Call:
lm(formula = y ~ x1)

Residuals:
    Min      1Q  Median      3Q     Max
-1.2133 -0.4700 -0.3200  0.5733  1.4867

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.06000    0.47687  -0.126    0.902
x1           2.01867    0.02876  70.198   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7875 on 13 degrees of freedom
Multiple R-squared:  0.9974,    Adjusted R-squared:  0.9972
F-statistic:  4928 on 1 and 13 DF,  p-value: < 2.2e-16
```

Por tanto la recta de regresión del precio (Y) frente al número de páginas por minuto (X1) es
Y =-0.006 + 2.0186 X1:
Podemos dibujar el gráfico de los residuos frente a los valores estimados utilizando estas instrucciones de R:

```
> plot(residuals(lm_11)~fitted.values(lm_11), main = "Diagrama de residuos",
   xlab = "Valores estimados", ylab = "Residuos")
> abline(0,0)      #imprime una línea horizontal a 0
```

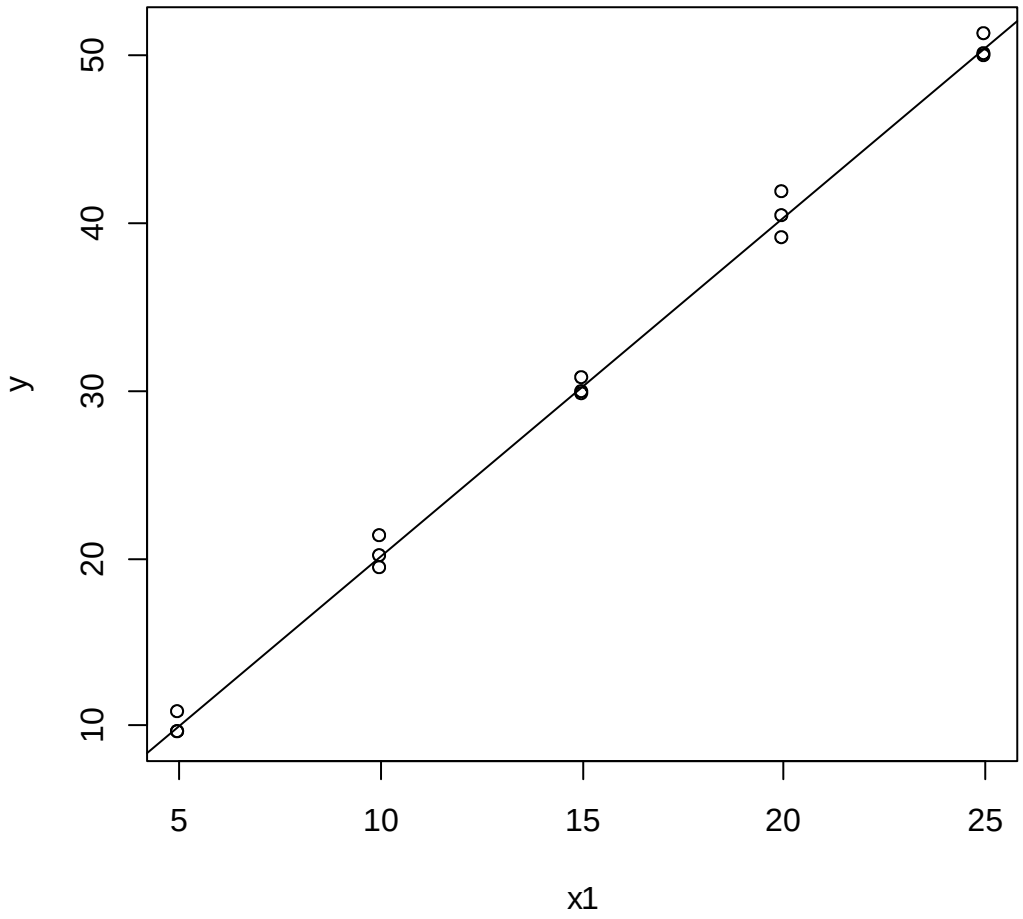


b.1) La recta de regresión del rendimiento informático (Y) frente al número de buffers (X_1) es:
Y =-0.006 + 2.0186 X1
Interpretación de los coeficientes:

- Pendiente de la recta (2.02): es el aumento de rendimiento informático por cada unidad de buffer añadida.
- Término independiente (-0.06): no tiene mucho sentido interpretarlo en este caso ya que representaría el rendimiento del sistema cuando no tenemos ningún buffer.

b.2) Para conseguir el gráfico de dispersión con el ajuste a la recta, con R hemos de hacer lo siguiente:

```
> plot(y~x1)
> abline(lm_11)
```



b.3)
El contraste se puede hacer automáticamente, haciendo uso del comando summary(lm_11) de R. La salida sería:

```
> summary(lm_11)

Call:
lm(formula = y ~ x1)

Residuals:
    Min      1Q  Median      3Q     Max
-1.2133 -0.4700 -0.3200  0.5733  1.4867

Coefficients:
            Estimate        Std. Error   t value   Pr(>|t|)
(Intercept) -0.06000    0.47687  -0.126    0.902
x1           2.01867    0.02876  70.198      <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7875 on 13 degrees of freedom
Multiple R-squared:  0.9974,    Adjusted R-squared:  0.9972
```

```
F-statistic:  4928 on 1 and 13 DF,  p-value: < 2.2e-16
```

El valor remarcado en negrita es el p-valor. Como el p-valor está muy próximo a 0, rechazaríamos la hipótesis nula para cualquier nivel de significación (por eso aparece ***, ver la leyenda) y aceptamos que la pendiente de la recta es representativa , distinta de cero.

b.4) El coeficiente de determinación vale: R-Sq = 99,7%. O sea, 0,997 y el coeficiente de correlación lineal será $r=\sqrt{0,997}=0,9985$.

b.5) El gráfico de residuos frente a valores estimados no tiene ningún tipo de estructura y está disperso alrededor del valor 0.
En conclusión el modelo de regresión es un buen modelo en este caso.

Actividad 2: Estudio sobre transmisión de información en redes informáticas.

Modelo de regresión lineal. Software R.

En un departamento de informática, un grupo de investigación dedicado al estudio de las comunicaciones por la red, desea conocer la relación entre el tiempo de transmisión de un fichero y la información útil del mismo. Para ello se han hecho algunos experimentos en los que se enviaban paquetes de distintas longitudes (bytes) de información útil y se medían los tiempos (en milisegundos) que tardaban desde el momento en que se enviaban hasta que llegaban al servidor. Los resultados del experimento se muestran en la tabla:

X (longitud)	100	110	120	150	190	200	225	265	280	300
Y (tiempo)	52	75	62	61	84	98	110	94	100	135

Se pide estudiar la relación entre las variables “tiempo” y “longitud” de los ficheros. Para ello, se pide:

a) Obtener la recta de regresión del tiempo en función de la longitud del fichero. Interpretad los resultados obtenidos.
b) Indicar el valor que toma el coeficiente de determinación y del coeficiente de correlación. Interpretar estos resultados
c) Estudiar la significación del modelo
d) Obtener el intervalo de confianza, al 95%, para la pendiente de la recta.
e) ¿Cuál será el tiempo de transmisión para un fichero que tiene una longitud de 250 bytes?

Solución

Introducimos los datos

```
> longitud<-c(100,110,120,150,190,200,225,265,280,300)
> tiempo<-c(52,75,62,61,84,98,110,94,100,135)
```

y podemos hacer los cálculos paso a paso:

a)

```
> mean(tiempo)
[1] 87.1
> var(longitud)
[1] 5332.222
> cov(tiempo,longitud)
[1] 1642.889
> beta1<-cov(longitud,tiempo)/var(longitud)
> beta1
[1] 0.3081059
> beta0<-mean(tiempo)-beta1*mean(longitud)
```

```
> beta0
[1] 27.32746
```

La recta de regresión resultante es: tiempo=27.3246 + 0.3081059 * longitud.

En este caso, la ordenada en el origen (27.327) no tiene sentido ya que sería el tiempo de transmisión de un fichero de 0 bytes de longitud. La pendiente de la recta nos indica que para cada byte de longitud el tiempo de transmisión aumenta en 0.308 ms.

b) Podemos calcular también el coeficiente de correlación lineal r y R^2

```
> r<-cov(longitud,tiempo)/sqrt(var(tiempo)*var(longitud))
> r
[1] 0.8823969
> r^2
[1] 0.7786243
```

El coeficiente de correlación lineal es de: $r = \frac{s_{xy}}{s_x\, s_x} = = 0.882$

Lo que nos indica que el ajuste es bastante bueno. El coeficiente de determinación:

$R^2 = r^2 = 0.779$, nos indica que nuestro modelo explica el 77.9% de la variabilidad de las observaciones.

c) Como se trata de un modelo de regresión lineal simple, para contrastar el modelo podemos hacer un contraste sobre la pendiente:

1) Hipótesis nula: $H_0: \beta_1 = 0$, es decir, la variable X no es explicativa.
   Hipótesis alternativa: $H_1: \beta_1 \neq 0$, es decir, la variable X es explicativa.

2) Nivel significativo: $\alpha = 0,05$

utilizando las funciones respectivas de R.

```
>r1<-lm(tiempo~longitud)
> summary(r1)

Call:
lm(formula = tiempo ~ longitud)

Residuals:
   Min     1Q  Median     3Q    Max
-14.976 -10.942  -2.084  12.274  15.241

Coefficients:
            Estimate   Std. Error   t value   Pr(>|t|)
(Intercept) 27.32746   11.96515     2.284     0.051754 .
longitud     0.30811    0.05808     5.304     0.000724 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 12.72 on 8 degrees of freedom
Multiple R-squared:  0.7786,    Adjusted R-squared:  0.751
F-statistic: 28.14 on 1 and 8 DF,  p-value: 0.0007244
```

$$t = \frac{\hat{\beta}_1}{s_{\hat{\beta}_1}} = 5.304$$

Estadístico de contraste como vemos en la salida de R:
el p-valor obtenido con R es 0.000724 por lo que se rechaza la hipótesis nula y se concluye con que la variable X es explicativa y por tanto, también lo es el modelo.

También se puede hacer a partir del valor crítico en lugar del p-valor. A partir de las tablas, se tiene que $\pm t_{\alpha/2,n-2} = \pm 2.306$ y dado que el estadístico de contraste 5.304 > 2.303, se rechaza la hipótesis nula, llegando a la conclusión comentada en el párrafo anterior.

d) El intervalo de confianza para la pendiente viene dado por:

$$\left[\hat{\beta}_1 - t_{\alpha/2,n-2}\, s_{\hat{\beta}_1} \leq \beta_1 \leq \hat{\beta}_1 + t_{\alpha/2,n-2}\, s_{\hat{\beta}_1}\right]$$

Utilizando los valores obtenidos en R

donde $\hat{\beta}_1 = 0.308$, $t_{0.025,8} = 2.306$, $s_{\hat{\beta}_1} = 0.05808$, de manera que se tiene:
el intervalo de confianza es:

```
> ic2<-c(beta1-t2*(0.05808),beta1+t2*(0.05808))
> ic2
[1] 0.1741731 0.4420386
```

e) Finalmente, para un fichero de longitud 250 bytes, podemos esperar un tiempo de transmisión de:

$$\hat{y} = 27.327 + 0.308 * 250 = 104.327\ ms$$

Actividad 3: Relación entre el precio y el tamaño de patallas TFT.

Regresión lineal simple. Bondad de ajuste. Recta de regresión. Gráfico de residuos. Software R.

Los precios de una pantalla TFT de una conocida marca son los siguientes:

Medida (pulgadas)	15	17	19	24
Precio (euros)	251	301	357	556

Calcular la recta de regresión para explicar el precio a partir del tamaño, indicar si es un buen ajuste y adjuntar los gráficos de la recta de regresión y de residuos.

Solución

Introducimos los datos

```
> mida<-c(15,17,19,24)
> preu<-c(251,301,357,556)
```

y podemos hacer los cálculos paso a paso:

```
> mean(mida)
[1] 18.75
> mean(preu)
[1] 366.25
> var(mida)
[1] 14.91667
> cov(preu,mida)
[1] 513.4167
> beta1<-cov(mida,preu)/var(mida)
> beta1
```

```
[1] 34.41899
> beta0<-mean(preu)-beta1*mean(mida)
> beta0
[1] -279.1061
```

La recta de regresión resultante es: Preu= -279.11 + 34.41 * Mida.

Podemos calcular también el coeficiente de correlación lineal r i R^2

```
> r<-cov(mida,preu)/sqrt(var(preu)*var(mida))
> r
[1] 0.994232
> r^2
[1] 0.9884974
```

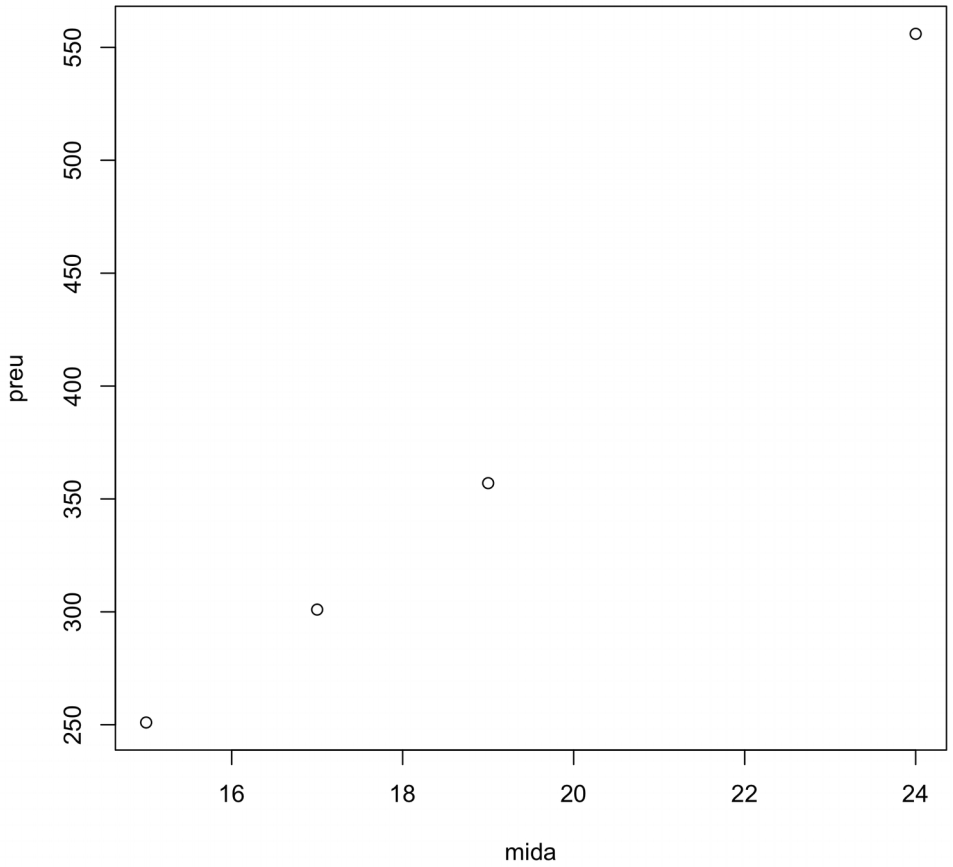
O también utilizando las funciones respectivas de R.

```
> r1<-lm(preu~mida)
> summary(r1)

Call:
lm(formula = preu ~ mida)

Residuals:
      1       2       3       4
 13.821  -5.017 -17.855   9.050

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -279.106     50.003  -5.582  0.03063 *
mida          34.419      2.625  13.110  0.00577 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 17.56 on 2 degrees of freedom
Multiple R-squared: 0.9885,   Adjusted R-squared: 0.9827
F-statistic: 171.9 on 1 and 2 DF,  p-value: 0.005768
```

En los dos casos tenemos que la recta de regresión es: Preu = 34.419 *Mida - 279.1061 y un R^2=0.9884974.

En cuanto a los gráficos podemos hacer:

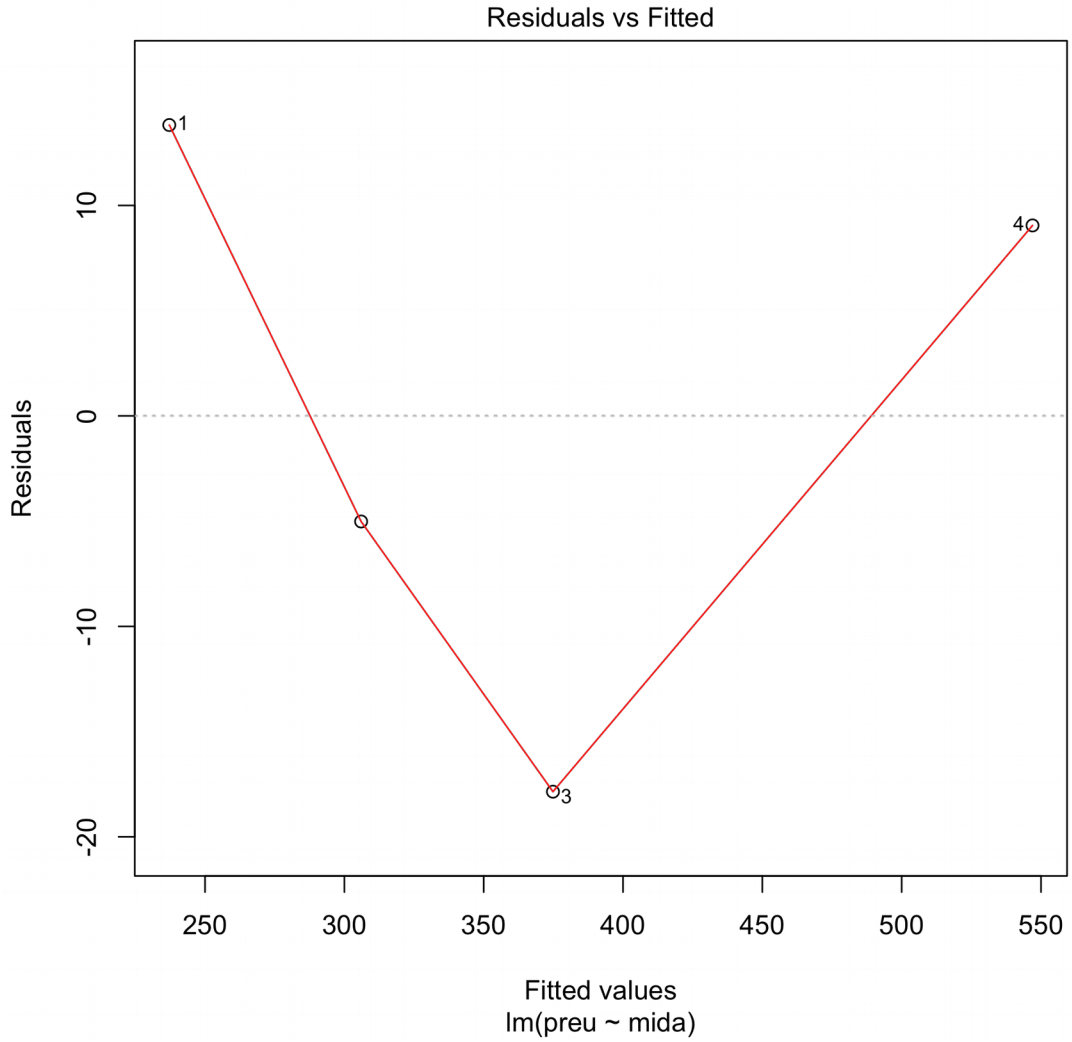
```
>  plot(preu~mida)
```



Si hacemos:

```
> plot(r1)
```

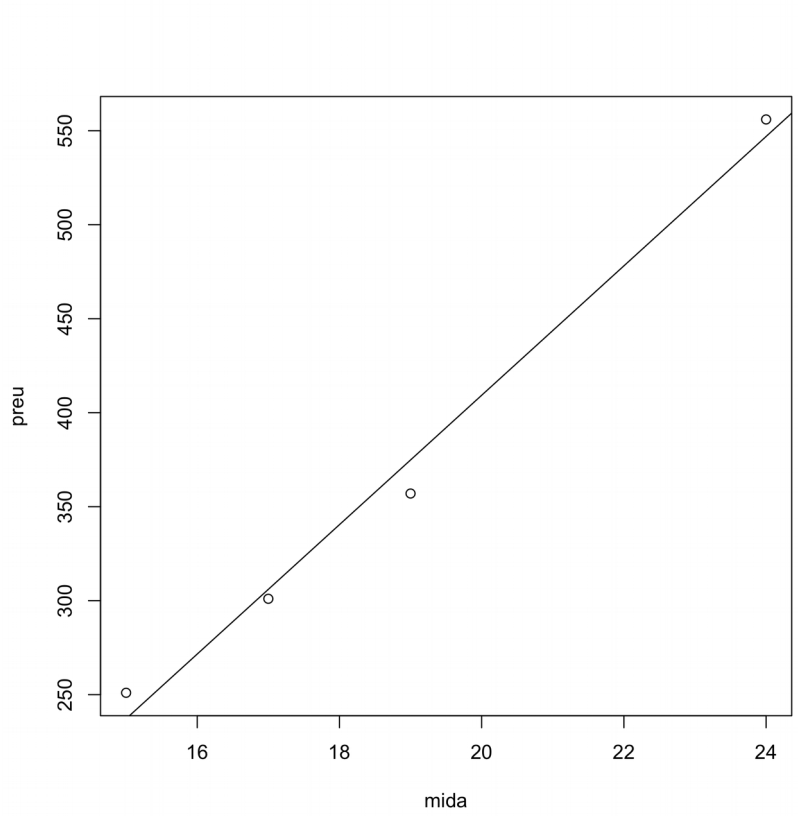
obtenemos el gráfico de residuos:



Y haciendo:

```
> plot(preu~mida)
> abline(r1)
```

obtenemos el gráfico de dispersión con la recta de regresión.



Actividad 4: Relación entre el precio y el tamaño de patallas TFT.

Regresión lineal simple. Intervalo de confianza para la pendiente. Software R.

Si considerásemos todos los modelos de TFT que hay en una nave de un gran almacén (son exactamente 12) encontraríamos la recta de regresión $Y=-250+30.5X$ (donde X indica el tamaño en pulgadas y Y el precio en euros). Además obtenemos que $s^2=82$ i $\sum(x_i-\bar{x})^2=39.$ Con esta información determinar un intervalo de confianza al 95% para la pendiente de la recta.

Solución

El intervalo de confianza viene dado por:

$$\left(\hat{\beta}_1-t_{\alpha/2,n-2}s_{\hat{\beta}_1},\hat{\beta}_1+t_{\alpha/2,n-2}s_{\hat{\beta}_1}\right).$$

Calcularemos cada elemento del intervalo:

a) Del enunciado extraemos $\hat{\beta}_1=30.5,$

```
> beta1_2<-30.5
```

b) $t_{0,025,n-2}=t_{0.025,10}=2.228139$ ya que

```
> t2<-qt(0.025,10,lower.tail=FALSE)
 > t2
[1] 2.228139
```

c) Calculemos ahora

$$s_{\hat{\beta}_1} = \frac{s^2}{\sum (X_i - \overline{X})^2} = \frac{82}{39} = 2.1026.$$

y por tanto el intervalo de confianza es:

```
> ic2<-c(beta1_2-2*(82/39),beta1_2+t2*(82/39))
> ic2
[1] 26.29487 35.18480
```

Actividad 5: Relación entre la velocidad de impresión y el precio de unas impresoras.

Regresión lineal simple. Bondad del ajuste. Gráfico de la recta de regresión. Gráfico de residuos. Coeficiente de correlación lineal. Coeficiente de determinación. Software R.

Disponemos de la siguiente muestra que contiene datos en los que Y representa el precio en euros de una serie de impresoras y X_1 representa el número de páginas por minuto que la impresora es capaz de imprimir.

X_1	6	6	6	6	8	8	8	8	12	12	12	12
Y	466	418	434	487	516	462	475	501	594	553	551	589

Queremos ajustar la variable Y, precio de la impresora a la variable X_1, número de páginas por minuto.
a) Hallar la recta de regresión del precio de la impresora en función del número de páginas por minuto e interpretar los resultados.
b) Hacer el gráfico de dispersión con el ajuste de la recta.
c) Contrastar la hipótesis sobre la pendiente de la recta.
d) Calcular el coeficiente de determinación y el coeficiente de regresión lineal.
e) Hacer un gráfico de los residuos frente a los valores estimados.

Solución

Para empezar introduciremos los datos. Lo haremos así:

```
> x1 <- c(6,6,6,6,8,8,8,8,12,12,12,12)
> y  <- c(466,418,434,487,516,462,475,501,594,553,551,589)
```

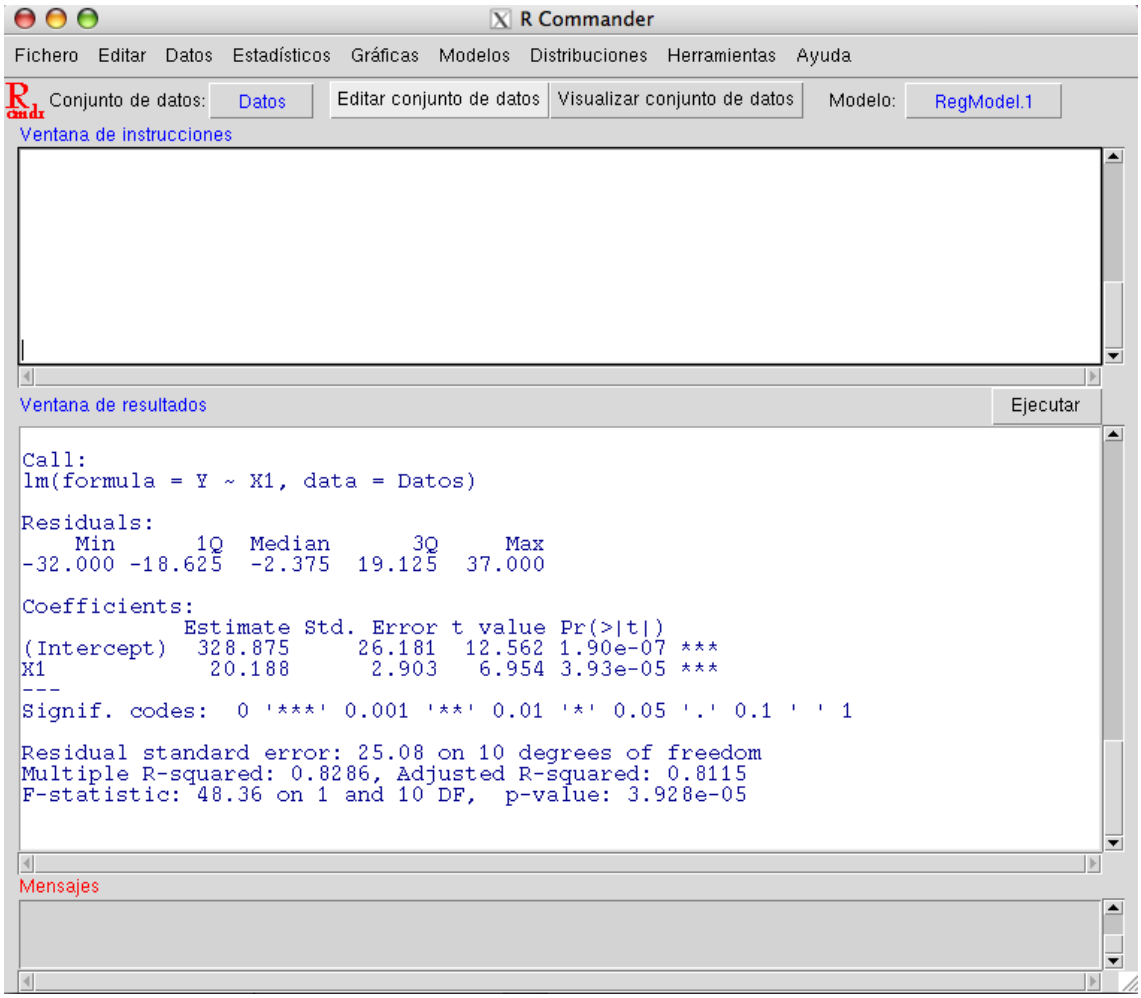
Ahora podemos calcular los parámetros utilizando la función lm() de R. Esta función estima modelos lineales, que es lo que queremos hacer.

```
> lm_11 <- lm(y~x1)
> summary(lm_11)

Call:
lm(formula = y ~ x1)

Residuals:
    Min      1Q  Median      3Q     Max
-32.000 -18.625  -2.375  19.125  37.000

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  328.875     26.181  12.562 1.90e-07 ***
x1            20.187      2.903   6.954 3.93e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 25.08 on 10 degrees of freedom
Multiple R-squared: 0.8286,     Adjusted R-squared: 0.8115
F-statistic: 48.36 on 1 and 10 DF,  p-value: 3.928e-05
```

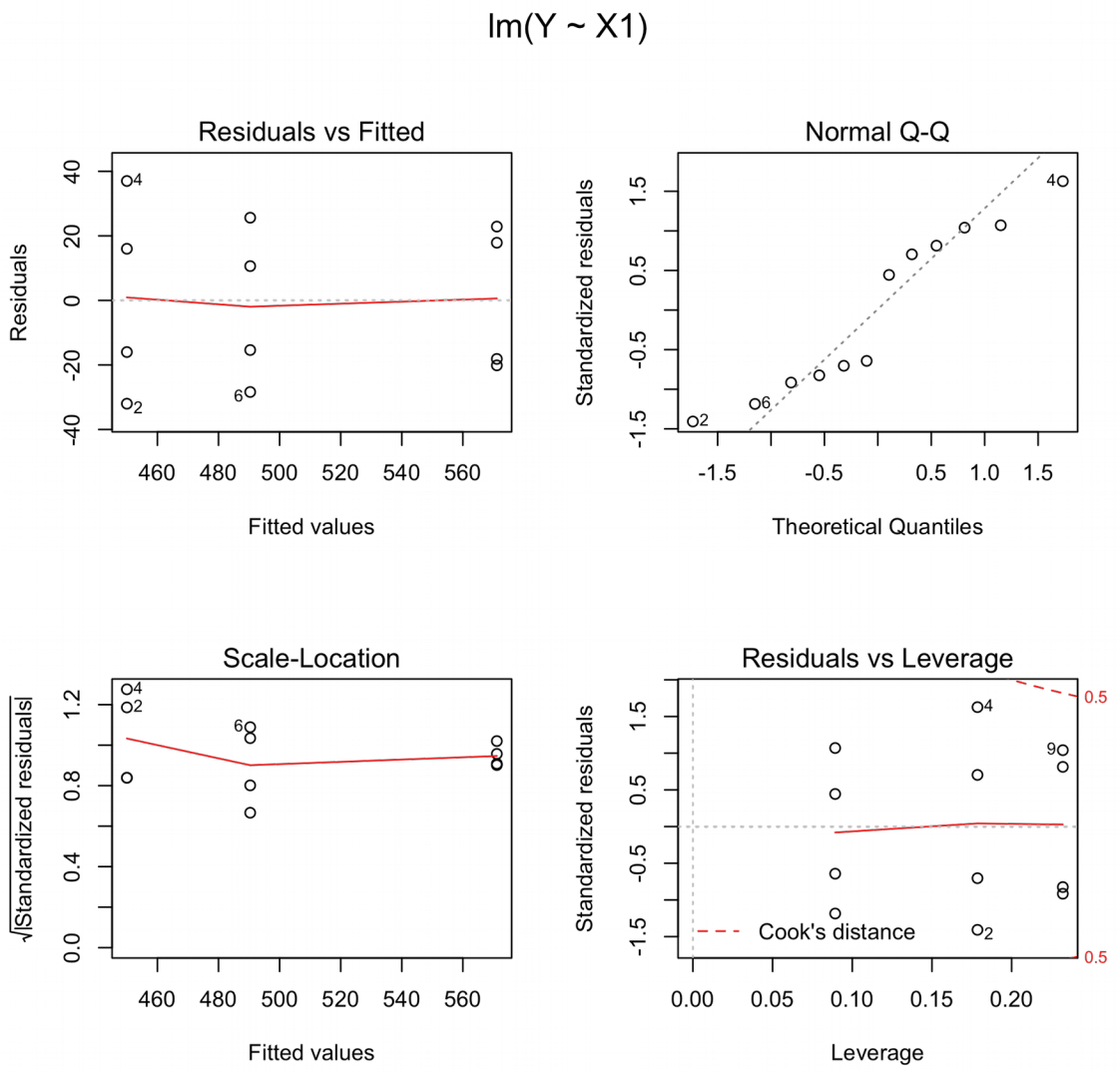
Por tanto la recta de regresión del precio (Y) frente al número de páginas por minuto (X1) es Y = 329 + 20,2 X1:

Estos valores nos indican que el precio base de la impresora es de 328.87 €, y quiere decir que una supuesta impresora que imprimiera 0 páginas por minuto valdría esa cantidad. Este valor carecería de sentido visto de esta manera.
El que sí que tiene sentido es el otro parámetro, la pendiente de la recta de regresión. Representa lo que aumentaría Y por cada unidad de X1. Es decir, por cada hoja por minuto de más que queramos (dentro del rango estudiado, está claro: 6 - 12, ya que resultaría peligroso aventurarse a hacer predicciones sobre más o menos páginas por minuto) tendríamos que pagar 20.19€ euros de más.

Con Rcomander obtendríamos una pantalla como esta (des de *Estadísticos -> Ajustes de modelos -> Regresión lineal*):

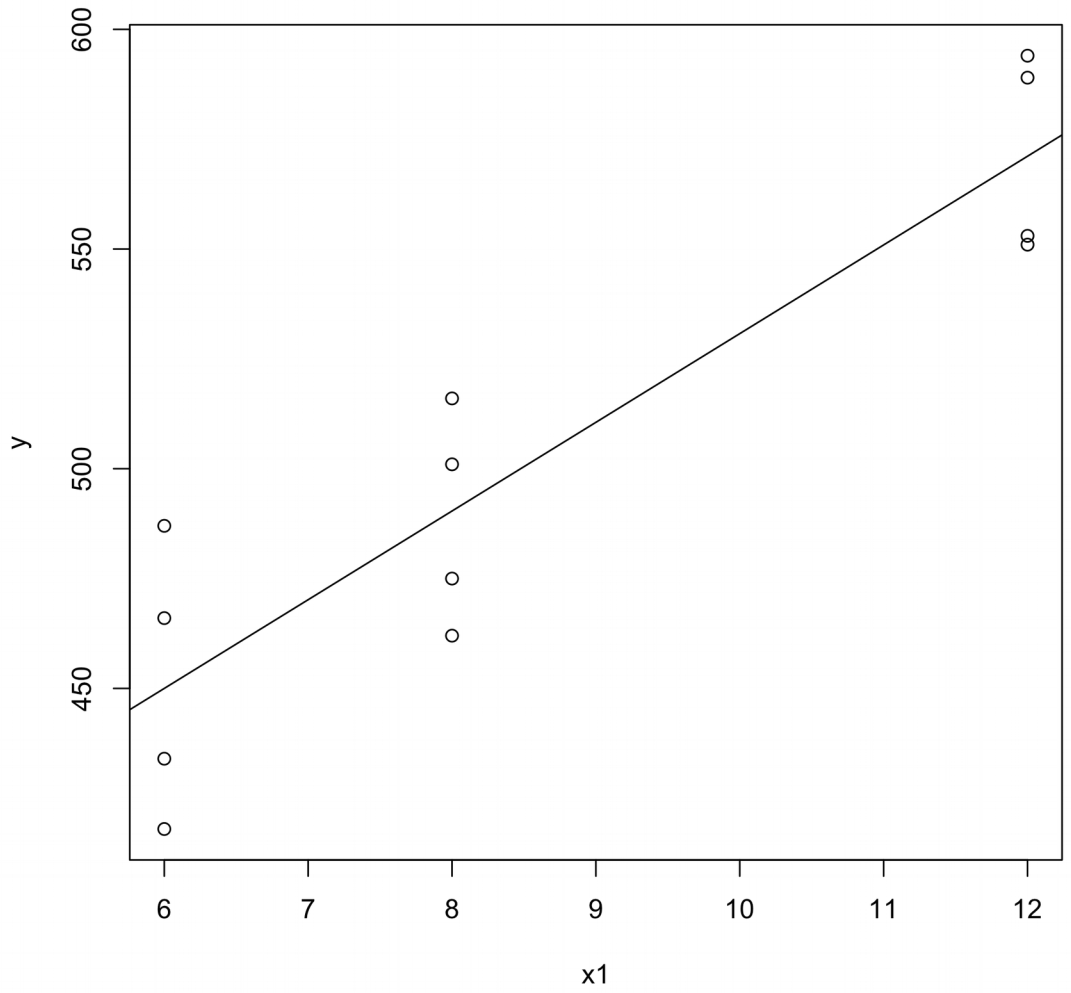


y desde *Modelos -> Gráficas -> Gráficas básicas de diagnóstico* obtendríamos



b) Para conseguir el gráfico pedido con R hemos de hacer lo siguiente:

```
#fa un gràfic de les observacions de la taula inicial
> plot(y~x1)
> abline(lm_11)   #afegeix al gràfic fet anteriorment la recta de
regressió
```



Donde observamos los datos facilitados por el enunciado (los puntos) y la recta de regresión calculada.

c) El contraste se puede hacer automáticamente, haciendo uso del comando summary(lm_11) de R. La salida sería:

```
> summary(lm_11)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  328.875     26.181  12.562 1.90e-07 ***
x1            20.187      2.903   6.954 3.93e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 25.08 on 10 degrees of freedom
Multiple R-squared: 0.8286,     Adjusted R-squared: 0.8115
F-statistic: 48.36 on 1 and 10 DF,  p-value: 3.928e-05
```

El valor remarcado en negrita es el p-valor. Como el p-valor está muy próximo a 0, rechazaríamos la hipótesis nula para cualquier nivel de significación (por eso aparece ***, ver la leyenda).

d) Como se trata de un modelo de regresión simple, el coeficiente de determinación es igual al cuadrado del coeficiente de correlación muestral ($R^2 = r^2$), por tanto lo primero que hay que hacer es calcular el coeficiente de correlación muestral y a partir de ese es que calcularemos el de determinación:

```
> r <- cov(x1,y) / (sd(x1)*sd(y))
> r
[1] 0.9103004
```

Ahora calculamos el coeficiente de determinación elevando al cuadrado el resultado anterior tal y como se explicó:

```
> R2 <- r^2
> R2
[1] 0.8286468
```

Tal y como se vió, la salida de summary(lm_11) también nos da la información que necesitamos, en este caso:

```
> summary(lm_11)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  328.875     26.181  12.562 1.90e-07 ***
x1            20.187      2.903   6.954 3.93e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 25.08 on 10 degrees of freedom
Multiple R-squared: 0.8286,     Adjusted R-squared: 0.8115
F-statistic: 48.36 on 1 and 10 DF,  p-value: 3.928e-05

En aquest cas el valor de R^2 és 0.8286468. A partir d'aquest podem
deduir el de r, ja que sabem que el pendent de la recta de regressió
és positiu. Aquest serà:

> +sqrt(0.8286)
[1] 0.9102747
```

e) Podemos dibujar el gráfico de los residuos frente a los valores estimados utilizando estas instrucciones de R:

```
> plot(residuals(lm_11)~fitted.values(lm_11), main = "Diagrama de
residus",
  xlab = "Valors estimats", ylab = "Residus")
> abline(0,0)      #imprimeix una línea horitzontal a 0
```

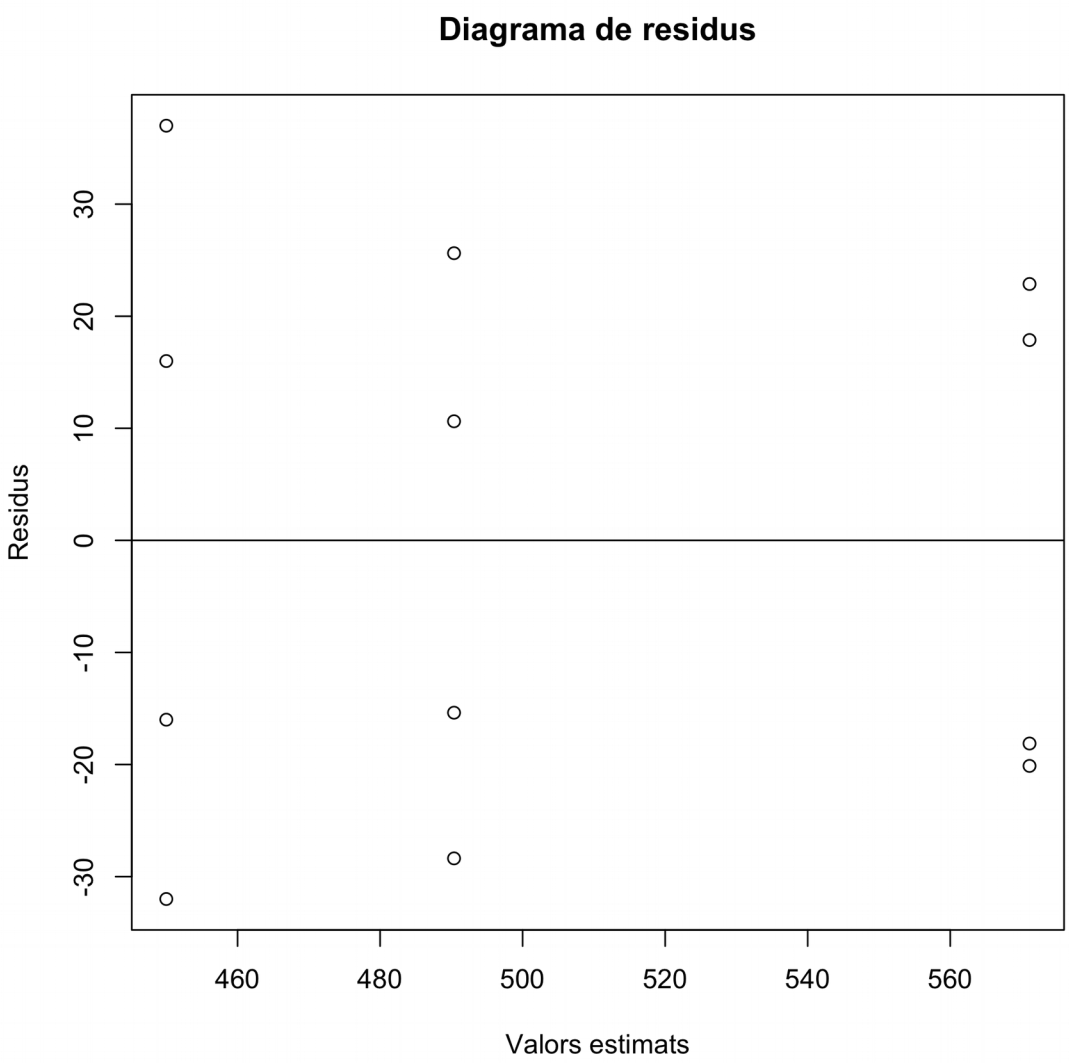


Diagrama de residus
Residus
30
20
10
0
-10
-20
-30
460
480
500
520
540
560
Valors estimats

Direcciones de interés

http://www.bioestadistica.uma.es/libro/node42.htm
Éste texto es la versión electrónica del manual de la Universidad de Málaga y habla sobre la regresión lineal.

http://www.uoc.edu/in3/e-math/docs/RegresionLineal.pdf
Math-block del proyecto e-math sobre regresión lineal con teoría y ejemplos con y sin Minitab.

http://www.udc.es/dep/mate/estadistica2/cap6.html
Apuntes y notas sobre regresión lineal de la Universidad de La Coruña.