Estadística climatológica con R

3. Regresión y Correlación

El propósito del ***análisis de regresión y correlación*** es el estudio de la relación existente entre dos variables aleatorias, una denominada *independiente*, *explicativa*, *de predicción* o *covariable*, bajo el control del experimentador, habitualmente representada por *X* y con valores en el eje de abscisas, y otra denominada *dependiente*, *explicada*, o *de interés*, habitualmente representada por *Y* y con valores en el eje de ordenadas.

El análisis de la regresión se ocupa de estudiar la relación existente entre dos o más variables aleatorias, mientras que ***el análisis de la correlación*** investiga el grado o fuerza de dicha relación.

Lo primero que debe realizar el investigador es representar las observaciones de ambas variables en un gráfico llamado *diagrama de dispersión* o *nube de puntos*. A partir de esta representación el investigador puede especificar la forma funcional de *la función de regresión.*

A menudo se supone que la relación que guardan la variable dependiente y la independiente es lineal. En estos casos, se utilizan los **modelos de regresión lineales**. Aunque las relaciones lineales aparecen de forma frecuente, también es posible considerar otro tipo de relación entre las variables, que se modelizan mediante otros modelos de regresión, como pueden ser el modelo de regresión *exponencial, potencial, logarítmico, polinomico, etc.*

Regresión: Consiste en la búsqueda de una "función" que exprese lo mejor posible el tipo de relación entre dos o más variables.

Correlación: Estudia el grado de dependencia entre las variables, es decir su objetivo es medir el grado de ajuste existente entre la función teórica (función ajustada) y la nube de puntos. Por lo tanto, una variable independiente que presente un alto grado de correlación con una variable dependiente será muy útil para predecir los valores de ésta última. Cuando la relación entre las variables es lineal, se habla de **correlación lineal**.

Si la relación es de tipo lineal entre dos variables aleatorias se denomina ***Regresión Lineal Simple***, mientras que cuando se consideran más de dos covariables se hablara de ***Regresión Lineal Múltiple***.

3.1 Modelo de regresión lineal simple

La situación general que se plantea en este modelo es la de 2 variables aleatorias, *X* e *Y*, estando interesados en inferir la existencia o no de una relación lineal entre ambas, de la forma

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

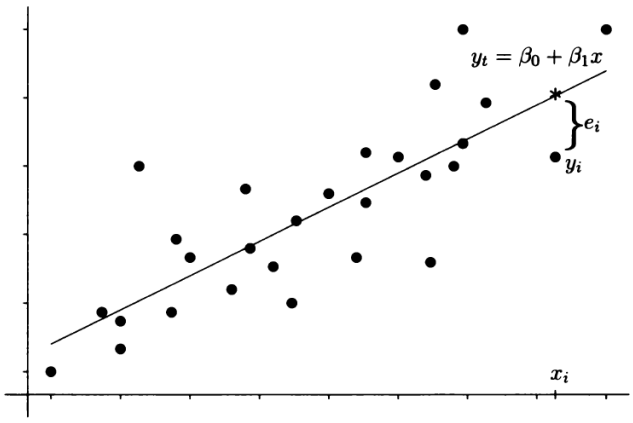


Fig. 3.1

Para unos valores x_i de las variables *X* obtendremos valores

$$y_i = \beta_0 + \beta_1 x_i + e_i$$

de la variable *Y*, los cuales no llegan a estar sobre la recta $y_t = \beta_0 + \beta_1 x$ debido al *error del muestreo* e_i. Los parámetros β_0 y β_1 se denominan *coeficientes de regresión*.

El modelo de regresión lineal supone que *los errores e_i son independientes y con distribución normal N(0,σ). (La distribución normal la veremos en el capitulo 4)*

Si tenemos una nubes de puntos (x_i,y_i) trataremos de determinar la recta $y_i = \beta_0 + \beta_1 x_i + e_i$ lo más próxima a la nube de puntos en el sentido de mínimos cuadrados, es decir determinar los valores de β_0 y β_1 que hagan mínima la suma de los cuadrados de las desviaciones e_i entre los valores observados y_i y los teóricos dados por la recta y_{ti} como se ve en la figura 3.1.

Resolviendo el sistema de ecuaciones se llega a

$$\beta_1 = \frac{n\sum_{i=1}^n x_i\, y_i - (\sum_{i=1}^n x_i)\,(\sum_{i=1}^n y_i)}{n\sum_{i=1}^n x_i^2 - (\sum_{i=1}^n x_i)^2}$$

y

$$\beta_0 = \frac{\sum_{i=1}^n y_i - \beta_1 \sum_{i=1}^n x_i}{n}.$$

Veremos cómo se hace un ajuste lineal con *R*.

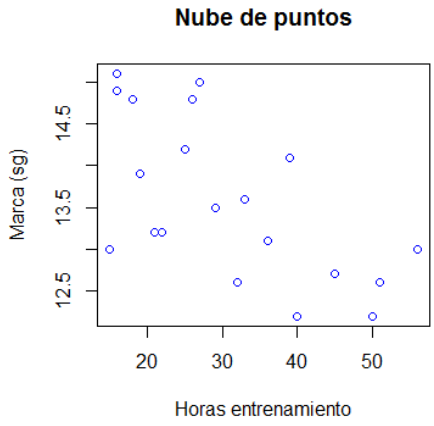
Ejemplo 3.1

Se preguntó a 20 corredores la marca que poseían en los 100 metros lisos y las horas semanales, que por término medio, dedicaban a entrenar y se obtuvieron los siguientes datos

Horas	21	32	15	40	27	18	26	50	33	51
Marca	13'2	12'6	13	12'2	15	14'8	14'8	12'2	13'6	12'6
Horas	36	16	19	22	16	39	56	29	45	25
Marca	13'1	14'9	13'9	13'2	15'1	14'1	13	13'5	12'7	14'2

Vamos a ver si existe una relación lineal entre las variables aleatorias: marca personal del corredor y horas de entrenamiento. Primero representaremos los datos gráficamente

```
> x <- c(21,32,15,40,27,18,26,50,33,51,36,16,19,22,16,39,56,29,45,25)
> y <- c(13.2,12.6,13,12.2,15,14.8,14.8,12.2,13.6,12.6,13.1,14.9,13.9,13.2,15.1,14.1,13,
13.5,12.7,14.2)

> plot(x,y, main='Nube de puntos', xlab='Horas entrenamiento',ylab='Marca (sg)', col=4)     #
ploteamos los puntos
```



Parece a simple vista que podría haber una relación lineal de pendiente negativa, es decir mejores marcas cuanto mayor es el tiempo dedicado, aunque la posible recta está no está muy bien definida en el gráfico.

Para calcular la recta de ajuste lineal de los datos en *R* utilizamos la función ***lm***. Esta recta la pintaremos sobre la nube de puntos. El ajuste se calcula de la forma:

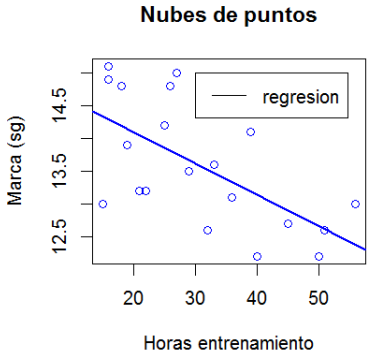
```
> ajuste <- lm( y ~ x) # regresión lineal
> ajuste

Call:
lm(formula = y ~ x)

Coefficients:
(Intercept)            x
   15.05908     -0.04786
```

Los coeficientes están escritos en

```
> ajuste$coefficients
> ajuste$coefficients
 (Intercept)           x
15.05908402 -0.04785987
```

Para añadir la recta estimada en el gráfico de puntos

```
> abline (ajuste)
> abline (ajuste, lwd=2, col=4) # la misma recta con grosor 2 y color 4
> legend (30, 15, c('regresion'), lty=1) )  # añade un rotulo en las coordenadas (30,15)
```



Precisión o bondad del ajuste por mínimos cuadrados

La nube de puntos del ejemplo anterior no parece muy concentrada alrededor de la recta de ajuste, lo que nos hace dudar de la bondad del ajuste.

Quizás el problema del ajuste es que ambas variables no están relacionadas linealmente. Un atleta no podrá hacer una marca más baja de un cierto límite por muchas horas que entrene. Si aumentásemos el número de horas enormemente la recta nos daría unos tiempos imposibles en la realidad. Es probable que para este tipo de variables fuese mejor una función de tipo *exponencial* del tipo $y_{ti} = ab^{xi}$ que se podría ajustar linealmente tomando logaritmos.

En ocasiones puede que el ajuste se corresponda mejor a una función *potencial* $y_{ti} = ax_i^b$.
Otras veces será necesario un *un polinomio de grado n* para conseguir un buen ajuste.

Necesitamos pues un valor que nos dé una medida de *la bondad del ajuste*. Como hemos usado el criterio de los mínimos cuadrados de las desviaciones entre los datos y los datos estimados por la curva de ajuste, parece razonable que una vez que hemos calculado los parámetros de cada función, elijamos la función que obtenga un menor valor para esa suma de cuadrados.

Este valor recibe el nombre de ***varianza residual***

$$V_r = \frac{1}{n}\sum_{i=1}^{n}(y_i - y_{t_i})^2$$

La función optima seria aquella para la que la varianza residual fuese cero, es decir aquella que pasase por todos los puntos y_i.

Vamos a tratar de encontrar otro valor que nos permita decidir si un ajuste es adecuado o no en sí mismo.

Se define ***el coeficiente de Determinación*** definido como

$$R^2 = 1 - \frac{V_r}{s_y^2}$$

siendo V_r la varianza residual y $s_y{}^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - a_y)^2$ la varianza marginal de las y_i. Este coeficiente <u>*está comprendido entre 0 y 1*</u>, hablándose de un buen ajuste si R^2 está cerca de 1 y malo si está próximo a 0.

Para el caso de un ajuste lineal existe un valor llamado ***Coeficiente de Correlación Lineal*** de Pearson, definido como

$$r = \frac{n\sum_{i=1}^{n} x_i\, y_i - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{\sqrt{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}\sqrt{n\sum_{i=1}^{n} y_i^2 - \left(\sum_{i=1}^{n} y_i\right)^2}}$$

Este coeficiente *oscila entre -1 y 1*. Si está cerca de estos extremos -1 (*correlación negativa*) o 1 (*correlación positiva*) es un buen ajuste, mientras que si está cerca de 0 nos diría que el ajuste lineal no nos explica la relación entre las variables *X* e *Y*. Aunque los valores 0 o 1 no se dan nunca, valores de |r| cercanos a 1 o 0 permiten interpretar el grado o fuerza de la relación entre ambas variables, siempre que esta relación sea *lineal*, si no no sería de utilidad.

El coeficiente de determinación es igual al cuadrado del coeficiente de correlación.

El *coeficiente de correlación* en *R* se calcula con la función ***cor***

> cor(x,y)

```
> cor(x,y)
[1] -0.6304069
```

Luego el *coeficiente de determinación* es R^2=cor(x,y)*cor(x,y) = 0,3974

Contraste de la regresión lineal simple

La recta de regresión lineal siempre se puede hacer pero en algunos casos explicara bien a la variable dependiente en función de la independiente y en otros casos no lo hará. En algunos casos la recta de regresión se puede usar para hacer predicciones de Y dados unos x concretos y en otros casos hará unas predicciones bastante malas.

Mediante *la Inferencia Estadística* y unos *test de hipótesis* habrá que validar o no la recta de regresión obtenida.

Análisis de la variación explicada frente a la no explicada por la recta de regresión

Entre dos funciones que ajustamos por mínimos cuadrados a una nube de puntos, debemos elegir aquella que obtenga *una menor varianza residual*.

En esta sección contrastaremos la regresión lineal simple mediante ***test de hipótesis***.

Contraste de hipótesis para β_1

Un test de hipótesis para analizar si puede considerarse valida la recta de regresión determinada, es contrastar si se puede aceptar que es cero o no el parámetro $\boldsymbol{\beta_1}$ de la ecuación de regresión lineal entre ambas variables.

Si se rechaza ***la hipótesis nula*** $\boldsymbol{H_0: \beta_1 = 0}$ y se acepta la alternativa $\boldsymbol{H_1: \beta_1 \neq 0}$ *la regresión lineal será aceptable*. También se dice que existe una relación lineal significativa ya que de hecho el test ha resultado significativo.

Se hace el cálculo e interpretación por medio del ***p-valor***. Si el ***p-valor*** menor que $\boldsymbol{\alpha}$ *se rechaza la hipótesis nula* al nivel de significación $\boldsymbol{\alpha}$, que se suele tomar como 0.05 (95%), *si es mayor se acepta* H_0.

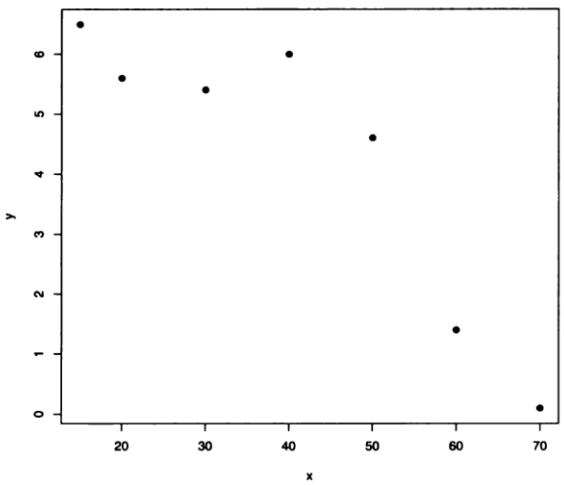
Ejemplo 3.2

Consideremos el ejemplo siguiente:

Se midió el contenido de oxígeno, variable *Y*, a diversas profundidades, variable *X*, en el lago Worther de Australia, obteniéndose los siguientes datos, en miligramos por litro

X	15	20	30	40	50	60	70
Y	6′5	5′6	5′4	6	4′6	1′4	0′1

La representación de los datos en un diagrama de dispersión es:



De esta figura parece desprenderse una relación lineal inversa entre profundidad y cantidad de oxígeno, es decir a mayor profundidad menor cantidad de oxígeno.

Primero incorporamos los datos y hacemos el ajuste lineal ***lm***

```
> x <- c(15,20,30,40,50,60,70)
> y <- c(6.5,5.6,5.4,6,4.6,1.4,0.1)

> plot(x,y, col=4,pch=19)    # ploteamos los puntos, pch=19 (círculos rellenos)
```

La recta de regresión se calculara:

```
> ajuste <- lm( y ~ x) # regresión lineal
Call:
lm(formula = y ~ x)

Coefficients:
(Intercept)            x
     8.6310      -0.1081
```

La recta de regresión ajustada es por tanto:

$$y = 8.6310 - 0.1081x$$

```
> abline(ajuste)  # añadimos la recta estimada en el gráfico de puntos
```

Ahora tenemos que analizar si la covariable *X* explica suficientemente bien a la variable dependiente *Y*; es decir si puede aceptarse o no la hipótesis nula de ser cero el coeficiente de regresión de *X*, es decir H_0: $\beta_1 = 0$. Para ellos ejecutamos:

```
> summary (ajuste)
```

```
Call:
lm(formula = y ~ x)

Residuals:
       1        2        3        4        5        6        7
-0.50907 -0.86841  0.01289  1.69419  1.37550 -0.74320 -0.96190

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  8.63102    1.07747   8.010  0.00049 ***
x           -0.10813    0.02399  -4.508  0.00635 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.204 on 5 degrees of freedom
Multiple R-squared:  0.8025,    Adjusted R-squared:  0.7631
F-statistic: 20.32 on 1 and 5 DF,  p-value: 0.006352
```

Esta salida contiene una información más completa sobre el análisis. Así, por ejemplo, encontramos información sobre los **residuos** (en el apartado ***Residuals***), que se definen como la diferencia entre el verdadero valor de la variable dependiente y el valor que pronostica el modelo de regresión. Cuanto más pequeños sean estos residuos mejor será el ajuste del modelo a los datos y más acertadas serán las predicciones que se realicen a partir de dicho modelo.

En la tabla ***Coefficients*** encontramos los valores de los parámetros que aparecían en la salida por defecto junto a su error estándar. Cada parámetro aparece acompañado de un ***p-valor*** que sirven para contrastar la **significación del parámetro** en cuestión, es decir, para resolver los siguientes contrastes de hipótesis:

$$H_0 \equiv \beta_0 = 0 \quad vs \quad H_1 \equiv \beta_0 \neq 0$$
$$H_0 \equiv \beta_1 = 0 \quad vs \quad H_1 \equiv \beta_1 \neq 0$$

Lo que se pretende mediante estos contrastes, es determinar si los efectos de la constante y de la variable independiente son realmente importantes para explicar la variable dependiente o si, por el contario, pueden considerarse nulos.

En nuestro ejemplo, los p-valores que nos ayudan a resolver estos contrastes son **0.00049** y **0.00635** ambos menores que 0.05. Así, considerando un nivel del significación del 5%, *rechazamos la hipótesis nula* en ambos contrastes, de manera que podemos suponer *ambos parámetros significativamente distintos de 0*.

Por último, en la parte final de la salida, encontramos el valor de $\mathbf{R^2}$ (*Multiple R-squared*) y de $\mathbf{R^2}$ **ajustado** (*Adjusted R-squared*), que son indicadores de la bondad del ajuste de nuestro modelo a los datos. $\mathbf{R^2}$ oscila entre 0 y 1, de manera que, valores de $\mathbf{R^2}$ próximos a 1 indican un buen ajuste del modelo lineal a los datos. Por otro lado, $\mathbf{R^2}$ ajustado es similar a $\mathbf{R^2}$, pero penaliza la introducción en el modelo de variables independientes poco relevantes a la hora de explicar la variable dependiente. Por tanto, $\mathbf{R^2}$ ajustado <= $\mathbf{R^2}$. En nuestro ejemplo, $\mathbf{R^2}$ = **0.8025** y $\mathbf{R^2}$ ajustado = **0.7631,** por lo que *podemos concluir que el modelo lineal se ajusta de forma aceptable a nuestros datos*.

La última línea de la salida incluye el ***p-valor*** conjunto de la prueba. Mediante este contraste se comprueba si, *de forma global*, el modelo lineal es apropiado para modelizar los datos.

En nuestro ejemplo, el *p-valor=0.00635* asociado a este contraste es inferior a *0.05* por lo que, al 5% de significación podemos rechazar la hipótesis nula y concluir con que β_1 es significativamente distinto de cero y afirmar que, efectivamente, *el modelo lineal es adecuado para nuestro conjunto de datos*.

Ejercicio 3.1

A veces en climatología se usa una serie climatológica de una estación para rellenar lagunas u homogeneizar estaciones cercanas. Se suelen usar estaciones meteorológicas cercanas de características geográficas similares, sobre las que actúan, en principio, las mismas situaciones meteorológicas. Normalmente se tiene una estación con datos depurados y comprobados, que sea homogénea, que consideramos más fiable, y la usamos para rellenar lagunas o depurar estaciones cercanas (veremos esto más detenidamente en el capítulo 5).

Consideremos dos series de temperaturas medias mensuales de dos estaciones. Tomaremos la estación de Pontevedra y la de Vigo.

- Examinar si existe una relación lineal entre ellas, representando primero la nube de puntos y después haciendo el modelo de regresión lineal.
- Escribir la recta de regresión hallada y pintarla sobre la nube de puntos.
- Ver la fuerza del ajuste lineal calculando el *coeficiente de correlación* y hacer el test de hipótesis obteniendo el *p-valor*.

Los datos de las estaciones están en los ficheros:

Temperaturas.Medias.Anuales_Pontevedra.txt
Temperaturas.Medias.Anuales_Vigo.txt

Una columna con los años y otra con las temperaturas, con cabecera ***ano temp***. Primero leemos los dos ficheros y cogemos solo las columnas de las temperaturas (***temp***).

```
> datos1 <-read.table('Temperaturas.Medias.Anuales_Pontevedra.txt' , header = TRUE)
> datos2 <-read.table('Temperaturas.Medias.Anuales_Vigo.txt', header = TRUE)

> x <- datos1$temp
> y <- datos2$temp
```

3.2 Modelo de regresión lineal múltiple

El modelo de regresión múltiple es la extensión a k variables explicativas del modelo de regresión simple. En general, una variable de interés Y depende de varias variables $X_1, X_2, \ldots, X_k$ y no sólo de una única variable de predicción X. Por ejemplo, para estudiar la contaminación atmosférica, parece razonable considerar más de una variable explicativa, como pueden la temperatura media anual, el número de fábricas, el número de habitantes, etc. Además de las variables observables, la variable de interés puede depender de otras desconocidas para el investigador. Un modelo de regresión representa el efecto de estas variables en lo que se conoce como *error aleatorio o perturbación*.

Un modelo de regresión teórico en el que las variables se pueden relacionar mediante una función de *tipo lineal*, se puede expresar de la siguiente forma:

$$Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k + e$$

donde

- $\boldsymbol{Y}$ es la variable de interés que vamos a predecir, también llamada variable respuesta o variable dependiente
- $\boldsymbol{X_1, X_2, \ldots, X_k}$ son las variables independientes, explicativas o de predicción
- $\boldsymbol{\beta_1, \beta_2, \ldots, \beta_k}$ son los coeficientes de regresión que vamos a estimar
- $\boldsymbol{e}$ es el error aleatorio o perturbación, que representa el efecto de todas las variables que pueden afectar a la variable dependiente y no están incluidas en el modelo de regresión.

Como hacíamos en el caso de la regresión lineal simple, estimaremos los coeficientes de regresión con objeto de determinar el mejor hiperplano de regresión muestral de entre todos de la forma

$$y_t = \widehat{\beta_0} + \widehat{\beta_1}\, x_1 + \ldots + \widehat{\beta_k}\, x_k.$$

Los estimadores de los coeficientes de regresión serán los de mínimos cuadrados, es decir aquellos que hagan mínima la suma de cuadrados

$$\sum_{j=1}^{n} e_j^2 = \sum_{j=1}^{n} \left(y_j - \widehat{\beta_0} - \widehat{\beta_1}\, x_{1j} - \widehat{\beta_2}\, x_{2j} - \ldots - \widehat{\beta_k}\, x_{kj} \right)^2$$

Se llega a un sistema de ecuaciones que hay que resolver.

En *R* usaremos la misma función ***lm*** pero uniendo las variables independientes con el signo +.

Ejemplo 3.3

Se consideró que el número de admisiones previas de un paciente X_1, y su edad X_2, podrían servir para calcular la estancia del paciente en el hospital, en días Y, para ciertos enfermos crónicos.

Se tomó una muestra aleatoria simple de 15 pacientes:

X_1	0	0	0	1	1	1	1	2	2	2	3	3	4	4	5
X_2	21	18	22	24	25	25	26	34	25	38	44	51	39	54	55
Y	15	15	21	28	30	35	40	35	30	45	50	60	45	60	50

Se quiere analizar si alguna o ambas variables independientes X_1, X_2, pueden servir para explicar a la variable dependiente Y, estimando previamente los coeficientes de regresión de las variables significativas.

Vamos a calcular primero los coeficientes de regresión. Primero se introducen los datos en vectores.

```
> x1 <- c(0,0,0,1,1,1,1,2,2,2,3,3,4,4,5)
> x2 <- c(21,18,22,24,25,25,26,34,25,38,44,51,39,54,55)
> y <- c(15,15,21,28,30,35,40,35,30,45,50,60,45,60,50)
```

Hacemos el ajuste mediante ***lm***

```
> ajuste <- lm (y ~ x1 + x2)
> ajuste
```

```
Call:
lm(formula = y ~ x1 + x2)

Coefficients:
(Intercept)           x1           x2
    2.08572      0.05699      1.05002
```

Es decir, el hiperplano de regresión muestral es

$$y_t = 2'0857 + 0'057\, x_1 + 1'05\, x_2.$$

Contraste de regresión lineal múltiple

Como en el caso de la regresión lineal simple se puede contrastar la adecuación global del modelo o la igualdad a cero de los coeficientes de regresión.

Antes estos dos tests eran equivalentes porque solo había una covariable independiente, ahora no, y en este caso son mucho más interesantes los segundos porque permitirán decidir cuáles de las covariables X_i son significativas y cuales no en la explicación de la variable dependiente Y, de manera que se puedan descartar algunas de estas variables independientes no significativas. Haremos esto con la función ***summary***.

Siguiendo el ejemplo anterior

> summary (ajuste)

```
Call:
lm(formula = y ~ x1 + x2)

Residuals:
    Min      1Q  Median      3Q     Max
-10.122  -3.543   1.542   2.317  10.557

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.08572    6.73931   0.309  0.76226
x1           0.05699    2.61310   0.022  0.98296
x2           1.05002    0.32621   3.219  0.00737 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.059 on 12 degrees of freedom
Multiple R-squared:  0.8503,    Adjusted R-squared:  0.8254
F-statistic: 34.08 on 2 and 12 DF,  p-value: 1.125e-05
```

Si nos fijamos en la última columna del p-valor de los tests sobre cada coeficiente de regresión, llegamos a la conclusión que *puede aceptarse la hipótesis nula de ser cero el coeficiente de regresión X_1, debiendo eliminar esta variable del modelo*. El p-valor del test sobre el modelo en conjunto, (el último valor en la última línea 1.125e-05) indica que es aceptable.

Vamos a repetir el proceso considerando solo la variable X_2.

> ajuste2 <- lm (y ~ x2)

> summary (ajuste2)

```
Call:
lm(formula = y ~ x2)

Residuals:
    Min      1Q  Median      3Q     Max
-10.088  -3.561   1.534   2.345  10.552

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)    1.977      4.373   0.452    0.659
x2             1.057      0.123   8.593 1.01e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.821 on 13 degrees of freedom
Multiple R-squared:  0.8503,    Adjusted R-squared:  0.8388
F-statistic: 73.84 on 1 and 13 DF,  p-value: 1.014e-06
```

Se ve que el *p-valor* asociado al modelo en conjunto es muy pequeño con lo que se confirma que la Edad del paciente, X_2, *es significativa* (ahora aún mas) para explicar a la variable dependiente, estancia en días en el hospital.

La recta de regresión finalmente ajustada

$$y_t = 1{,}977 + 1{,}057x_2$$

permite predecir la estancia de los pacientes en el hospital en estudio.

3.3 Otros modelos de regresión

Aunque los modelos de regresión lineal (tanto simple como múltiple) funcionan bien en una amplia mayoría de situaciones, en ocasiones es necesario considerar modelos más complejos para conseguir un mejor ajuste a los datos.

Algunas veces cuando la relación entre las variables dependientes e independientes no es lineal, es útil incluir términos polinomiales para ayudar a explicar la variación de nuestra variable dependiente.

Un ejemplo de este tipo de modelos es *la regresión cuadrática* (polinomio de segundo grado). El modelo más sencillo de regresión cuadrática es el siguiente:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i}^2$$

Para ajustar un modelo de regresión cuadrático en R basta con indicar en el argumento ***formula*** de la función ***lm*** que una de las variables independientes está elevada al cuadrado mediante el símbolo ^2.

En el ejemplo anterior

```
> ajuste <- lm (y ~ x1 + x2^2)
> ajuste
Call:
lm(formula = y ~ x1 + x2^2)

Coefficients:
(Intercept)           x1           x2
    2.08572      0.05699      1.05002
```

Con lo que el modelo queda así:

$$y_t = 2{,}0857 + 0{,}057x_1 + 1{,}05\ x_2^2$$