Correlación lineal y Regresión lineal simple

Joaquín Amat Rodrigo *j.amatrodrigo@gmail.com*

Junio, 2016

Índice

Introducción

La información y ejemplos se han obtenido en su gran mayoría de los libros OpenIntro Statistics, Introduction to Statistical Learning, TheRBook y Handbook of Biological Statistics

La correlación lineal y la regresión lineal son métodos estadísticos que estudian la relación lineal existente entre dos variables. Antes de profundizar en cada uno de ellos conviene destacar algunas diferencias:

- La correlación consiste en cuantificar como de relacionadas están dos variables, mientras que la regresión lineal consiste en generar una ecuación (modelo) que, basándose en la relación existente entre ambas variables, permita predecir el valor de una a partir de la otra.
- El cálculo de correlación entre dos variables es independiente del orden o asignación de cada variable a X e Y, se mide únicamente la relación entre ambas sin considerar dependencias. En el caso de la regresión lineal, el modelo varía según qué variable se considere dependiente de la otra (lo cual no implica causa-efecto).
- A nivel experimental, la correlación se suele emplear cuando ninguna de las variables se ha controlado, simplemente se han medido ambas y se desea saber si están relacionadas. En el caso de estudios de regresión lineal, es más común que una de las variables se controle (tiempo, concentración de reactivo, temperatura…) y se mida la otra.
- Por norma general, los estudios de correlación lineal preceden a la generación de modelos de regresión lineal, primero se calcula si ambas variables están correlacionadas y, en caso de estarlo, se procede a generar el modelo de regresión.

Coeficientes de correlación

Para estudiar la relación lineal existente entre dos variables continuas es necesario disponer de parámetros que permitan cuantificar dicha relación. Uno de estos parámetros es la *covarianza*, que indica el grado de variación conjunta de dos variables aleatorias.

$$Covarianza\ muestral = Cov(X,Y) = \frac{\sum_{i=1}^{n}(x_i - \overline{x})\,(y_i - \overline{y})}{N-1}$$

La covarianza depende de las escalas en que se midan las variables estudiadas y por lo tanto no es comparable entre distintos pares de variables. La solución pasa por estandarizar la covarianza generando lo que se conoce como *coeficientes de correlación*. Existen diferentes tipos, de entre los que destacan el *coeficiente de Pearson*, *Rho de Spearman* y *Tau de Kendall*.

- Todos ellos varían entre +1 y -1. Siendo +1 una correlación positiva perfecta y -1 una correlación negativa perfecta.
- Se emplean como medida de fuerza de asociación (tamaño del efecto):
  - 0: asociación nula.
  - 0.1: asociación pequeña.
  - 0.3: asociación mediana.
  - 0.5: asociación moderada.
  - 0.7: asociación alta.
  - 0.9: asociación muy alta.

Las principales diferencias entre estos tres coeficientes de asociación son:

- La correlación de *Pearson* funciona bien con variables cuantitativas que sigan una distribución normal. *En el libro Handbook of Biological Statatistics se menciona que sigue siendo bastante robusto a pesar de la falta de normalidad*. Es más sensible a los valores extremos que las otras dos alternativas.
- La correlación de *Spearman* se emplea cuando los datos son ordinales, de intervalo, o bien cuando no se satisface la condición de normalidad para variables continuas y los datos se pueden transformar a rangos. Es un método no paramétrico.
- La correlación de *Kendall* es otra alternativa no paramétrica para el estudio de la correlación que trabaja con rangos. Se emplea cuando se dispone de pocos datos y muchos de ellos están en el mismo nivel, es decir, cuando hay muchas ligaduras.

Además del valor obtenido para el coeficiente de correlación, es necesario calcular su significancia. Solo si el *p-value* es significativo se puede aceptar que existe correlación, y esta será de la magnitud que indique el coeficiente. Por muy cercano que sea el valor del coeficiente de correlación a +1 o -1, si no es significativo se ha de interpretar que la correlación de ambas variables es 0, ya que el valor observado puede deberse a simple al aleatoriedad.

El test paramétrico de significancia estadística empleado para el coeficiente de correlación es el *t-test*. Al igual que ocurre siempre que se trabaja con muestras, por un lado está el parámetro estimado (en este caso el coeficiente de correlación) y por otro su significancia a la hora de considerar la población entera. Si se calcula el coeficiente de correlación entre diferentes muestras de X e Y de la misma población, el valor va a variar dependiendo de las

muestras utilizadas, de ahí que se tenga que calcular la significancia de la correlación obtenida y que también sea recomendable calcular su intervalo de confianza.

$$t = \frac{r\sqrt{N-2}}{\sqrt{1-r^2}}, df = N-2$$

Para este test de hipótesis la H_0 considera que las variables son independientes (coeficiente de correlación poblacional = 0) mientras que la H_a considera que existe relación (coeficiente de correlación poblacional ≠ 0)

La correlación entre dos variables, además del valor del coeficiente de correlación y de sus significancia, también tiene un tamaño de efecto asociado. Se conoce como *coeficiente de determinación* R^2. Se interpreta como la cantidad de varianza de *Y* explicada por *X*. En el caso del coeficiente de *Pearson* y el de *Spearman*, R^2 se obtiene elevando al cuadrado el coeficiente de correlación. En el caso de Kendall no se puede calcular de este modo. *(No he encontrado como se calcula)*

Mediante *bootstrapping* también se puede calcular la significancia de un coeficiente de correlación. Es una alternativa no paramétrica al t-test. *(Ver capítulo dedicado al resampling)*.

Coeficiente de Pearson

El coeficiente de correlación de Pearson es la covarianza estandarizada, y su ecuación difiere dependiendo de si se aplica a una muestra *Coeficiente de Pearson muestral (r)* o si se aplica la población *Coeficiente de Pearson poblacional (ρ)*.

$$\rho = \frac{Cov(X,Y)}{\sigma_x \sigma_y}$$

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})\,(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2 \sum_{i=1}^{n}(y_i - \overline{y})^2}}$$

Condiciones

- La relación que se quiere estudiar entre ambas variables es lineal (de lo contrario, el coeficiente de Pearson no la puede detectar).
- Las dos variables deben de ser cuantitativas.
- Normalidad: ambas variables se tienen que distribuir de forma normal. *Varios textos defienden su robustez cuando las variables se alejan moderadamente de la normal.*
- Homocedasticidad: La varianza de *Y* debe ser constante a lo largo de la variable X. Esto se puede identificar si en el *scatterplot* los puntos mantienen la misma dispersión en las distintas zonas de la variable X. *Esta condición no la he encontrado mencionada en todos los libros.*

Características

- Toma valores entre [-1, +1], siendo +1 una correlación lineal positiva perfecta y -1 una correlación lineal negativa perfecta.
- Es una medida independiente de las escalas en las que se midan las variables.
- No varía si se aplican transformaciones a las variables.
- No tiene en consideración que las variables sean dependientes o independientes.
- El coeficiente de correlación de Pearson no equivale a la pendiente de la recta de regresión.
- Es sensible a *outliers*, por lo que se recomienda en caso de poder justificarlos, excluirlos del análisis.

Interpretación

Además del valor obtenido para el coeficiente, es necesario calcular su significancia. Solo si el *p-value* es significativo se puede aceptar que existe correlación y esta será de la magnitud que indique el coeficiente. Por muy cercano que sea el valor del coeficiente de correlación a +1 o -1, si no es significativo, se ha de interpretar que la correlación de ambas variables es 0 ya que el valor observado se puede deber al azar. (Ver más adelante como calcular la significancia).

Coeficiente de Spearman (Spearman's rho)

El coeficiente de *Spearman* es el equivalente al coeficiente de *Pearson* pero con una previa transformación de los datos a rangos. Se emplea como alternativa cuando los valores son ordinales o bien cuando los valores son continuos pero no satisfacen la condición de

normalidad requerida por el coeficiente de Pearson y se pueden ordenar transformándolos en rangos. Al trabajar con rangos, es menos sensible que *Pearson* a valores extremos. Existe una diferencia adicional con respecto a *Pearson*. El coeficiente de Spearman requiere que la relación entre las variables sea monótona es decir, que cuando una variable crece la otra también lo hace, o cuando una crece la otra decrece (que la tendencia sea constante). Este concepto no es exactamente el mismo que linealidad.

$$r_s = 1 - \frac{6\sum d_i^2}{n(n^2 - 1)}$$

Siendo d_i la distancia entre los rangos de cada observación $(x_i - y_i)$ y n el número de observaciones.

Coeficiente Tau de Kendall

Trabaja con rangos por lo que requiere que las variables cuya relación se quiere estudiar sean ordinales o que se puedan transformar en rangos. Al ser no paramétrico, es otra alternativa al *Coeficiente de correlación de Pearson* cuando no se cumple la condición de normalidad. Parece ser más aconsejable que el coeficiente de *Spearman* cuando el número de observaciones es pequeño o los valores se acumulan en una región por lo que el número de ligaduras al generar los rangos es alto.

$$\tau = \frac{C - D}{\frac{1}{2}n(n-1)}$$

Siendo *C* es el número de pares concordantes, aquellos en los que el rango de la segunda variable es mayor que el rango de la primera variable. *D* el número de pares discordantes, cuando el rango de la segunda es igual o menor que el rango de la variable primera.

Tau represents a probability; that is, it is the difference between the probability that the two variables are in the same order in the observed data versus the probability that the two variables are in different orders.

Jackknife correlation

El coeficiente de correlación de Pearson resulta efectivo en ámbitos muy diversos, sin embargo, tiene la desventaja de no ser robusto frente a *outliers* a pesar de que se satisface la condición de normalidad. Si dos variables tienen un pico o un valle común en una única observación, por ejemplo por un error de lectura, la correlación va a estar dominada por este registro a pesar de que entre las dos variables no haya correlación real alguna. Lo mismo puede ocurrir en la dirección opuesta. Si dos variables están altamente correlacionadas excepto para una observación para la que los valores son muy dispares, entonces la correlación existente quedará enmascarada. Una forma de evitarlo es recurrir a la *Jackknife correlation*, que consiste en calcular todos los posibles coeficientes de correlación entre dos variables si se excluye cada vez una de las observaciones. El promedio de todas las *Jackknife correlations* calculadas atenuará en cierta medida el efecto del *outlier*.

$$\overline{\theta}_{(A,B)} = Promedio\ Jackknife\ correlation(A,B) = \frac{1}{n}\sum_{i=1}^{n}\hat{r}_i$$

Donde n es el número de observaciones y $\hat{r}_i$ es el coeficiente de correlación de Pearson estimado entre las variables A y B, habiendo excluido la observación i.

Además del promedio, se puede estimar su error estándar (SE) y así obtener intervalos de confianza para la *Jackknife correlation* y su correspondiente *p-value*.

$$SE = \sqrt{\frac{n-1}{n}\sum_{i=1}^{n}(\hat{r}_i - \overline{\theta})^2}$$

Intervalo de confianza del 95% ($Z = 1.96$)

$$promedio\ Jackknife\ correlation(A,B)\ \pm\ 1.96 * SE$$

$$\overline{\theta} - 1.96\sqrt{\frac{n-1}{n}\sum_{i=1}^{n}(\hat{r}_i - \overline{\theta})^2}\ ,\overline{\theta} + 1.96\sqrt{\frac{n-1}{n}\sum_{i=1}^{n}(\hat{r}_i - \overline{\theta})^2}$$

P-value para la hipótesis nula de que $\overline{\theta} = 0$:

$$Z_{calculada} = \frac{\overline{\theta} - H_0}{SE} = \frac{\overline{\theta} - 0}{\sqrt{\frac{n-1}{n}\sum_{i=1}^{n}(\hat{r}_i - \overline{\theta})^2}}$$

$$p_{value} = P(Z > Z_{calculada})$$

Cuando se emplea este método es conveniente calcular la diferencia entre el valor de correlación obtenido por *Jackknife correlation* ($\overline{\theta}$) y el que se obtiene si se emplean todas las observaciones ($\overline{r}$). Ha esta diferencia se le conoce como *Bias*. Su magnitud es un indicativo de cuanto está influenciada la estimación de la correlación entre dos variables debido a un valor atípico o *outlier*.

$$Bias = (n-1) * (\overline{\theta} - \hat{r})$$

Si se calcula la diferencia entre cada correlación ($\hat{r}_i$) estimada en el proceso de *Jackknife* y el valor de correlación ($\hat{r}$) obtenido si se emplean todas las observaciones, se puede identificar que observaciones son más influyentes.

Cuando el estudio requiere minimizar al máximo la presencia de falsos positivos, a pesar de que se incremente la de falsos negativos se puede seleccionar como valor de correlación el menor de entre todos los calculados en el proceso de *Jackknife*.

$$Correlacion = min\{\hat{r}_1, \hat{r}_2, \ldots, \hat{r}_n\}$$

A pesar de que el método de *Jackknife* permite aumentar la robustez de la correlación de Pearson, si los *outliers* son muy extremos su influencia seguirá siendo notable. Siempre es conveniente una representación gráfica de los datos para poder identificar si hay valores atípicos y eliminarlos. Otras alternativas robustas son la correlación de *Spearman* o el método de *Bootsrapping*.

Ejemplo correlación lineal

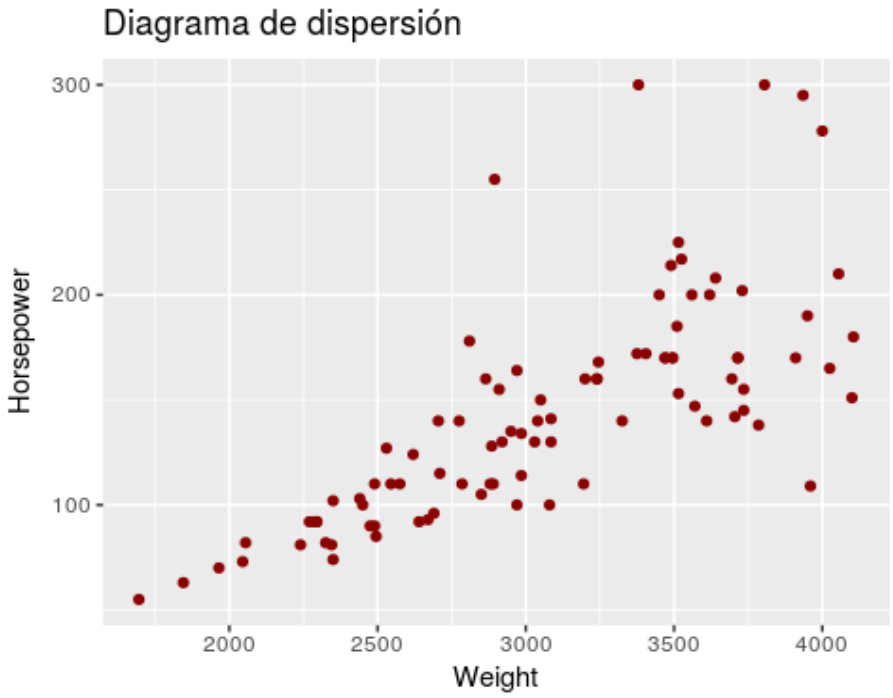
Se dispone de un data set con información sobre diferentes coches. Se quiere estudiar si existe una correlación entre el peso de un vehículo (Weight) y la potencia de su motor (Horsepower).

R contiene funciones que permiten calcular los diferentes tipos de correlaciones y sus niveles de significancia: `cor()` y `cor.test()`. La segunda función es más completa ya que además de calcular el coeficiente de correlación indica su significancia (*p-value*) e intervalo de confianza.

```
require(MASS)
require(ggplot2)
data("Cars93")
```

En primer lugar se representan las dos variables mediante un diagrama de dispersión (*scatterplot*) para intuir si existe relación lineal o monotónica. Si no la hay, no tiene sentido calcular este tipo de correlaciones.

```
ggplot(data = Cars93, aes(x = Weight, y = Horsepower)) +
 geom_point(colour = "red4") +  ggtitle("Diagrama de dispersión")
```

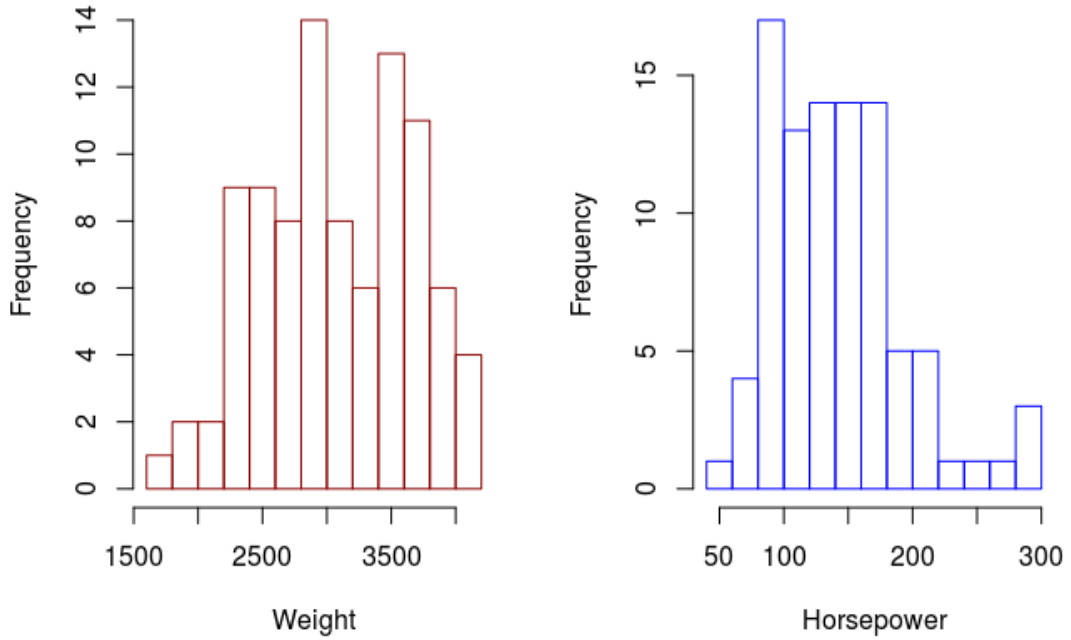


El diagrama de dispersión parece indicar una posible relación lineal positiva entre ambas variables.

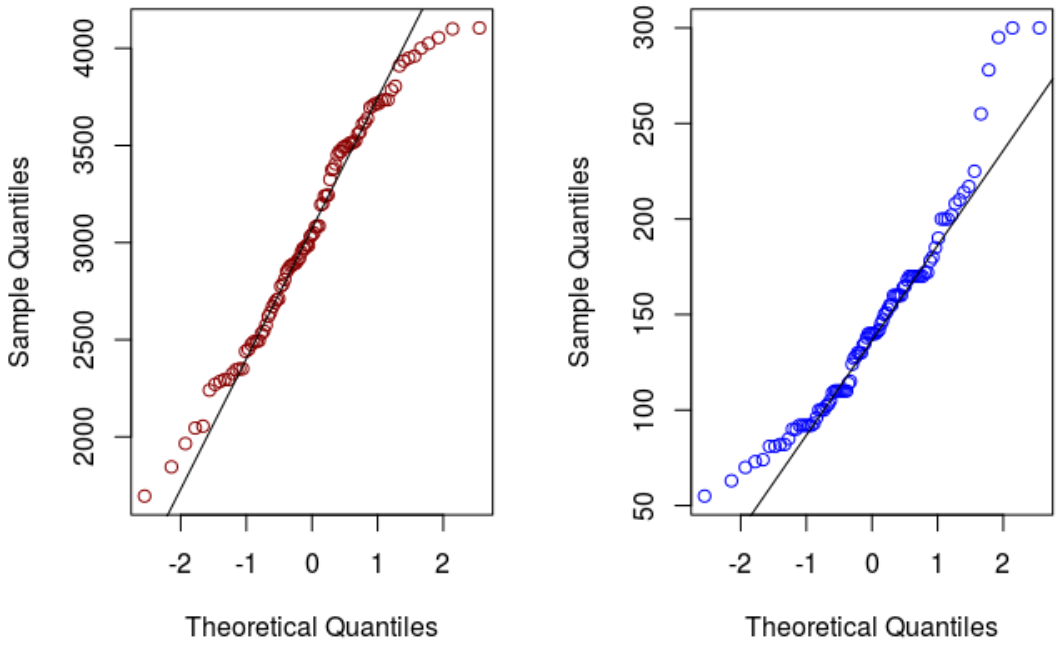
Para poder elegir el coeficiente de correlación adecuado, se tiene que analizar el tipo de variables y la distribución que presentan. En este caso, ambas variables son cuantitativas continuas y pueden transformarse en rangos para ordenarlas, por lo que *a priori* los tres coeficientes podrían aplicarse. La elección se hará en función de la distribución que presenten las observaciones.

1.Análisis de normalidad

```
# representación gráfica
par(mfrow = c(1, 2))
hist(Cars93$Weight, breaks = 10, main = "", xlab = "Weight", border = "darkred")
hist(Cars93$Horsepower, breaks = 10, main = "", xlab = "Horsepower", border =
"blue")
```




```
qqnorm(Cars93$Weight, main = "", col = "darkred")
qqline(Cars93$Weight)
qqnorm(Cars93$Horsepower, main = "", col = "blue")
qqline(Cars93$Horsepower)
```




```
# Test de hipótesis para el análisis de normalidad
shapiro.test(Cars93$Weight)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  Cars93$Weight
## W = 0.97432, p-value = 0.06337
```

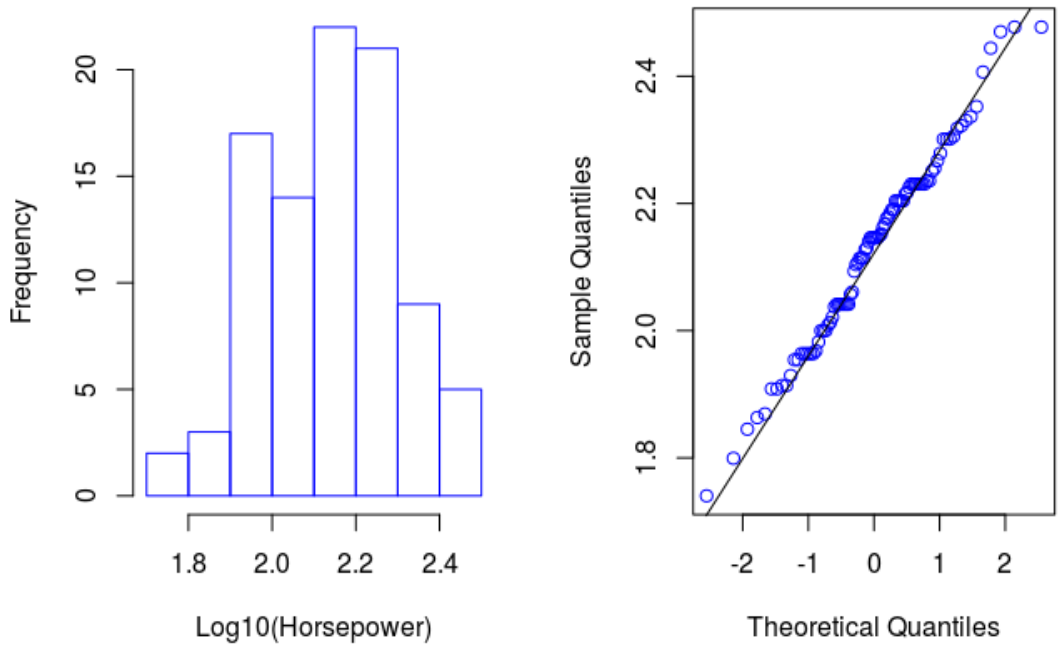
```
shapiro.test(Cars93$Horsepower)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  Cars93$Horsepower
## W = 0.93581, p-value = 0.0001916
```

El análisis gráfico y el contraste de normalidad muestran que para la variable *Horsepower* no se puede asumir normalidad y que la variable *Weight* está en el límite. Siendo estrictos, este hecho excluye la posibilidad de utilizar el *coeficiente de Pearson*, dejando como alternativas el de *Spearman* o *Kendall*. Sin embargo, dado que la distribución no se aleja

mucho de la normalidad y de que el coeficiente *de Pearson* tiene cierta robustez, a fines prácticos sí que se podría utilizar siempre y cuando se tenga en cuenta este hecho en los resultados. Otra posibilidad es tratar de transformar las variables para mejorar su distribución.

```
# representación gráfica
par(mfrow = c(1, 2))
hist(log10(Cars93$Horsepower), breaks = 10, main = "", xlab = "Log10(Horsepower)",
    border = "blue")
qqnorm(log10(Cars93$Horsepower), main = "", col = "blue")
qqline(log10(Cars93$Horsepower))
```




```
par(mfrow = c(1, 1))
shapiro.test(log10(Cars93$Horsepower))
```

```
##  Shapiro-Wilk normality test
##
## data:  log10(Cars93$Horsepower)
## W = 0.98761, p-value = 0.5333
```

La trasformación logarítmica de la variable *horsepower* consigue una distribución de tipo normal.

2.Homocedasticidad

La homocedasticidad implica que la varianza se mantenga constante. Puede analizarse de forma gráfica representando las observaciones en un diagrama de dispersión y viendo si mantiene una homogeneidad en su dispersión a lo largo del eje X. Una forma cónica es un claro indicativo de falta de homocedasticidad. *En algunos libros se menciona el test de Goldfeld-Quandt o el de Breusch-Pagan como test de hipótesis para la homocedasticidad en correlación y regresión.*

Tal como muestra el diagrama de dispersión generado al inicio del ejercicio, sí hay un patrón cónico. Esto debe de tenerse en cuenta si se utiliza *Pearson* puesto que viola una de sus condiciones.

3.Cálculo de correlación

Debido a la falta de homocedasticidad, los resultados generados por *Pearson* no son precisos, desde el punto de vista teórico *Spearman* o *Kendall* son más adecuados. Sin embargo, en la bibliografía emplean Pearson, así que se van a calcular tanto *Pearson* como *Spearman*.

```
cor(x = Cars93$Weight, y = log10(Cars93$Horsepower), method = "pearson")
```

```
## [1] 0.809672
```

```
cor(x = Cars93$Weight, y = log10(Cars93$Horsepower), method = "spearman")

# La función cor() también acepta matrices o data frames y calcula todas las
# correlaciones dos a dos.
```

```
## [1] 0.8042527
```

Ambos test muestran una correlación alta (>0.8). Sin embargo para poder considerar que existe realmente correlación entre las dos variables es necesario calcular su significancia, de lo contrario podría deberse al azar.

4.Significancia de la correlación

Por muy alto que sea un coeficiente de correlación, si no es significativa se ha de considerar inexistente.

```
cor.test(x = Cars93$Weight, y = log10(Cars93$Horsepower), alternative =
"two.sided", conf.level = 0.95, method = "pearson")
```

```
##
##  Pearson's product-moment correlation
##
## data:  Cars93$Weight and log10(Cars93$Horsepower)
## t = 13.161, df = 91, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.7256502 0.8699014
## sample estimates:
##      cor
## 0.809672
```

```
cor.test(x = Cars93$Weight, y = log10(Cars93$Horsepower), alternative =
"two.sided",
    conf.level = 0.95, method = "spearman")
```

```
## Warning in cor.test.default(x = Cars93$Weight, y =
## log10(Cars93$Horsepower), : Cannot compute exact p-value with ties
```

```
##
##  Spearman's rank correlation rho
##
## data:  Cars93$Weight and log10(Cars93$Horsepower)
## S = 26239, p-value < 2.2e-16
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##       rho
## 0.8042527
```

Ambos coeficientes de correlación son significativos.

5.Coeficiente de determinación R^2 (tamaño del efecto)

```
R2_pearson <- cor(x = Cars93$Weight, y = log10(Cars93$Horsepower), method =
"pearson")^2
R2_pearson
```

```
## [1] 0.6555688
```

```
R2_spearman <- cor(x = Cars93$Weight, y = log10(Cars93$Horsepower), method =
"spearman")^2
R2_spearman
```

```
## [1] 0.6468225
```

6.Conclusión

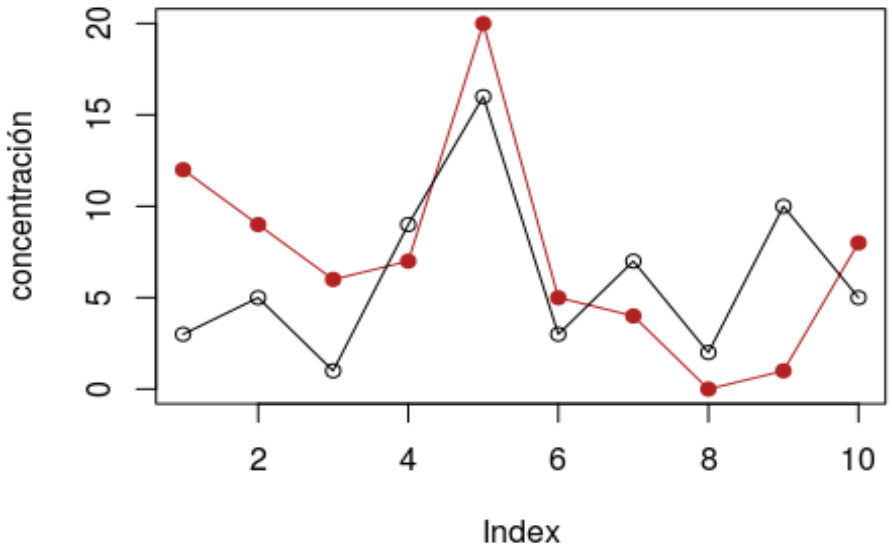
Existe una correlación significativa entre el peso del vehículo y la potencia de su motor (r=0.739, p-value < 2.2e-16).

Ejemplo Jackknife correlation

Un equipo de investigadores quiere estudiar si existe correlación en la presencia de dos sustancias (A y B) en el agua de los ríos. Para ello han realizado una serie de mediciones en las que se cuantifica la concentración de las dos sustancias en 10 muestras independientes de agua. Se sospecha que el instrumento de lectura sufre alguna avería que provoca que algunas lecturas se disparen, por esta razón se quiere emplear un método de correlación robusto. El objetivo de este ejemplo es ilustrar el método de Jackknife, por lo que se asume que se cumplen las condiciones para la correlación de Pearson.

```
# Datos simulados de dos variables A y B
a <- c(12, 9, 6, 7, 2, 5, 4, 0, 1, 8)
b <- c(3, 5, 1, 9, 5, 3, 7, 2, 10, 5)

# Se introduce un outlier
a[5] <- 20
b[5] <- 16
datos <- data.frame(a, b)
```

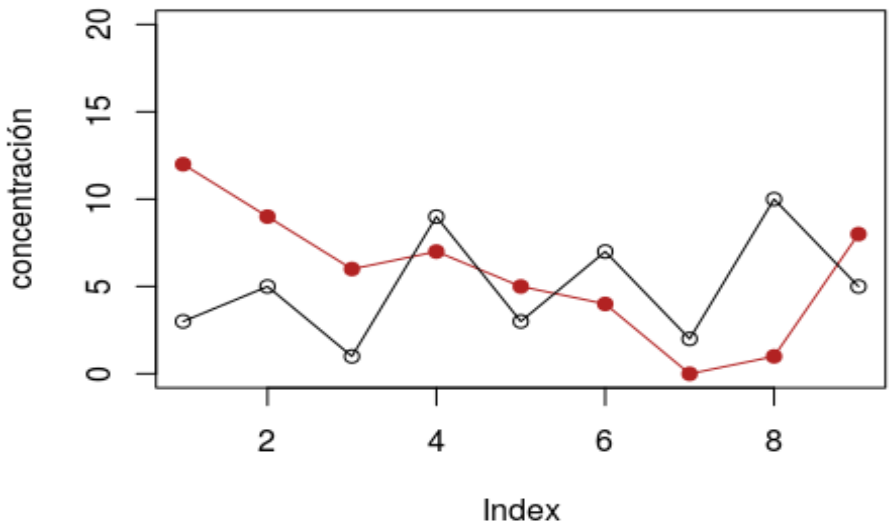
```r
plot(datos$a, type = "o", pch = 19, col = "firebrick", ylab = "concentración")
lines(datos$b, type = "o")
```




```r
cor(datos$a, datos$b, method = "pearson")
```

```
## [1] 0.5249277
```

```r
# Se elimina el outlier
a <- a[-5]
b <- b[-5]
datos_sin_outlier <- data.frame(a, b)
plot(datos_sin_outlier$a, type = "o", pch = 19, col = "firebrick",
 ylim = c(0, 20), ylab = "concentración")
lines(datos_sin_outlier$b, type = "o")
```



```r
cor(datos_sin_outlier$a, datos_sin_outlier$b, method = "pearson")
```

```
## [1] -0.1790631
```

La observación numero 5 tiene una gran influencia en el resultado de la correlación, siendo de 0.52 en su presencia y de -0.18 si se excluye.

```r
# FUNCIÓN PARA APLICAR JACKKNIFE A LA CORRELACIÓN DE PEARSON

correlacion_jackknife <- function(matriz, method = "pearson") {
    n <- nrow(matriz)  #   número de observaciones
    valores_jackknife <- rep(NA, n)

    for (i in 1:n) {
        # Loop para excluir cada observación y calcular la correlación
        valores_jackknife[i] <- cor(matriz[-i, 1], matriz[-i, 2], method = method)
    }

    promedio_jackknife <- mean(valores_jackknife)
    standar_error <- sqrt(((n - 1)/n) * sum((valores_jackknife -
                                             promedio_jackknife)^2))
    bias <- (n - 1) * (promedio_jackknife - cor(matriz[, 1], matriz[, 2],
                                                method = method))
    return(list(valores_jackknife = valores_jackknife,
                promedio = promedio_jackknife,
                se = standar_error, bias = bias))
}

correlacion <- correlacion_jackknife(datos)
correlacion$promedio
```

```
## [1] 0.4854695
```

```r
correlacion$valores_jackknife
```

```
##  [1]  0.6409823  0.5394608  0.5410177  0.5414076 -0.1790631  0.5121559
##  [7]  0.5504217  0.4528914  0.7237978  0.5316224
```

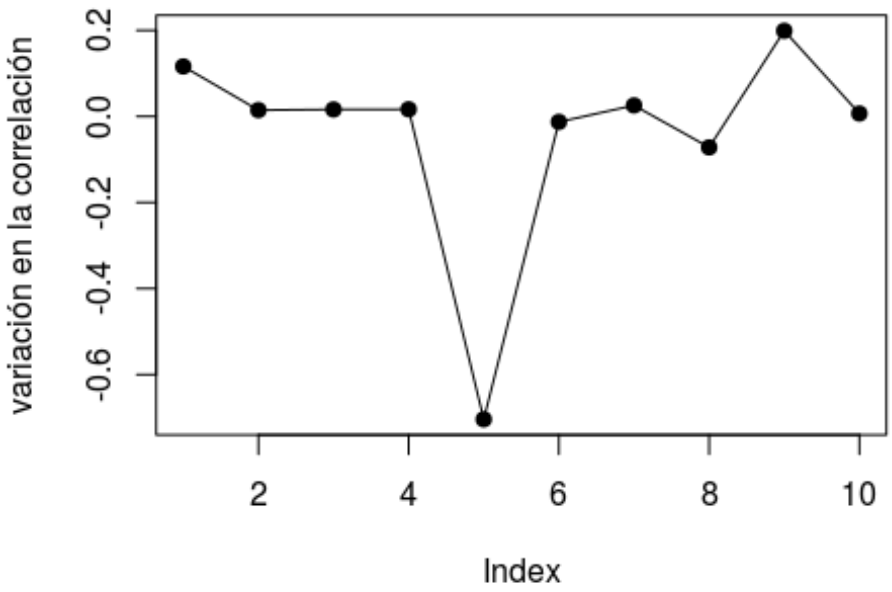
```r
correlacion$se
```

```
## [1] 0.697034
```

```r
correlacion$bias
```

```
## [1] -0.3551246
```

```
plot((correlacion$valores_jackknife - cor(datos$a, datos$b, method = "pearson")),
     type = "o", pch = 19, ylab = "variación en la correlación")
```



El método *Jackknife correlation* solo ha sido capaz de amortiguar una pequeña parte de la influencia del *outlier*, sin embargo, si ha permitido identificar que observación está afectando en mayor medida.

Matriz de correlaciones

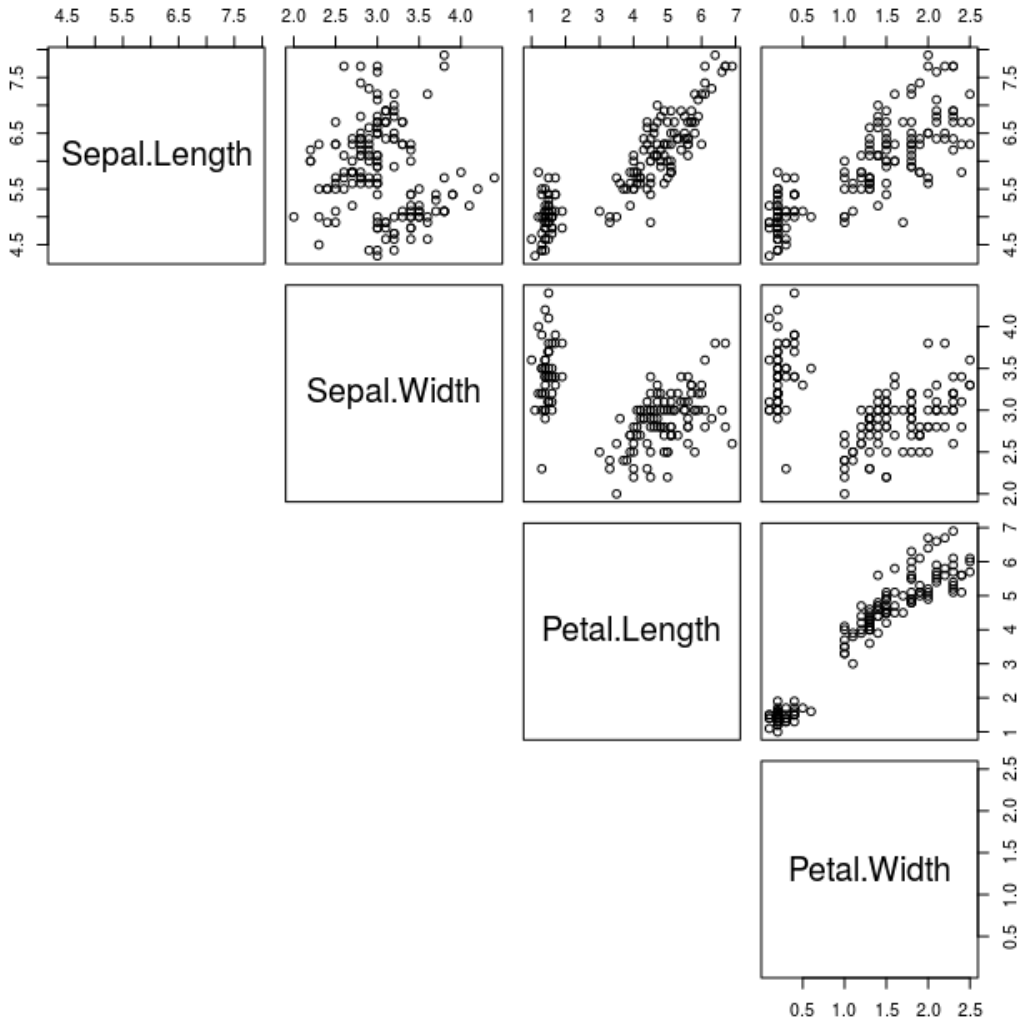
Cuando se dispone de múltiples variables y se quiere estudiar la relación entre todas ellas se recurre al cálculo de matrices con el coeficiente de correlación de cada par de variables. También se generan gráficos de dispersión dos a dos. En R existen diferentes funciones que permiten realizar este estudio, las diferencias entre ellas son el modo en que se representan gráficamente los resultados.

Se quiere estudiar la relación entre el tamaño de diferentes elementos de las flores. Para ello, se dispone del set de datos iris que consta de cuatro variables numéricas.

```r
data(iris)
# Se seleccionan únicamente las variables numéricas
datos <- iris[, c(1, 2, 3, 4)]
head(datos)
```

```
##   Sepal.Length Sepal.Width Petal.Length Petal.Width
## 1          5.1         3.5          1.4         0.2
## 2          4.9         3.0          1.4         0.2
## 3          4.7         3.2          1.3         0.2
## 4          4.6         3.1          1.5         0.2
## 5          5.0         3.6          1.4         0.2
## 6          5.4         3.9          1.7         0.4
```

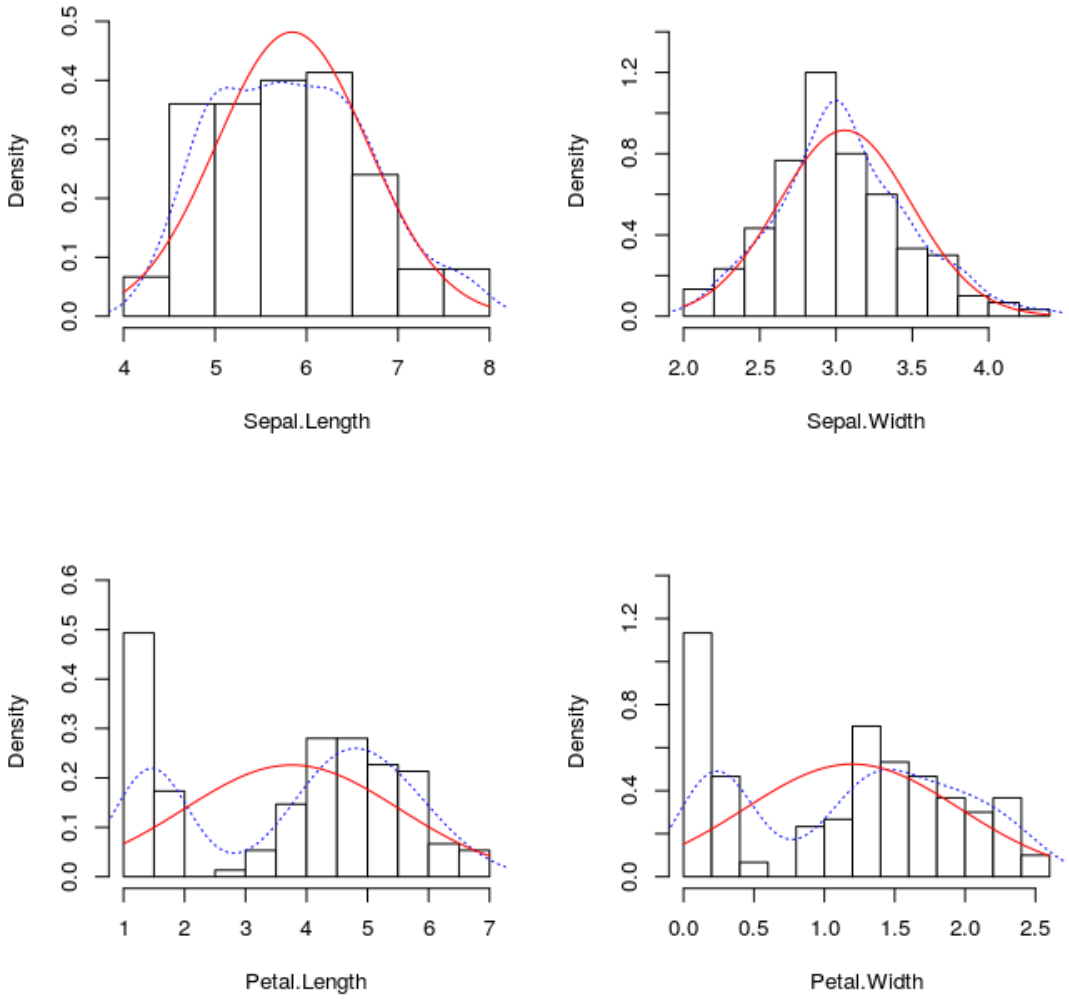
```r
pairs(x = datos, lower.panel = NULL)
# No se muestra la diagonal inferior ya que es lo mismo que la superior
```



```r
cor(x = datos, method = "pearson")
```

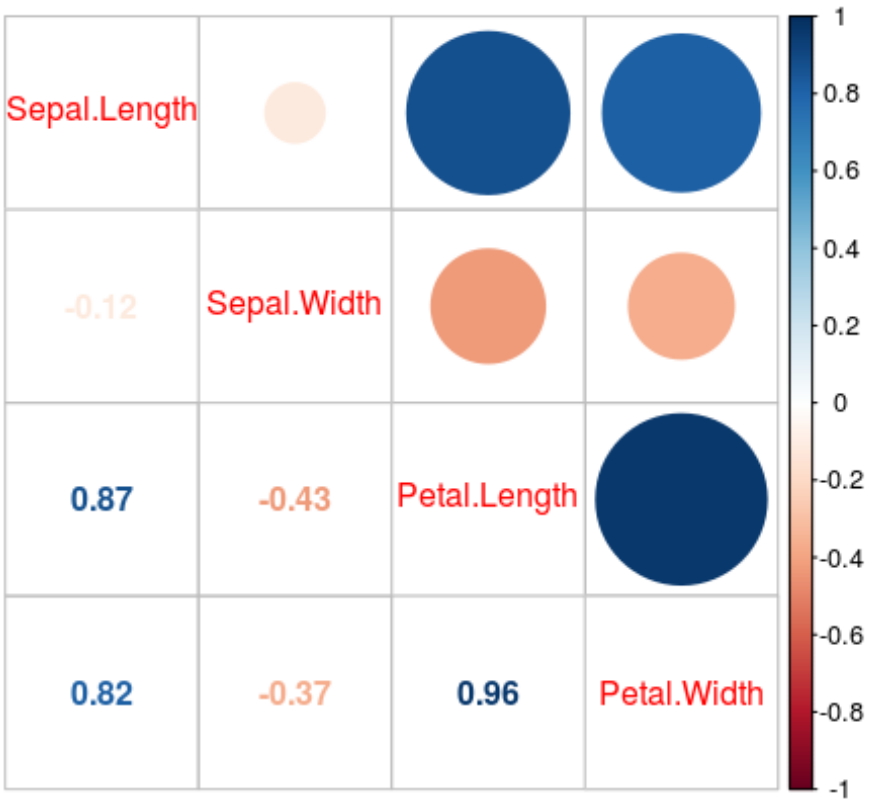
```
##              Sepal.Length Sepal.Width Petal.Length Petal.Width
## Sepal.Length    1.0000000  -0.1175698    0.8717538   0.8179411
## Sepal.Width    -0.1175698   1.0000000   -0.4284401  -0.3661259
## Petal.Length    0.8717538  -0.4284401    1.0000000   0.9628654
## Petal.Width     0.8179411  -0.3661259    0.9628654   1.0000000
```

```r
require(psych)
multi.hist(x = datos, dcol = c("blue", "red"), dlty = c("dotted", "solid"),
    main = "")
```



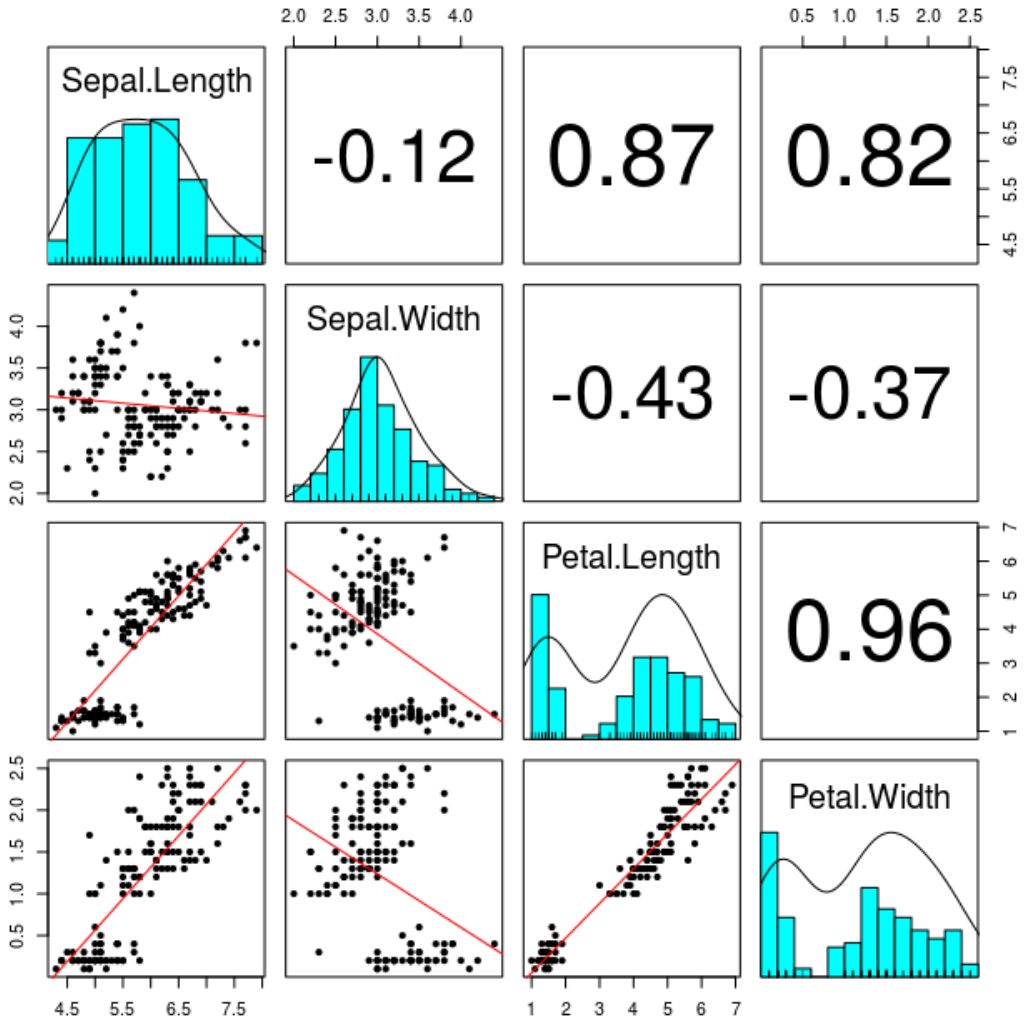
La función `corrplot()` del paquete *corrplot* recibe como argumento la matriz de correlaciones generada por la función `cor()` y genera diferentes tipos de *heat maps* mucho más visuales que la matriz numérica.

```
require(corrplot)
corrplot.mixed(corr = cor(x = datos, method = "pearson"))
```



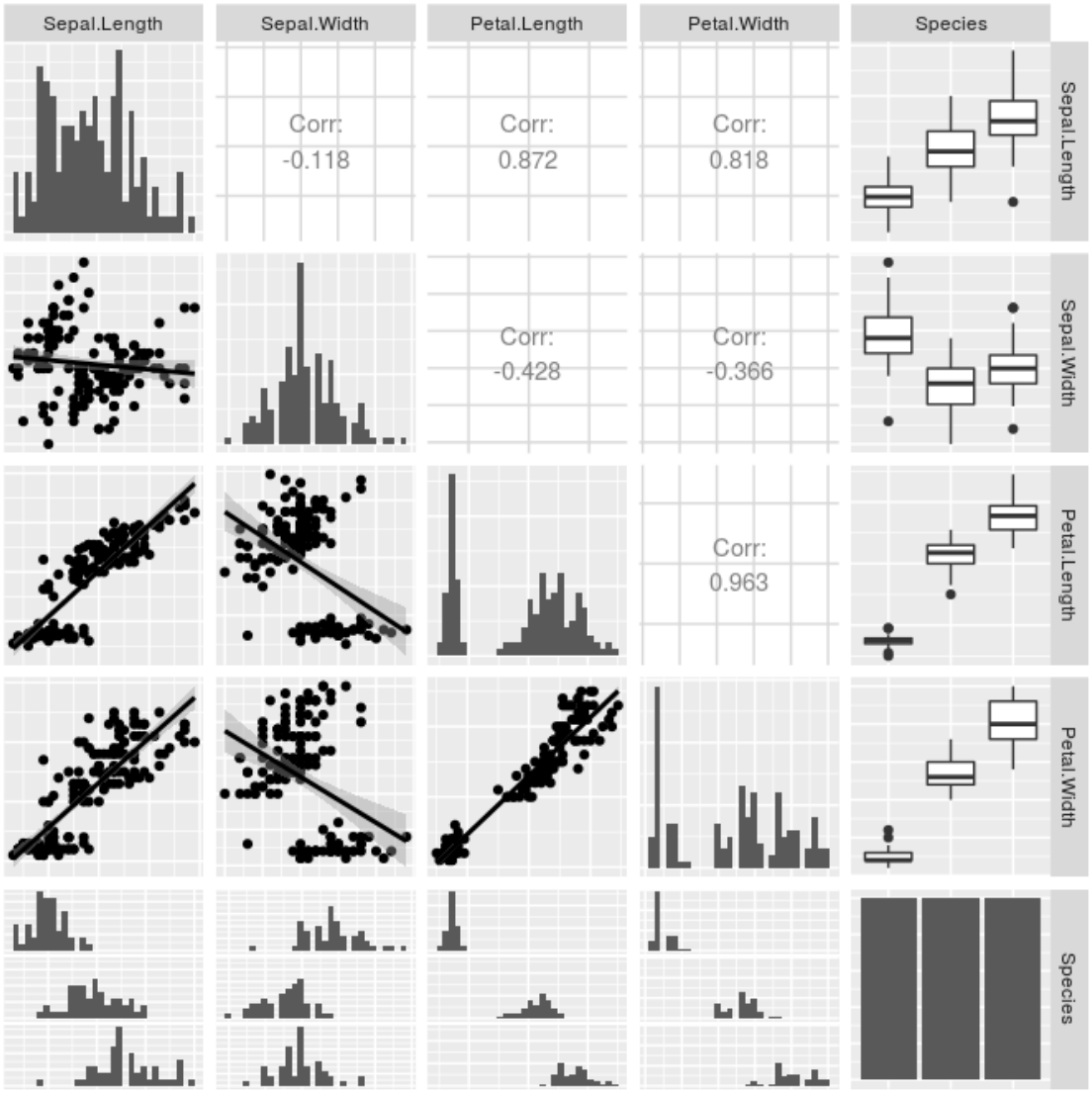
Otros paquetes permiten representar a la vez los diagramas de dispersión y los valores de correlación para cada par de variables. Además de la distribución de cada una de las variables.

```
require(psych)
pairs.panels(x = datos, ellipses = FALSE, lm = TRUE, method = "pearson")
```



La función `ggpairs()` del paquete *GGally* basada en *ggplot2* representa los diagramas de dispersión, el valor de la correlación e incluso interpreta el tipo de variable para que, en caso de ser categórica, representarla en forma de *boxplot*.

```
require(GGally)
ggpairs(iris, lower = list(continuous = "smooth"), diag = list(continuous = "bar"),
    axisLabels = "none")
```



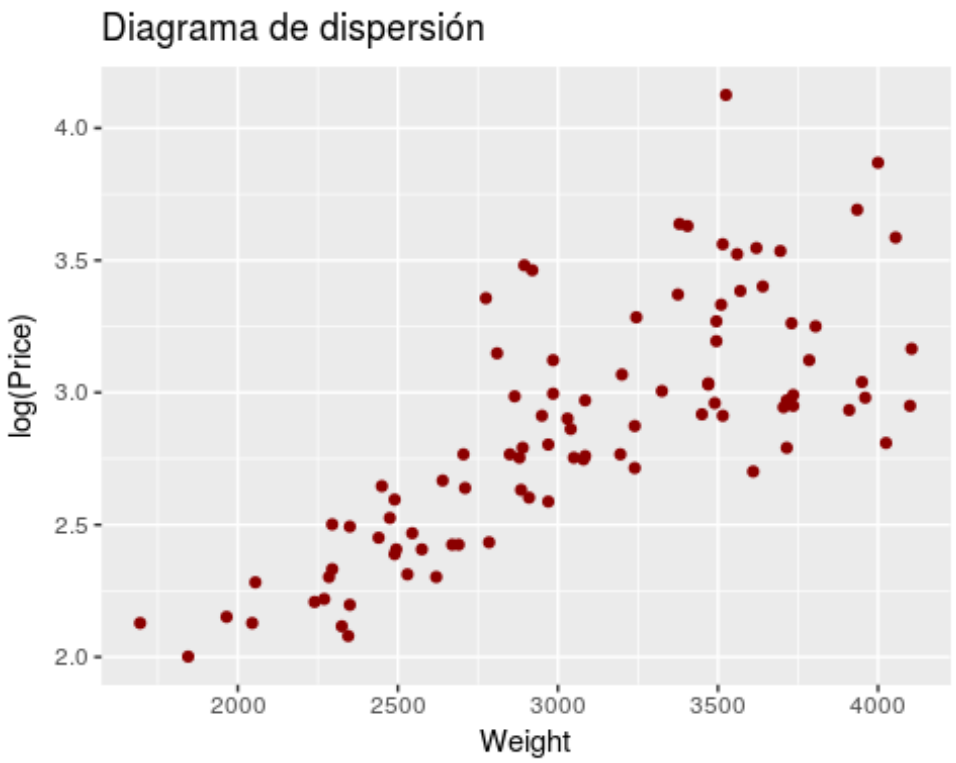
Correlación parcial

Como se ha explicado, la correlación estudia la relación (lineal o monotónica) existente entre dos variables. Puede ocurrir que la relación que muestran dos variables se deba a una tercera variable que influye sobre las otras dos, a este fenómeno se le conoce como *confounding*. Por ejemplo, si se correlaciona el tamaño del pie de una persona con su inteligencia, se encuentra una correlación positiva alta. Sin embargo, dicha relación se debe a una tercera variable que está relacionada con las otras dos, la edad. La correlación parcial permite estudiar la relación lineal entre dos variables bloqueando el efecto de una tercera (o más) variables. Si el valor de correlación de dos variables es distinto al valor de correlación parcial de esas mismas dos variables cuando se controla una tercera, significa que la tercera variable influye en las otras dos. La función en `pcor.test()` del paquete *ppcor* permite estudiar correlaciones parciales.

Ejemplo

Se quiere estudiar la relación entre las variables precio y peso de los automóviles. Se sospecha que esta relación podría estar influenciada por la variable potencia del motor, ya que a mayor peso del vehículo se requiere mayor potencia y, a su vez, motores más potentes son más caros.

```
require(MASS)
require(ggplot2)
data("Cars93")
# Se emplea el log del precio por que mejora la linealidad
ggplot(data = Cars93, aes(x = Weight, y = log(Price))) +
 geom_point(colour = "red4") +
 ggtitle("Diagrama de dispersión")
```



El gráfico permite intuir que existe una relación lineal entre el peso de un coche y el logaritmo de su precio.

```
require(ppcor)
cor.test(x = Cars93$Weight, y = log(Cars93$Price), method = "pearson")
```

```
##
##  Pearson's product-moment correlation
##
## data:  Cars93$Weight and log(Cars93$Price)
## t = 11.279, df = 91, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.6629513 0.8370563
## sample estimates:
##      cor
## 0.763544
```

```
pcor.test(x = Cars93$Weight, y = log(Cars93$Price), z = Cars93$Horsepower, method =
"pearson")
```

```
##    estimate      p.value statistic  n gp  Method
## 1 0.4047414 6.288649e-05  4.199019 93  1 pearson
```

La correlación entre el peso y el logaritmo del precio es alta (r=0.764) y significativa (p-value < 2.2e-16). Sin embargo, cuando se estudia su relación bloqueando la variable potencia de motor, a pesar de que la relación sigue siendo significativa (p-value = 6.288649e-05) pasa a ser baja (r=0.4047).

Conclusión

La relación lineal existente entre el peso y el logaritmo del precio está influenciada por el efecto de la variable potencia de motor. Si se controla el efecto de la potencia, la relación lineal existente es baja (r=0.4047).

Regresión lineal simple

La información aquí presente recoge los principales conceptos de la regresión lineal. Se puede encontrar una descripción mucho más detallada en el libro Introduction to Statistical Learning.

La regresión lineal simple consiste en generar un modelo de regresión (ecuación de una recta) que permita explicar la relación lineal que existe entre dos variables. A la variable dependiente o respuesta se le conoce como *y* y a la variable predictora o independiente como *x*.

El modelo de regresión lineal simple se describe de acuerdo a la ecuación:

$$Y = \beta_0 + \beta_1 X_1 + \epsilon$$

Siendo β_0 la ordenada en el origen, β_1 la pendiente y ϵ el error aleatorio. Este último representa la diferencia entre el valor ajustado por la recta y el valor real, recoge el efecto de todas aquellas variables que influyen en x pero que no se emplean en el modelo como predictores. Al error aleatorio también se le conoce como residuo.

En la gran mayoría de casos, los valores β_0 y β_1 poblacionales son desconocidos, por lo que a partir de una muestra se obtienen sus estimaciones $\hat{\beta}_0$ y $\hat{\beta}_1$. Estas estimaciones se conocen como *least square coefficient estimates* ya que toman aquellos valores que minimizan la suma de cuadrados residuales, dando lugar a la recta que pasa más cerca de todos los puntos.

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

$$\hat{\beta}_1 = \frac{Sy}{Sx} R$$

$$\hat{\beta}_0 = \hat{y} - \hat{\beta}_1 \hat{x}$$

Donde *Sy y Sx son las desviaciones típicas de cada variable y R el coeficiente de correlación.*

$\hat{\beta}_0$ es el valor que tiene la variable y para x = 0, es decir, la intersección de la recta con el eje y. Es un dato necesario para generar la recta pero en muchas ocasiones no tiene interpretación práctica, ya que hay situaciones en las que x no puede adquirir el valor 0.

Una recta de regresión puede emplearse para diferentes propósitos y dependiendo de ellos es necesario satisfacer distintas condiciones. En caso de querer medir la relación lineal entre dos variables, la recta de regresión lo va a indicar de forma directa (ya que calcula la correlación). Sin embargo, en caso de querer predecir el valor de una variable en función de la otra, no solo se necesita calcular la recta, sino que además hay que asegurar que el modelo sea bueno.

Inferencia mediante regresión lineal. Intervalo de confianza para $\beta 0$ y $\beta 1$

En la mayoría de casos, el estudio de regresión se aplica a una muestra pero el objetivo último es obtener un modelo lineal que explique la relación entre las dos variables en toda la población. Esto significa que el modelo generado es una estimación de la relación poblacional a partir de la relación que se observa en la muestra y por lo tanto está sujeta a variaciones. Para cada uno de los parámetros de la ecuación de regresión lineal ($\beta 0$ y $\beta 1$) se puede calcular su significancia (p-value) y su intervalo de confianza. El test estadístico empleado es el t-student.

El test de significancia para la pendiente ($\beta 1$) del modelo lineal considera como hipótesis lo siguiente:

- H_0: No hay relación lineal entre ambas variables por lo que la pendiente del modelo lineal es cero. $\beta 1 = 0$
- H_a: Sí hay relación lineal entre ambas variables por lo que la pendiente del modelo lineal es distinta de cero. $\beta 1 != 0$

Cálculo de *p-value*:

$$(t) = \frac{\hat{\beta}1 - 0}{SE(\hat{\beta}1)}$$

- El error estandar de la pendiente es complejo de calcular, suele estar incluido en el output generado por el ordenador al crear la recta de mínimos cuadrados. Descripción detallada en *Introduction to Stantistical Learning cap.3 Linear regression*.
- Grados de libertad (df) = número observaciones - 2
- p-value = P(|t| > valor calculado de t)

Intervalo de confianza para la pendiente ($\beta 1$) del modelo lineal:

$$\hat{\beta}1 \pm t_{df} SE(\hat{\beta}1)$$

En el caso de la ordenada en el origen $\beta 0$ el proceso a seguir es el mismo que para la pendiente.

En R, cuando se genera el modelo de regresión lineal, se devuelve junto con el valor de la pendiente y la ordenada en el origen el valor del estadístico *t* obtenido para cada uno y los *p-value* correspondientes. Esto permite saber además de la estimación de $\beta 0$ y $\beta 1$, si son significativamente distintos de 0. Si se desea conocer los intervalos de confianza para cada uno de los parámetros se puede calcular con la función `conint()`.

Residuos del modelo

El residuo de una estimación se define como la diferencia entre el valor observado y el valor esperado acorde al modelo. A la hora de sumarizar el conjunto de residuos hay dos posibilidades:

- El sumatorio del valor absoluto de cada residuo.
- El sumatorio del cuadrado de cada residuo. Esta es la aproximación más empleada (mínimos cuadrados) ya que magnifica las desviaciones más extremas.

- En R, cuando se genera un modelo los residuos también se calculan automáticamente y se almacenados dentro del modelo.

Cuanto mayor sea el sumatorio del cuadrado de los residuos menor es la precisión con la que el modelo puede predecir el valor de la variable dependiente a partir de la variable predictora. Los residuos son muy importantes puesto que en ellos se basan las diferentes medidas de la bondad de ajuste del modelo.

Bondad de ajuste del modelo

Una vez que se ha ajustado un modelo es necesario verificar su eficiencia, ya que aun siendo la línea que mejor se ajusta a las observaciones de entre todas las posibles, el modelo puede ser malo. Las medidas más utilizadas para medir la calidad del ajuste son: *el error residual estimado o error residual estándar*, *el test F* y *el coeficiente de determinación* R^2.

- **Error residual estándar:** Mide la desviación promedio de cualquier punto estimado por el modelo (generado a partir de una muestra) respecto de la verdadera recta de regresión poblacional. Tiene las mismas unidades que la variable dependiente y. Una forma de saber si el valor del error residual estándar es grande consiste en dividirlo entre el valor medio de la variable respuesta, obteniendo así un % de la desviación.
- **Coeficiente de determinación R^2:** Describe la proporción de variabilidad observada en la variable dependiente *Y* explicada por el modelo y relativa a la variabilidad total. Su valor está acotado entre 0 y 1. Al ser adimensional presenta la ventaja frente al error residual estándar de ser más fácil de interpretar.

$$R^2 = \frac{Suma\ de\ cuadrados\ totales - Suma\ de\ cuadrados\ residuales}{Suma\ de\ cuadrados\ totales} =$$

$$1 - \frac{Suma\ de\ cuadrados\ residuales}{Suma\ de\ cuadrados\ totales}$$

- En los modelos de regresión lineal simple el valor de R^2 se corresponde con el cuadrado del *coeficiente de correlación de Pearson (r)* entre *X* e *Y*, no siendo así en regresión múltiple. Existe una modificación de R^2 conocida como R^2-ajustado que se emplea

principalmente en los modelos de regresión múltiple. Introduce una penalización cuantos más predictores se incorporan al modelo. En los modelos lineales simples no se emplea.

- **Test F:** El test F es un test de hipótesis que considera como hipótesis nula que todos los coeficientes de correlación estimados son = 0, frente a la hipótesis nula de que al menos uno de ellos no lo es. Se emplea en modelos de regresión múltiple para saber si al menos alguno de los predictores introducidos en el modelo contribuye de forma significativa. En modelos lineales simples, dado que solo hay un predictor, el *p-value* del test F es igual al *p-value* del t-test del predictor.

Condiciones para la regresión lineal

1. **Linealidad:** La relación entre ambas variables debe ser lineal. Para comprobarlo se puede recurrir a:
   - Graficar ambas variables a la vez (*scatterplot* o diagrama de dispersión), superponiendo la recta del modelo generado por regresión lineal.
   - Calcular los residuos para cada observación acorde al modelo generado y graficarlos (*scatterplot*). Deben distribuirse de forma aleatoria en torno al valor 0.

2. **Distribución Normal de los residuos:** Los residuos se tiene que distribuir de forma normal, con media igual a 0. Esto se puede comprobar con un histograma, con la distribución de cuantiles (ppnorm + qqline) o con un test de hipótesis de normalidad. Los valores extremos suelen ser una causa frecuente por la que se viola la condición de normalidad.

3. **Varianza de residuos constante (homocedasticidad):** La varianza de los residuos a de ser aproximadamente constante a lo largo del eje X. Esto se comprueba graficando (scatterplot) los residuos de cada observación. Formas cónicas son un claro indicio de falta de homocedasticidad.

4. **Valores atípicos y de alta influencia:** Hay que estudiar con detenimiento los valores atípicos o extremos ya que pueden generar una falsa correlación que realmente no existe, o ocultar una existente.

5. **Independencia, Autocorrelación**: Las observaciones deben ser independientes unas de otras. Esto es importante a tenerlo en cuenta cuando se trata de mediciones temporales. Puede detectarse estudiando si los residuos siguen un patrón o tendencia.

Dado que las condiciones se verifican a partir de los residuos, primero se suele generar el modelo y después se valida.

Predicción de valores

Una vez generado un modelo que se pueda considerar válido, es posible predecir el valor de la variable dependiente y para nuevos valores de la variable predictora x. Es importante tener en cuenta que las predicciones deben *a priori* limitarse al rango de valores dentro del que se encuentran las observaciones con las que se ha generado el modelo. Esto es importante puesto que solo en esta región se tiene certeza de que se cumplen las condiciones para que el modelo sea válido. Para calcular las predicciones se emplea la ecuación generada por regresión.

Dado que el modelo generado se ha obtenido a partir de una muestra y por lo tanto las estimaciones de los coeficientes de regresión tienen un error asociado, también lo tienen los valores de las predicciones. Existen dos formas de medir la incertidumbre asociada con una predicción:

- **Intervalos de confianza:** Responden a la pregunta ¿Cuál es el intervalo de confianza del valor promedio de la variable respuesta y para un determinado valor del predictor x?

$$\hat{y} \pm t_{\alpha/2,n-2} \sqrt{MSE \left(\frac{1}{n} + \frac{(x_k - \overline{x})^2}{\sum (x_i - \overline{x}^2)} \right)}$$

- **Intervalos de predicción:** Responden a la pregunta ¿Dentro de que intervalo se espera que esté el valor de la variable respuesta y para un determinado valor del predictor x?

$$\hat{y} \pm t_{\alpha/2,n-2} \sqrt{MSE \left(1 + \frac{1}{n} + \frac{(x_k - \overline{x})^2}{\sum (x_i - \overline{x})^2} \right)}$$

Si bien ambas preguntas parecen similares, la diferencia se encuentra en que los intervalos de confianza se aplican al valor promedio que se espera de y para un determinado valor de x, mientras que los intervalos de predicción no se aplican al promedio. Por esta razón los segundos siempre son más amplios que los segundos.

En R se puede emplear la función `predict()` que recibe como argumento el modelo calculado, un *dataframe* con los nuevos valores del predictor (x) y el tipo de intervalo (*confidence* o *prediction*).

Ejemplo

Un analista de deportes quiere saber si existe una relación entre el número de bateos que realiza un equipo de béisbol y el número de runs que consigue. En caso de existir y de establecer un modelo, podría predecir el resultado del partido.

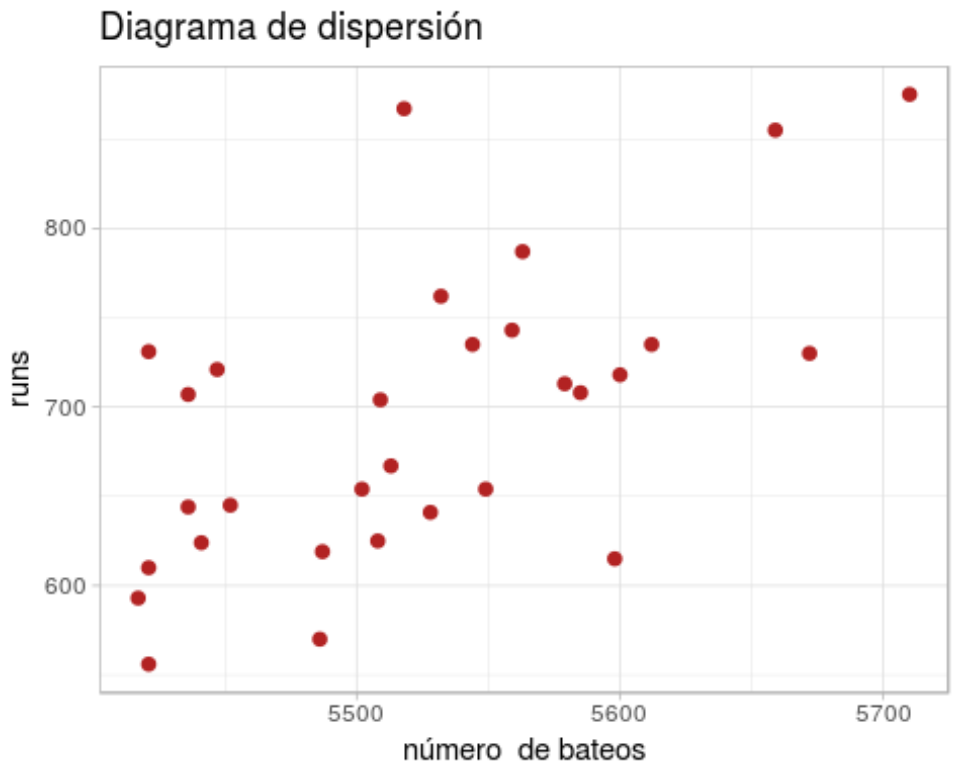
```
equipos <- c("Texas", "Boston", "Detroit", "Kansas", "St.", "New_S.", "New_Y.",
"Milwaukee", "Colorado", "Houston", "Baltimore", "Los_An.", "Chicago",
"Cincinnati", "Los_P.", "Philadelphia", "Chicago", "Cleveland", "Arizona",
"Toronto", "Minnesota", "Florida", "Pittsburgh", "Oakland", "Tampa", "Atlanta",
"Washington", "San.F", "San.I", "Seattle")
numero_bateos <- c(5659, 5710, 5563, 5672, 5532, 5600, 5518, 5447, 5544, 5598,
    5585, 5436, 5549, 5612, 5513, 5579, 5502, 5509, 5421, 5559, 5487, 5508,
    5421, 5452, 5436, 5528, 5441, 5486, 5417, 5421)
runs <- c(855, 875, 787, 730, 762, 718, 867, 721, 735, 615, 708, 644, 654, 735,
    667, 713, 654, 704, 731, 743, 619, 625, 610, 645, 707, 641, 624, 570, 593,
    556)
datos <- data.frame(equipos, numero_bateos, runs)
head(datos)
```

```
##   equipos numero_bateos runs
## 1   Texas          5659  855
## 2  Boston          5710  875
## 3 Detroit          5563  787
## 4  Kansas          5672  730
## 5     St.          5532  762
## 6  New_S.          5600  718
```

1.Representación gráfica de las observaciones

El primer paso antes de generar un modelo de regresión es representar los datos para poder intuir si existe una relación y cuantificar dicha relación mediante un coeficiente de correlación. Si en este paso no se detecta la posible relación lineal, no tiene sentido seguir adelante generando un modelo lineal (*se tendrían que probar otros modelos*).

```
require(ggplot2)
ggplot(data = datos, mapping = aes(x = numero_bateos, y = runs)) +
 geom_point(color = "firebrick", size = 2) +
 labs(title = "Diagrama de dispersión", x = "número  de bateos") +
 theme_light()
```




```
cor.test(x = datos$numero_bateos, y = datos$runs, method = "pearson")
```

```
##
##  Pearson's product-moment correlation
##
## data:  datos$numero_bateos and datos$runs
## t = 4.0801, df = 28, p-value = 0.0003388
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.3209675 0.7958231
## sample estimates:
##      cor
## 0.610627
```

El gráfico y el test de correlación muestran una relación lineal, de intensidad considerable (r = 0.61) y significativa (*p-value* = 0.0003388). Tiene sentido intentar generar un modelo de regresión lineal que permita predecir el número de *runs* en función del número de bateos del equipo.

2.Cálculo del modelo de regresión lineal simple

```
modelo_lineal <- lm(runs ~ numero_bateos, datos)
# lm() devuelve el valor de la variable y para x=0 (intersección) junto con
# la pendiente de la recta, para ver la información del modelo se requiere
# summary().
summary(modelo_lineal)
```

```
##
## Call:
## lm(formula = runs ~ numero_bateos, data = datos)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -125.58  -47.05  -16.59   54.40  176.87
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -2789.2429   853.6957  -3.267 0.002871 **
## numero_bateos     0.6305     0.1545   4.080 0.000339 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 66.47 on 28 degrees of freedom
## Multiple R-squared:  0.3729, Adjusted R-squared:  0.3505
## F-statistic: 16.65 on 1 and 28 DF,  p-value: 0.0003388
```

La primera columna (*Estimate*) devuelve el valor estimado para los dos parámetros de la ecuación del modelo lineal ($\hat{\beta}_0$ y $\hat{\beta}_1$) que equivalen a la ordenada en el origen y la pendiente.

Se muestran los errores estándar, el valor del estadístico *t* y el *p-value* (dos colas) de cada uno de los dos parámetros. Esto permite determinar si los parámetros son significativamente distintos de 0, es decir, que tienen importancia en el modelo. En los modelos de regresión lineal simple, el parámetro importante suele ser la pendiente.

Para el modelo generado, tanto la ordenada en el origen como la pendiente son significativas (*p-values* < 0.05).

El valor de R^2 indica que el modelo calculado explica el 37.29% de la variabilidad presente en la variable respuesta (*runs*) mediante la variable independiente (*número de bateos*).

El *p-value* obtenido en el test F (0.0003388) determina que sí es significativamente superior la varianza explicada por el modelo en comparación a la varianza total. Es el parámetro que determina si el modelo es significativo y por lo tanto se puede aceptar.

El modelo lineal generado sigue la ecuación **runs = -2789.2429 + 0.6305 bateos**.

Por cada unidad que se incrementa el número de bateos, el número de runs aumenta **en promedio** 0.6305 unidades.

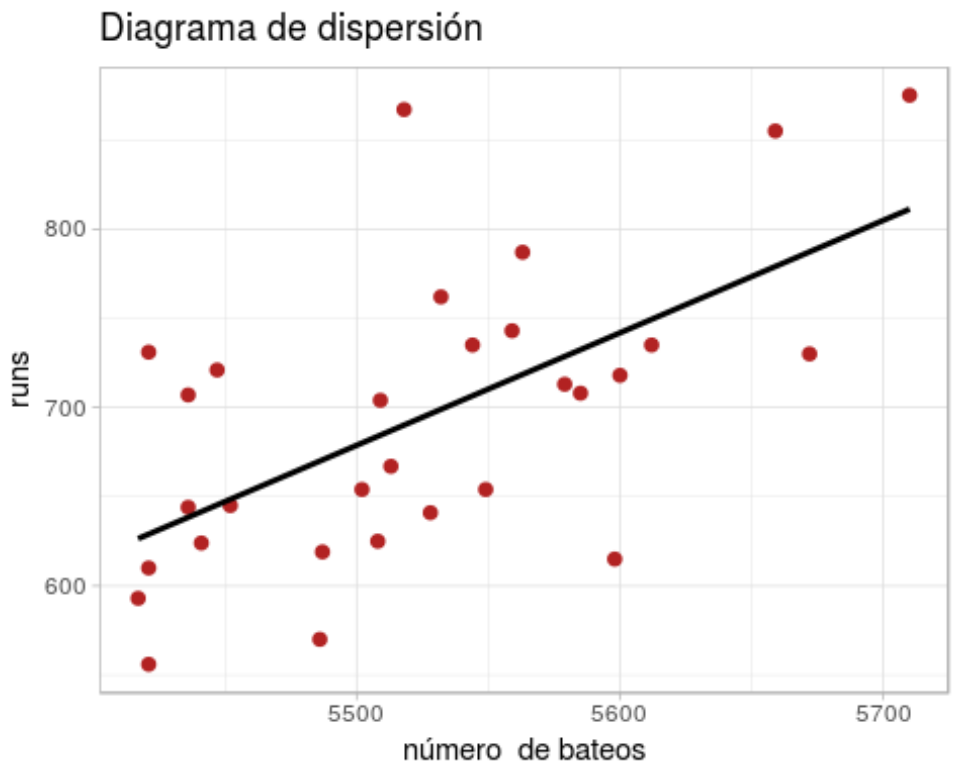
3.Intervalos de confianza para los parámetros del modelo

```
confint(modelo_lineal)
```

```
##                       2.5 %        97.5 %
## (Intercept)   -4537.9592982 -1040.5264727
## numero_bateos     0.3139863     0.9471137
```

4.Representación gráfica del modelo

```
ggplot(data = datos, mapping = aes(x = numero_bateos, y = runs)) + geom_point(color
= "firebrick",
    size = 2) + labs(title = "Diagrama de dispersión", x = "número  de bateos") +
    geom_smooth(method = "lm", se = FALSE, color = "black") + theme_light()
```



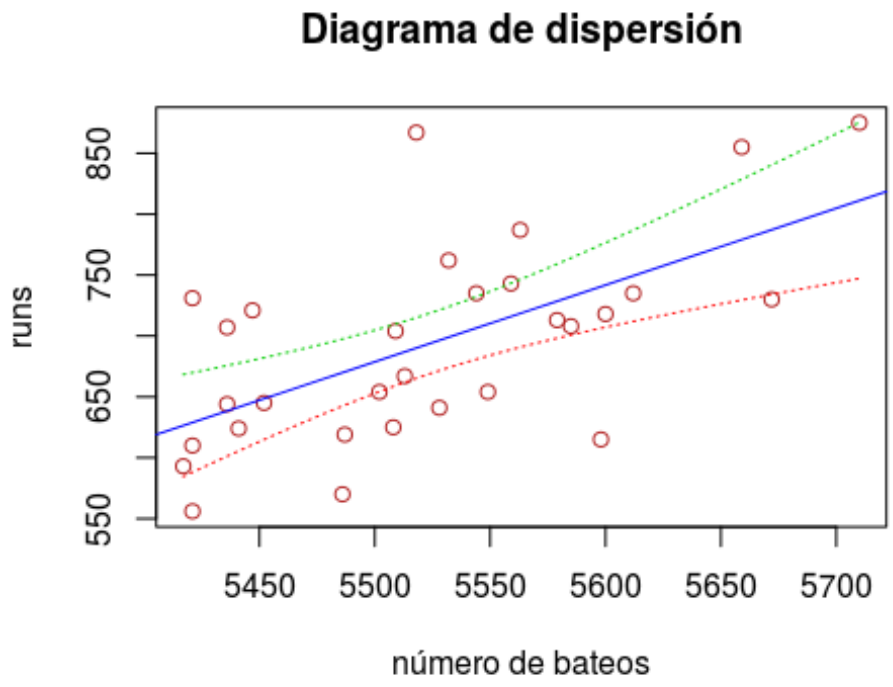
Además de la línea de mínimos cuadrados es recomendable incluir los límites superior e inferior del intervalo de confianza. Esto permite identificar la región en la que, según el modelo generado y para un determinado nivel de confianza, se encuentra el valor promedio de la variable dependiente.

Para poder representar el intervalo de confianza a lo largo de todo el modelo se recurre a la función `predict()` para predecir valores que abarquen todo el eje *X*. Se añaden al gráfico líneas formadas por los límites superiores e inferiores calculados para cada predicción.

```
# Se genera una secuencia de valores x_i que abarquen todo el rango de las
# observaciones de la variable X
puntos <- seq(from = min(datos$numero_bateos), to = max(datos$numero_bateos),
    length.out = 100)
# Se predice el valor de la variable Y junto con su intervalo de confianza
# para cada uno de los puntos generados. En la función predict hay que
# nombrar a los nuevos puntos con el mismo nombre que la variable X del
# modelo. Devuelve una matriz.
limites_intervalo <- predict(object = modelo_lineal, newdata =
data.frame(numero_bateos = puntos),
    interval = "confidence", level = 0.95)
head(limites_intervalo, 3)
```

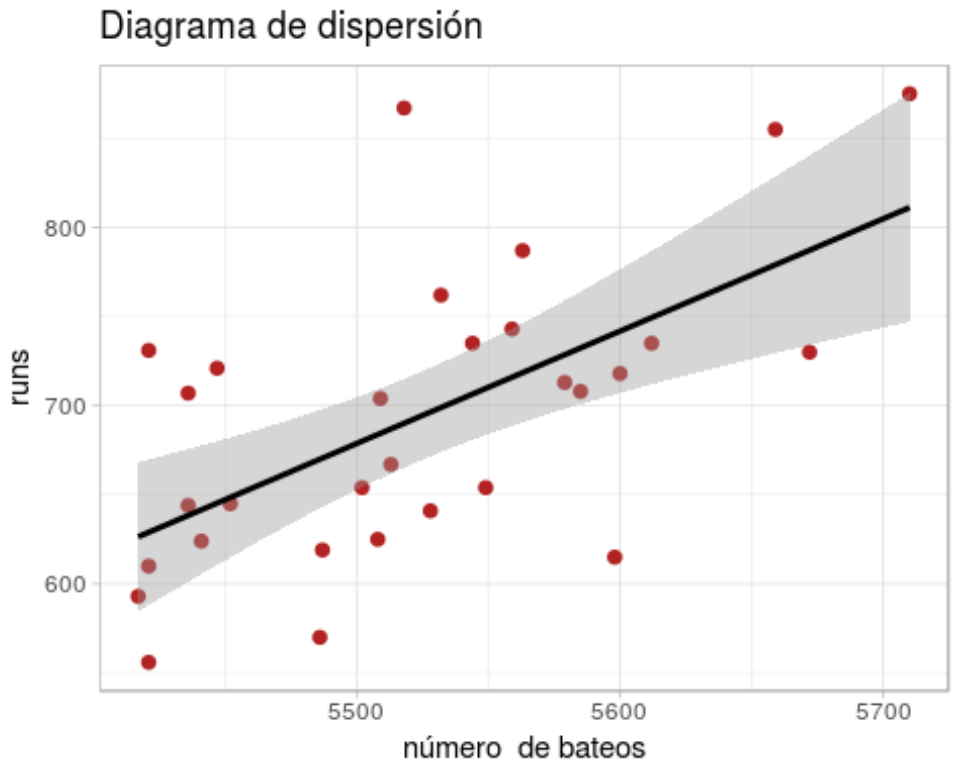
```
##        fit      lwr      upr
## 1 626.4464 584.5579 668.3350
## 2 628.3126 587.1743 669.4509
## 3 630.1788 589.7830 670.5745
```

```
# Finalmente se añaden al gráfico las lineas formadas por los límites
# superior e inferior.
plot(datos$numero_bateos, datos$runs, col = "firebrick", ylab = "runs", xlab =
"número de bateos", main = "Diagrama de dispersión")
abline(modelo_lineal, col = 4)
lines(x = puntos, y = limites_intervalo[, 2], type = "l", col = 2, lty = 3)
lines(x = puntos, y = limites_intervalo[, 3], type = "l", col = 3, lty = 3)
```



La función `geom_smooth()` del paquete *ggplot2* genera la regresión y su intervalo de forma directa.

```
ggplot(data = datos, mapping = aes(x = numero_bateos, y = runs)) +
 geom_point(color = "firebrick", size = 2) +
 labs(title = "Diagrama de dispersión", x = "número  de bateos") +
 geom_smooth(method = "lm", se = TRUE, color = "black") +
 theme_light()
# Add linear regression line, by default includes 95% confidence region.
```



5.Verificar condiciones para poder aceptar un modelo lineal

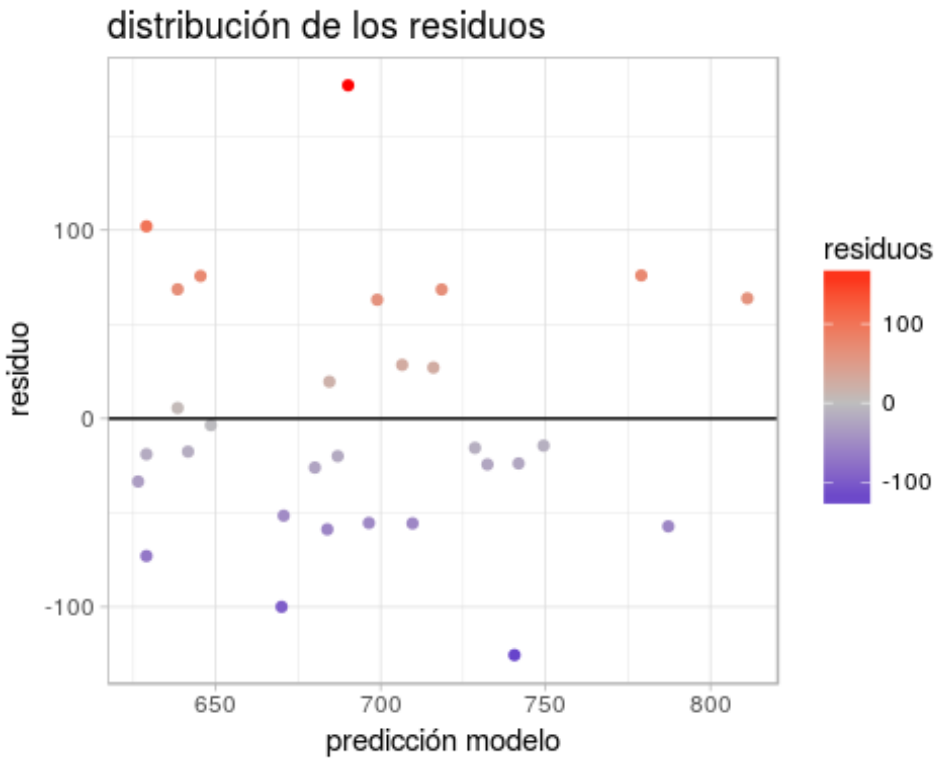
Relación lineal entre variable dependiente e independiente:

Se calculan los residuos para cada observación y se representan (scatterplot). Si las observaciones siguen la línea del modelo, los residuos se deben distribuir aleatoriamente entorno al valor 0.

```
# La función lm() calcula y almacena los valores predichos por el modelo y
# los residuos.
datos$prediccion <- modelo_lineal$fitted.values
datos$residuos <- modelo_lineal$residuals
head(datos,4)
```

```
##   equipos numero_bateos runs prediccion  residuos
## 1   Texas          5659  855   779.0395  75.96048
## 2  Boston          5710  875   811.1976  63.80243
## 3 Detroit          5563  787   718.5067  68.49328
## 4  Kansas          5672  730   787.2367 -57.23667
```

```
ggplot(data = datos, aes(x = prediccion, y = residuos)) +
 geom_point(aes(color = residuos)) +
 scale_color_gradient2(low = "blue3", mid = "grey", high = "red") +
  geom_hline(yintercept = 0) +
  labs(title = "distribución de los residuos", x = "predicción modelo", y =
"residuo") +
 theme_light()
```

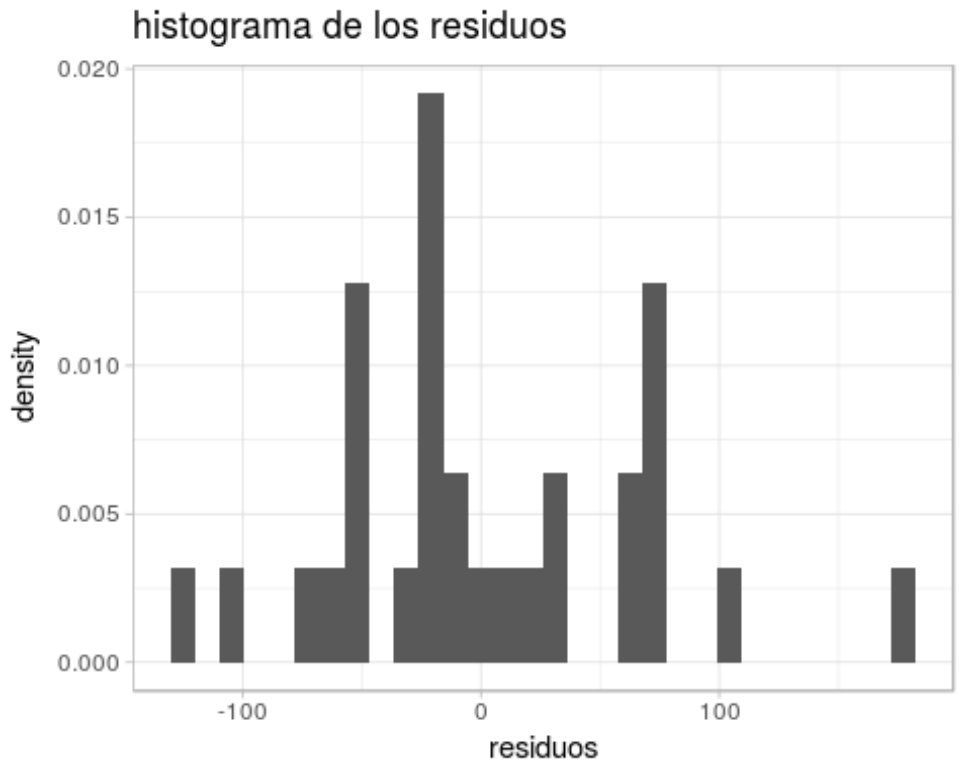


Los residuos se distribuyen de forma aleatoria entorno al 0 por lo que se acepta la linealidad.

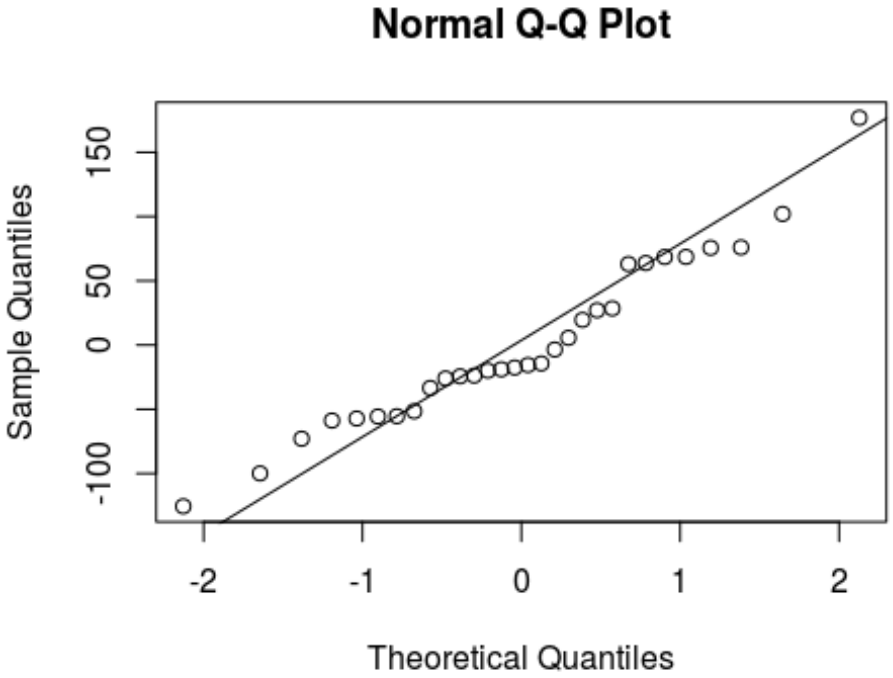
Distribución normal de los residuos:

Los residuos se deben distribuir de forma normal con media 0. Para comprobarlo se recurre a histogramas, a los cuantiles normales o a un test de contraste de normalidad.

```
ggplot(data = datos, aes(x = residuos)) +
 geom_histogram(aes(y = ..density..)) +
 labs(title = "histograma de los residuos") +
 theme_light()
```




```
qqnorm(modelo_lineal$residuals)
qqline(modelo_lineal$residuals)
```



```
shapiro.test(modelo_lineal$residuals)
```

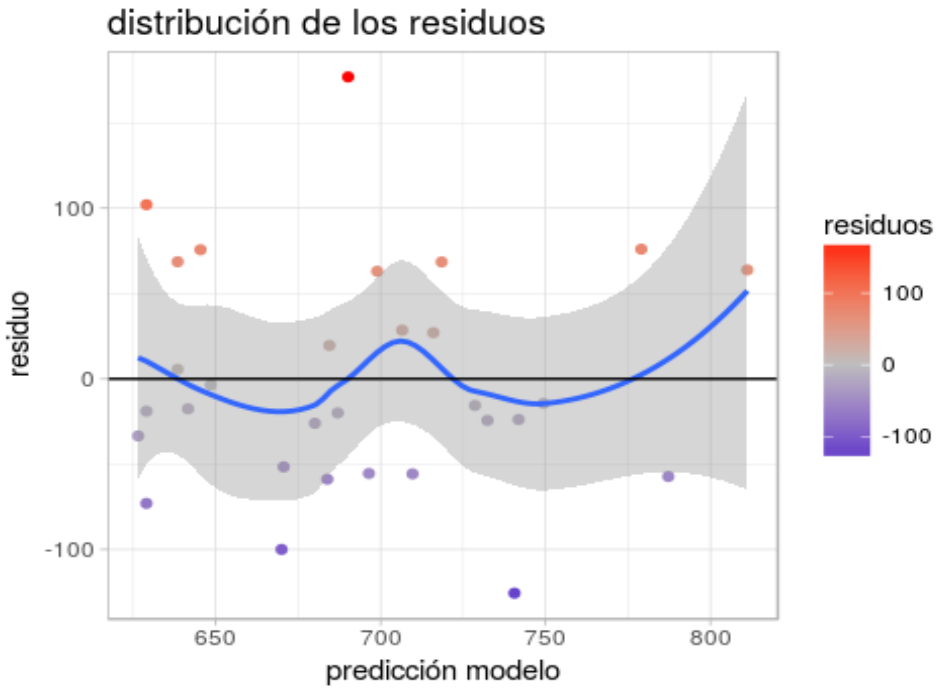
```
##  Shapiro-Wilk normality test
##
## data:  modelo_lineal$residuals
## W = 0.96144, p-value = 0.337
```

Tanto la representación gráfica como el diagnóstico analítico confirman la distribución normal de los residuos.

Varianza constante de los residuos (Homocedasticidad):

La variabilidad de los residuos debe de ser constante a lo largo del eje X. Un patrón cónico es indicativo de falta de homogeneidad en la varianza. También existen test de hipótesis *Test de Breusch-Pagan* y *Test de Golfeld-Quandt* (no explicados aquí).

```
ggplot(data = datos, aes(x = prediccion, y = residuos)) +
 geom_point(aes(color = residuos)) +
 scale_color_gradient2(low = "blue3", mid = "grey", high = "red") +
 geom_smooth() +
 labs(title="distribución de los residuos",x ="predicción modelo", y = "residuo") +
 geom_hline(yintercept = 0) +
 theme_light()
```

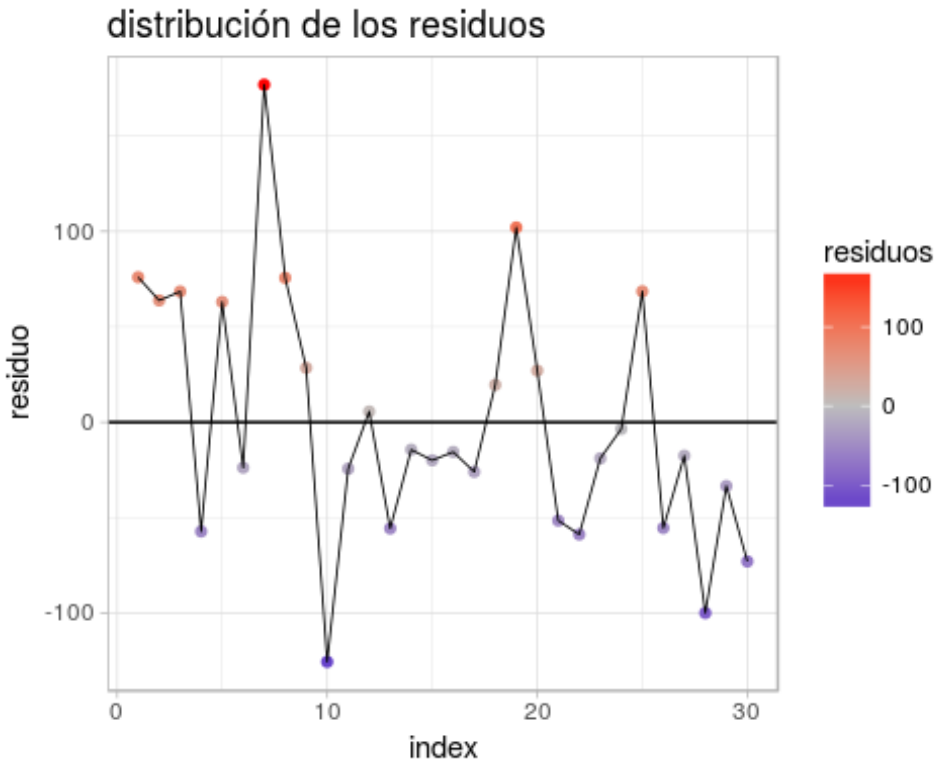


No existe ningún patrón que haga sospechar falta de homocedasticidad.

Autocorrelación de residuos:

Cuando se trabaja con intervalos de tiempo, es muy importante comprobar que no existe aoutocorrelación de los residuos, es decir que son independientes. Esto puede hacerse detectando visualmente patrones en la distribución de los residuos cuando se ordenan según se han registrado o con el test de Durbin-Watson `dwt()` del paquete *Car*.

```
ggplot(data = datos, aes(x = seq_along(residuos), y = residuos)) +
geom_point(aes(color = residuos)) +
  scale_color_gradient2(low = "blue3", mid = "grey", high = "red") +
 geom_line(size = 0.3) +
 labs(title = "distribución de los residuos", x = "index", y = "residuo") +
 geom_hline(yintercept = 0) +
 theme_light()
```



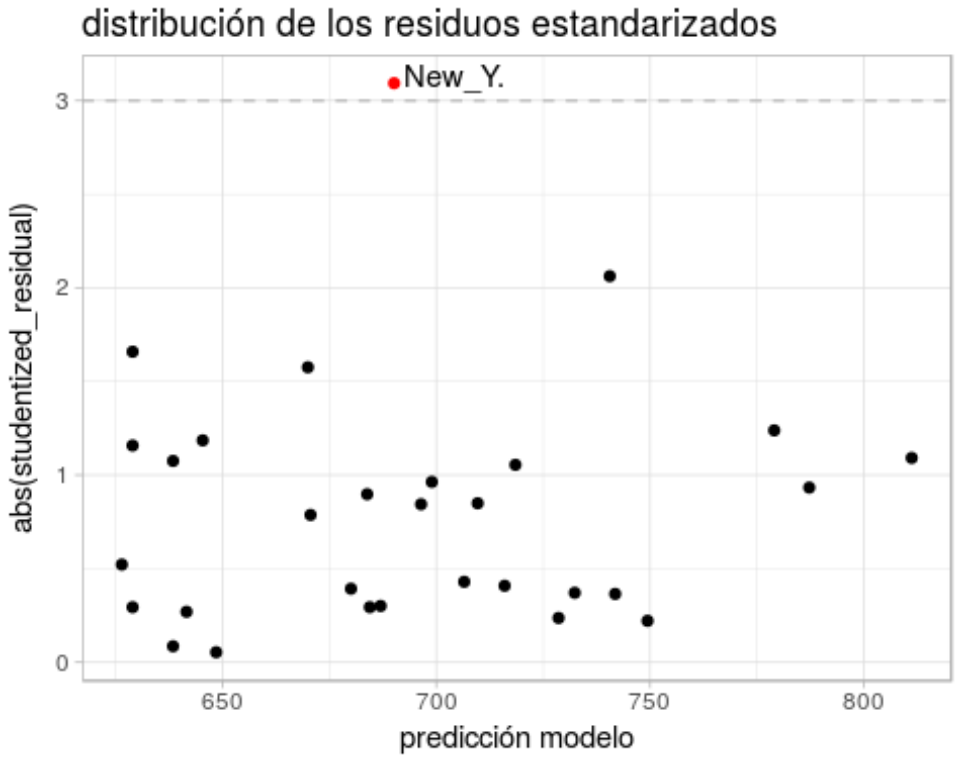
En este caso, la representación de los residuos no muestra ninguna tendencia.

6.Identificación de valores atípicos (*outliers*) o influyentes

Independientemente de que el modelo se haya podido aceptar, siempre es conveniente identificar si hay algún posible *outlier* u observación altamente influyente, puesto que podría estar condicionando en gran medida el modelo. La eliminación de este tipo de observaciones debe de analizarse con detalle y dependiendo de la finalidad del modelo. Si el fin es predictivo, un modelo sin *outliers* ni observaciones altamente influyentes puede ser más útil para predecir con mayor precisión la mayoría de casos. Sin embargo, es muy importante prestar atención a estos valores ya que de no ser errores de medida pueden ser los casos más interesantes. El modo adecuado a proceder cuando se sospecha de algún posible valor atípico o influyente es calcular el modelo de regresión incluyendo y excluyendo dicho valor.

Si se quiere hacer un estudio analítico de los posibles valores atípicos se recurre a los *studentized residuals* que se obtienen al dividir los residuos de cada observación entre una estimación de su error estándar. Si se supera un valor absoluto de 3 debe de considerarse esa observación como posible valor atípico. En R se pueden calcular con la función `rstudent()`.

```
library(ggrepel)
library(dplyr)
datos$studentized_residual <- rstudent(modelo_lineal)
ggplot(data = datos, aes(x = prediccion, y = abs(studentized_residual))) +
    geom_hline(yintercept = 3, color = "grey", linetype = "dashed") +
    #se identifican en rojo observaciones con residuos estandarizados absolutos > 3
    geom_point(aes(color = ifelse(abs(studentized_residual) >3, 'red', 'black'))) +
    scale_color_identity() +
    #se muestra el equipo al que pertenece la observación atípica
    geom_text_repel(data = filter(datos, abs(studentized_residual) > 3), aes(label
= equipos)) +
    labs(title = "distribución de los residuos estandarizados", x = "predicción
modelo") +
    theme_light()
```




```
datos %>% filter(abs(studentized_residual) > 3)
```

```
##   equipos numero_bateos runs prediccion residuos studentized_residual
## 1  New_Y.          5518  867    690.132  176.868             3.092876
```

```
which(abs(datos$studentized_residual) > 3)
```

```
## [1] 7
```

El estudio de los residuos estandarizados identifica al equipo de New_Y. como una posible observación atípica. Esta observación ocupa la posición 7 en la tabla de datos.

Además de que un valor sea atípico, es necesario estudiar como de influyente es en el conjunto del modelo. Si un valor es atípico pero no influyente, no es crítico que se considere su eliminación. Dos de las medidas más empleadas para cuantificar la influencia son:

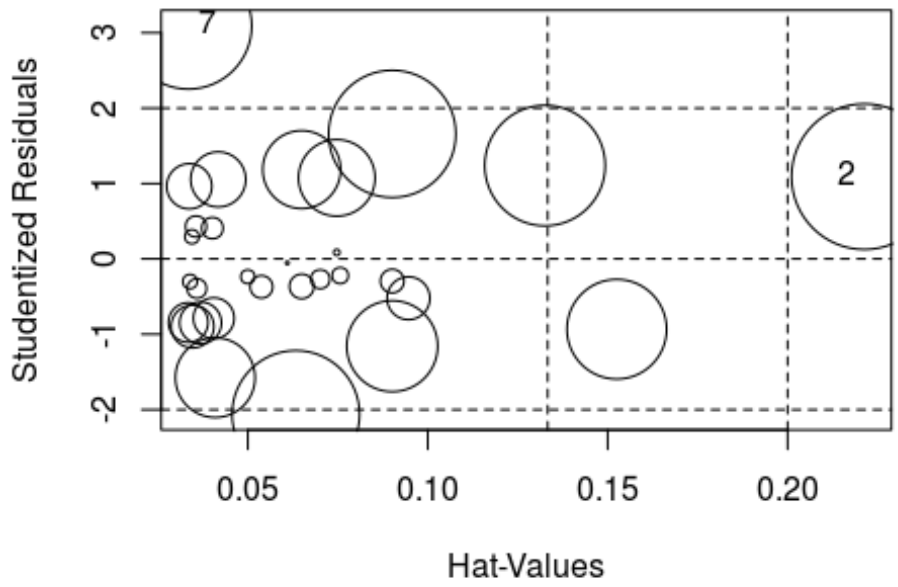
- Leverages (hat): Se consideran observaciones influyentes aquellas cuyos valores *hat* superen $2.5x((p+1)/n)$, siendo p el número de predictores y n el número de observaciones.
- Distancia Cook (cook.d): Se consideran influyentes valores superiores a 1.

En R se dispone de la función `outlierTest()` del paquete *car* y de las funciones `influence.measures()`, `influence.plot()` y `hatvalues()` para identificar las observaciones más influyentes en el modelo.

```
library(car)
summary(influence.measures(model = modelo_lineal))
```

```
## Potentially influential observations of
##   lm(formula = runs ~ numero_bateos, data = datos) :
##
##   dfb.1_ dfb.nmr_ dffit cov.r   cook.d hat
## 2 -0.53   0.54     0.58  1.27_*  0.17   0.22_*
## 7  0.05  -0.04     0.58  0.61_*  0.13   0.03
```

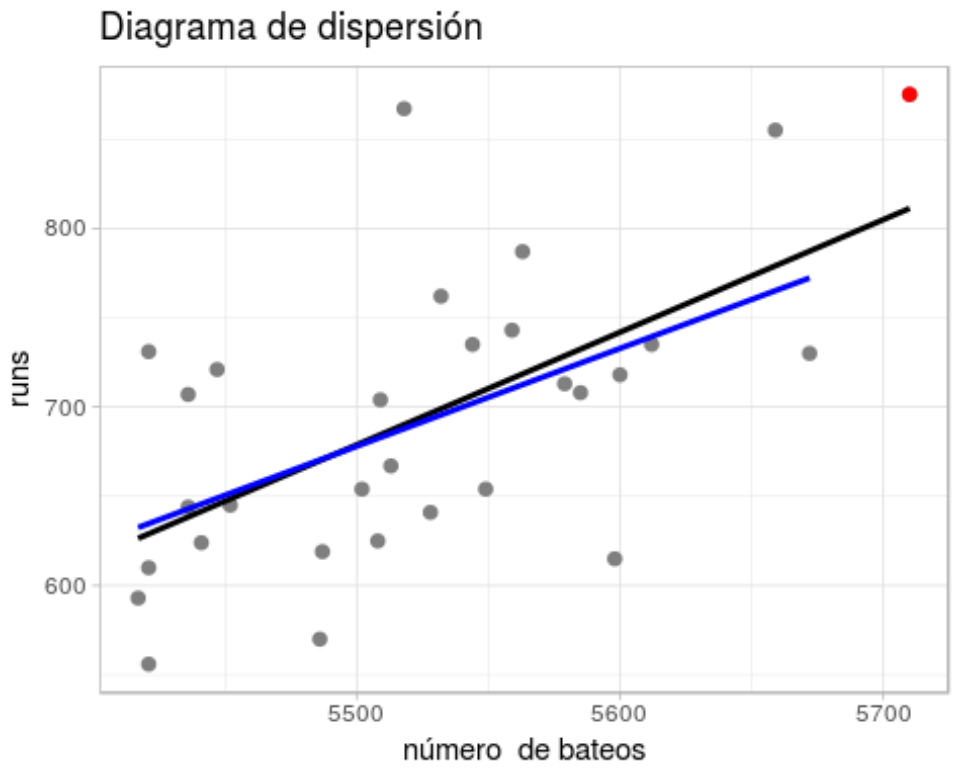
```
influencePlot(model = modelo_lineal)
```




```
##    StudRes        Hat     CookD
## 2 1.091428 0.22133381 0.1681516
## 7 3.092876 0.03349684 0.1269339
```

Las funciones `influence.measures()` e `influence.plot()` detectan la observación 7 como atípica pero no significativamente influyente. Sí detectan como influyente la observación que ocupa la segunda posición. Para evaluar hasta qué punto condiciona el modelo, se recalcula la recta de mínimos cuadrados excluyendo esta observación

```
ggplot(data = datos, mapping = aes(x = numero_bateos, y = runs)) +
 geom_point(color = "grey50", size = 2) +
 geom_smooth(method = "lm", se = FALSE, color = "black") + # se resalta el valor
excluido
 geom_point(data = datos[2, ], color = "red", size = 2) + # se añade la nueva recta
de mínimos cuadrados
 geom_smooth(data = datos[-2, ], method = "lm", se = FALSE, color = "blue") +
 labs(title = "Diagrama de dispersión", x = "número  de bateos") +
 theme_light()
```



La eliminación del valor identificado como influyente apenas cambia la recta de mínimos cuadrados. Para conocer con exactitud el resultado de excluir la observación se comparan las pendientes de ambos modelos.

```
lm(formula = runs ~ numero_bateos, data = datos)$coefficients
```

```
##   (Intercept) numero_bateos
##   -2789.24289       0.63055
```

```
lm(formula = runs ~ numero_bateos, data = datos[-2, ])$coefficients
```

```
##   (Intercept) numero_bateos
## -2335.7478247     0.5479527
```

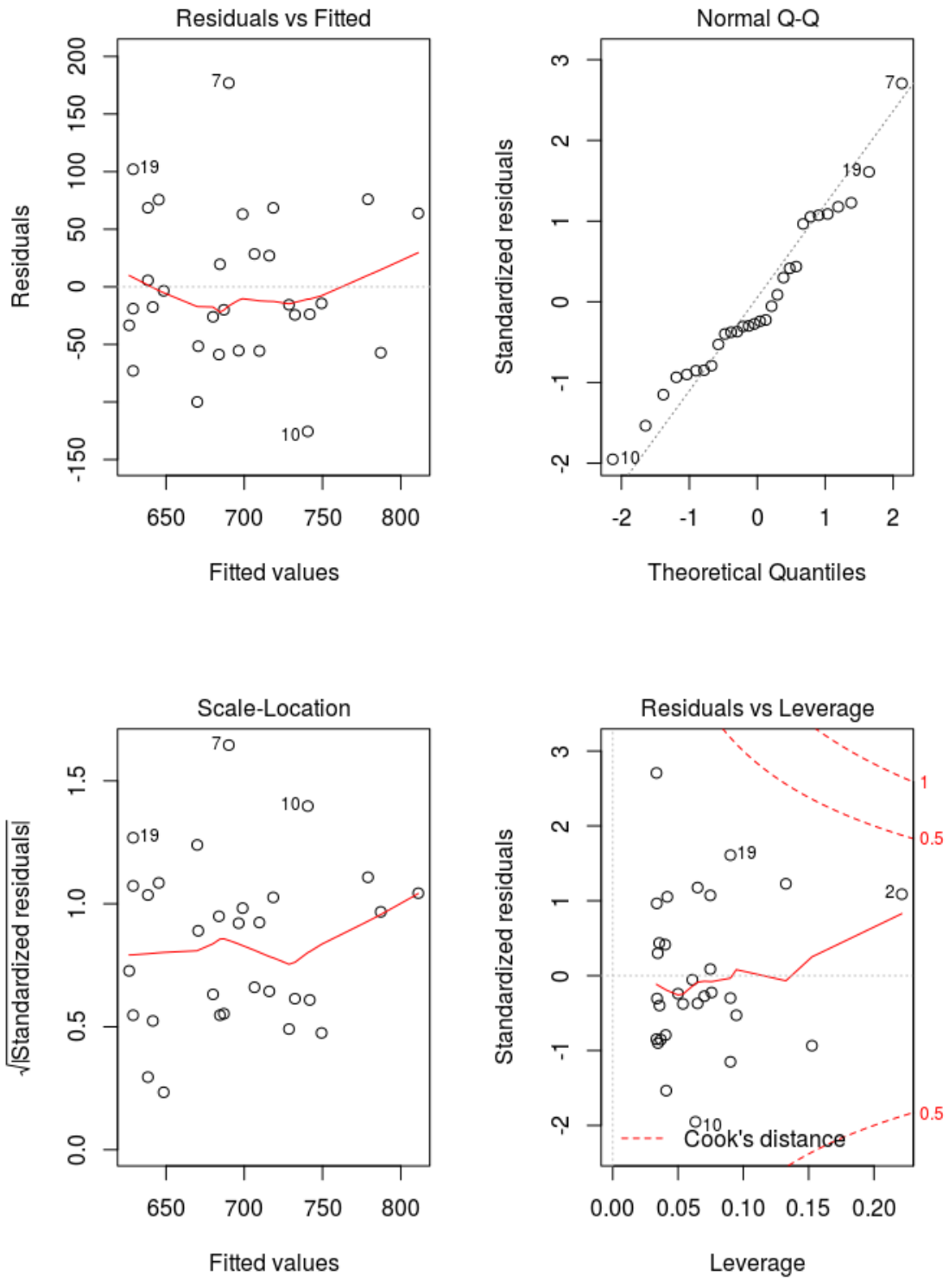
Conclusión

Dado que se satisfacen todas las condiciones para considerar válido un modelo de regresión lineal por mínimos cuadrados y que el *p-value* indica que el ajuste es significativo, se puede aceptar el modelo lineal. A pesar de ello, el valor de R^2 no es muy alto por lo que el número de bateos no es muy buen predictor del número de runs.

Evaluación de los residuos de un modelo lineal simple mediante gráficos R

Como se ha descrito en el apartado anterior, la evaluación de las condiciones que hacen válido un modelo de regresión lineal simple se hace en gran medida mediante representaciones gráficas de los residuos. El objeto devuelto por la función `lm()` puede pasarse como argumento a la función `plot()` obteniendo 4 gráficos que permiten evaluar los residuos.

```
par(mfrow = c(1, 2))
plot(modelo_lineal)
```




```
par(mfrow = c(1, 1))
```

Si bien es una forma muy rápida de obtener los gráficos, no son estéticamente muy "bonitos". A continuación se describe como obtener las mismas representaciones mediante el sistema gráfico *ggplot2*. *Información obtenida de https://rpubs.com/therimalaya/43190 y de https://drsimonj.svbtle.com/visualising-residuals.*

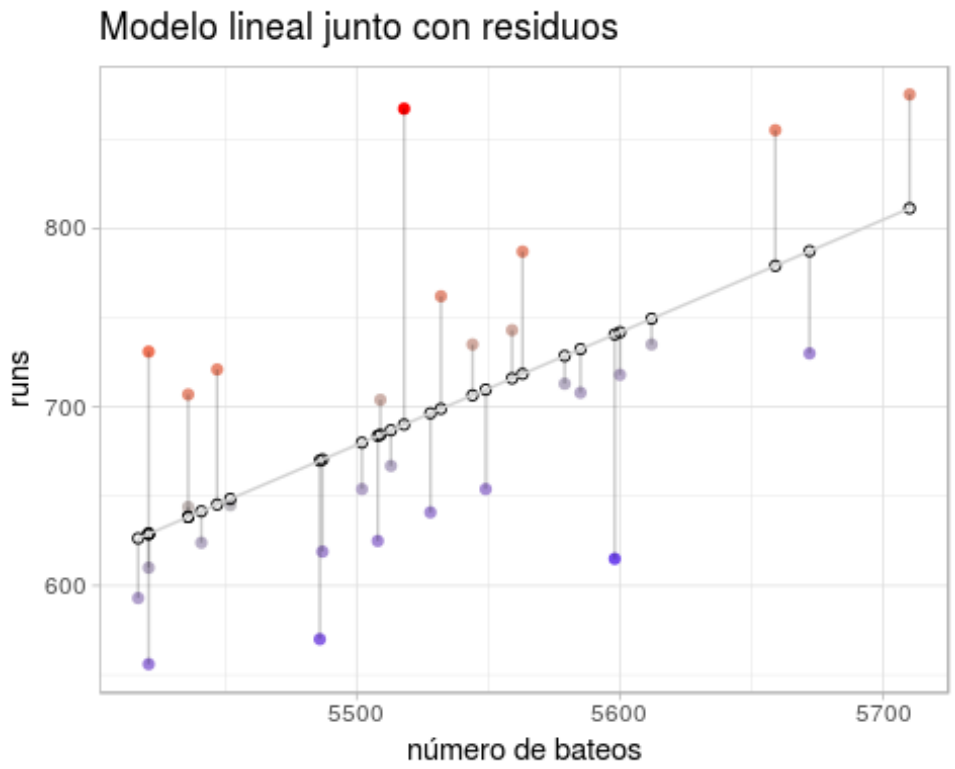
Modelo + residuos

```
equipos <- c("Texas", "Boston", "Detroit", "Kansas", "St.", "New_S.", "New_Y.",
"Milwaukee", "Colorado", "Houston", "Baltimore", "Los_An.", "Chicago",
"Cincinnati", "Los_P.", "Philadelphia", "Chicago", "Cleveland", "Arizona",
"Toronto", "Minnesota", "Florida", "Pittsburgh", "Oakland", "Tampa", "Atlanta",
"Washington", "San.F", "San.I", "Seattle")
numero_bateos <- c(5659, 5710, 5563, 5672, 5532, 5600, 5518, 5447, 5544, 5598,
    5585, 5436, 5549, 5612, 5513, 5579, 5502, 5509, 5421, 5559, 5487, 5508,
    5421, 5452, 5436, 5528, 5441, 5486, 5417, 5421)
runs <- c(855, 875, 787, 730, 762, 718, 867, 721, 735, 615, 708, 644, 654, 735,
    667, 713, 654, 704, 731, 743, 619, 625, 610, 645, 707, 641, 624, 570, 593,
    556)
datos <- data.frame(equipos, numero_bateos, runs)

# ajuste del modelo lineal simple
modelo <- lm(formula = runs ~ numero_bateos, data = datos)

# cálculo de los valores predichos y de los residuos
datos$prediccion <- predict(modelo)
datos$residuos <- residuals(modelo)
head(datos, 4)
```

```
##   equipos numero_bateos runs prediccion  residuos
## 1   Texas          5659  855   779.0395  75.96048
## 2  Boston          5710  875   811.1976  63.80243
## 3 Detroit          5563  787   718.5067  68.49328
## 4  Kansas          5672  730   787.2367 -57.23667
```

```
# representar las observaciones el color de los puntos se asocia al tamaño
# del residuo
p <- ggplot(data = datos, aes(x = numero_bateos, y = runs))
p <- p + geom_point(aes(color = residuos))
p <- p + scale_color_gradient2(low = "blue", mid = "grey", high = "red")
# añadir los valores acorde al modelo generado representandolos con otro
# tipo de puntos
p <- p + geom_point(aes(x = numero_bateos, y = prediccion), shape = 1)

# añadir segmentos que unan cada observación con su residuo
p <- p + geom_segment(aes(xend = numero_bateos, yend = prediccion), alpha = 0.2)
```

```
# añadir recta de mínimos cuadrados
p <- p + geom_smooth(method = "lm", se = FALSE, colour = "lightgrey", size = 0.5)

p <- p + labs(title = "Modelo lineal junto con residuos", x = "número de bateos")
p <- p + theme_light() + guides(color = FALSE)
p
```



Análisis gráfico de residuos

La combinación de los paquetes *broom* y *ggplot2* (ambos de Hadley Wickham) permiten obtener representaciones gráficas de modelos estadísticos. El paquete *broom* permite convertir objetos de análisis estadísticos tales como *lm, t.test, anova*... en tablas ordenadas *Tidy Data Frames*.

Una vez se ha obtenido el *data frame* a partir de un objeto estadístico, es muy sencillo generar representaciones gráficas.

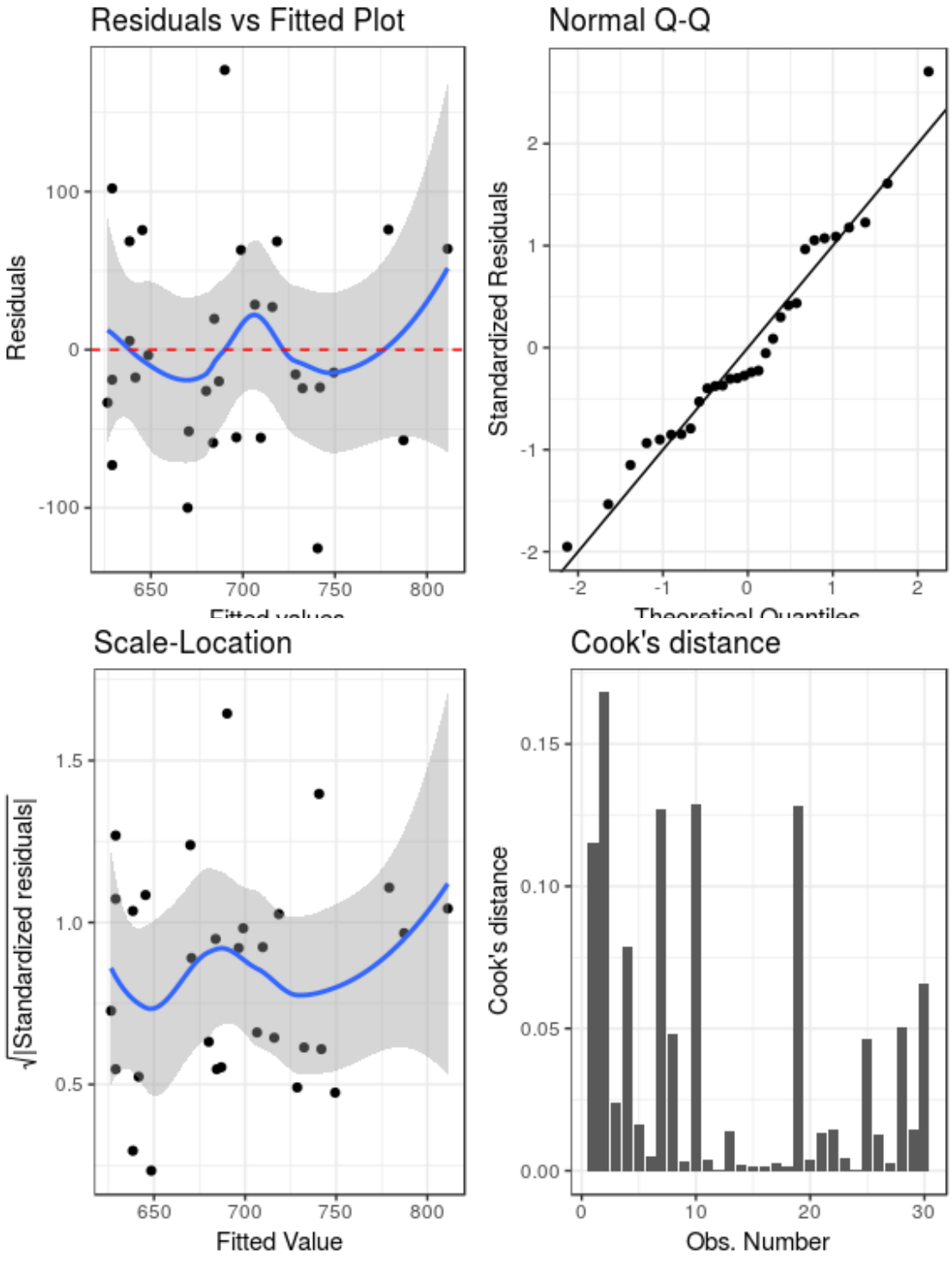
```
library(broom)
broom_modelo <- augment(modelo)
head(broom_modelo, n = 3)
```

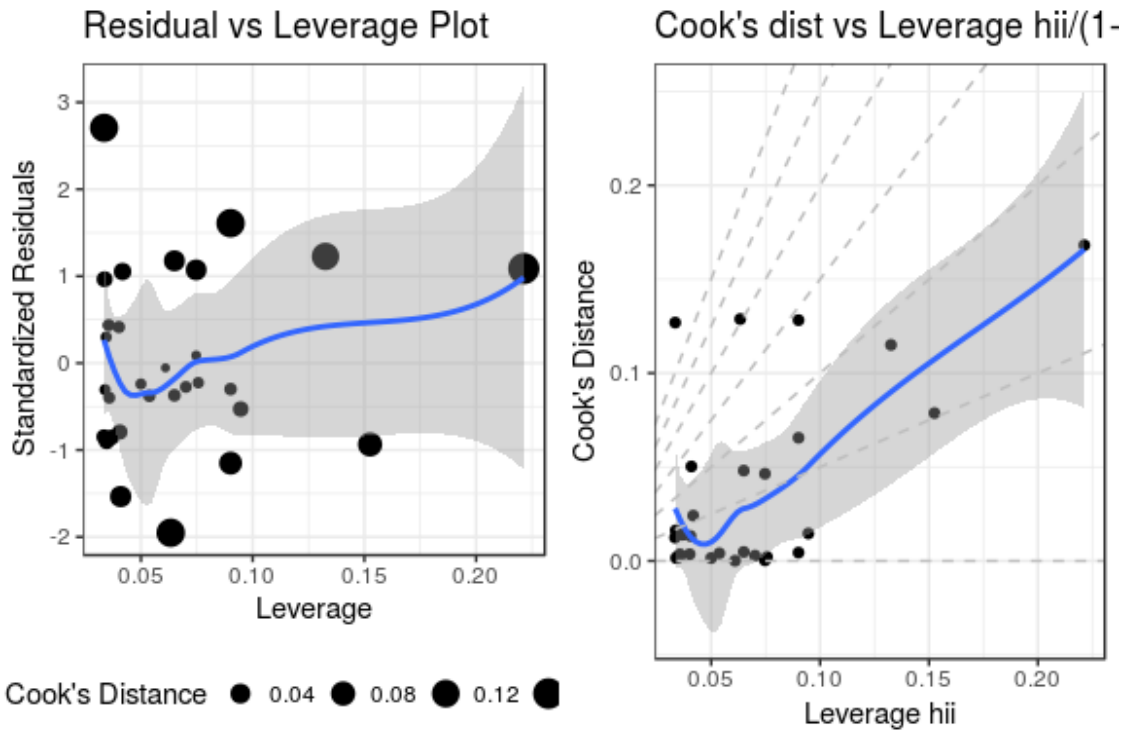
```
##   runs numero_bateos  .fitted  .se.fit   .resid       .hat   .sigma
## 1  855          5659 779.0395 24.20303 75.96048 0.13257176 65.84776
## 2  875          5710 811.1976 31.27290 63.80243 0.22133381 66.24702
## 3  787          5563 718.5067 13.58497 68.49328 0.04176659 66.33977
##      .cooksd .std.resid
## 1 0.11503787   1.226949
## 2 0.16815163   1.087721
## 3 0.02414704   1.052611
```

Se genera un *data frame* que contiene los datos originales más los datos que suelen emplearse de un modelo. Cuando se pasa como argumento a la función `ggplot()` un objeto estadístico tal como *lm* este proceso tiene lugar automáticamente, permitiendo acceder a las nuevas variables (.fitted, .se.fit, .resid, .hat, .sigma, .cooksd .std.resid).

```
diagnostico_residuos <- function(modelo) {
    library(gridExtra)
    p1 <- ggplot(data = modelo, aes(.fitted, .resid)) +
          geom_point()
    p1 <- p1 + stat_smooth(method = "loess") +
               geom_hline(yintercept = 0, col = "red", linetype = "dashed")
    p1 <- p1 + xlab("Fitted values") +
               ylab("Residuals")
    p1 <- p1 + ggtitle("Residuals vs Fitted Plot") +
               theme_bw()

    p2 <- ggplot(modelo, aes(qqnorm(.stdresid, plot.it = FALSE)[[1]], .stdresid)) +
          geom_point(na.rm = TRUE)
    p2 <- p2 + geom_abline() +
               xlab("Theoretical Quantiles") +
               ylab("Standardized Residuals")
    p2 <- p2 + ggtitle("Normal Q-Q") +
               theme_bw()

    p3 <- ggplot(modelo, aes(.fitted, sqrt(abs(.stdresid)))) +
          geom_point(na.rm = TRUE)
    p3 <- p3 + stat_smooth(method = "loess", na.rm = TRUE) +
               xlab("Fitted Value")
    p3 <- p3 + ylab(expression(sqrt("|Standardized residuals|")))
    p3 <- p3 + ggtitle("Scale-Location") +
               theme_bw()
```

```
    p4 <- ggplot(modelo, aes(seq_along(.cooksd), .cooksd)) +
          geom_bar(stat = "identity", position = "identity")
    p4 <- p4 + xlab("Obs. Number") + ylab("Cook's distance")
    p4 <- p4 + ggtitle("Cook's distance") + theme_bw()

    p5 <- ggplot(modelo, aes(.hat, .stdresid)) +
          geom_point(aes(size = .cooksd), na.rm = TRUE)
    p5 <- p5 + stat_smooth(method = "loess", na.rm = TRUE)
    p5 <- p5 + xlab("Leverage") + ylab("Standardized Residuals")
    p5 <- p5 + ggtitle("Residual vs Leverage Plot")
    p5 <- p5 + scale_size_continuous("Cook's Distance", range = c(1, 5))
    p5 <- p5 + theme_bw() + theme(legend.position = "bottom")

    p6 <- ggplot(modelo, aes(.hat, .cooksd)) +
          geom_point(na.rm = TRUE) +
          stat_smooth(method = "loess", na.rm = TRUE)
    p6 <- p6 + xlab("Leverage hii") + ylab("Cook's Distance")
    p6 <- p6 + ggtitle("Cook's dist vs Leverage hii/(1-hii)")
    p6 <- p6 + geom_abline(slope = seq(0, 3, 0.5), color = "gray", linetype =
"dashed")
    p6 <- p6 + theme_bw()

    grid.arrange(p1, p2, ncol = 2)
    grid.arrange(p3, p4, ncol = 2)
    grid.arrange(p5, p6, ncol = 2)
}

diagnostico_residuos(modelo = modelo)
```


Residuals vs Fitted Plot
Residuals
100
0
-100
650
700
750
800
Normal Q-Q
Standardized Residuals
2
1
0
-1
-2
-2
-1
0
1
2
Scale-Location
√|Standardized residuals|
1.5
1.0
0.5
650
700
750
800
Fitted Value
Cook's distance
Cook's distance
0.15
0.10
0.05
0.00
0
10
20
30
Obs. Number



Bibliografía

Introduction to Statistical Learning

OpenIntro Statistics

An introduction to Logistic Regresion Analysis and Reporting. Chao-Ying Joanne Peng

http://www.ats.ucla.edu/stat/r/dae/logit.htm

http://ww2.coastal.edu/kingw/statistics/R-tutorials/index.html

R Tutorials by William B. King, Ph.D http://ww2.coastal.edu/kingw/statistics/R-tutorials/

Points of Significance: Association, correlation and causation. Naomi Altman & Martin Krzywinski Nature Methods

Points of Significance: Simple linear regression Naomi Altman & Martin Krzywinski. Nature Methods

Resampling Data: Using a Statistical Jackknife S. Sawyer | Washington University | March 11, 2005

http://www.biostat.jhsph.edu/~bcaffo/651/files/lecture12.pdf

https://en.wikipedia.org/wiki/Resampling_(statistics)#Jackknife

The Trusty Jackknife Method identifies outliers and bias in statistical estimates by I. Elaine Allen and Christopher A. Seaman