Mayo 2018

Cristina Gil Martínez

REGRESIÓN LINEAL SIMPLE

Apuntes personales sobre regresión lineal simple

CONTENIDO

INTRODUCCIÓN

La **regresión lineal** es un método útil para predecir una respuesta cuantitativa *Y* partiendo de una sola variable predictora *X*, asumiendo que hay una relación aproximadamente lineal entre *X* e *Y*. Matemáticamente, esta relación lineal se representa como

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

donde $\boldsymbol{\beta_0}$ (ordenada en el origen, valor esperado de *Y* cuando *X* = 0) y $\boldsymbol{\beta_1}$ (pendiente, incremento medio de Y asociado con el aumento de *X* en una unidad) son las dos constantes o parámetros desconocidos en el modelo. Asumimos que el residuo o error $\boldsymbol{\epsilon}$ (diferencia entre lo observado y estimado por el modelo) es independiente de *X*.

ESTIMACIÓN DE LOS COEFICIENTES DE REGRESIÓN

La verdadera recta de regresión poblacional suele ser desconocida, pero teniendo acceso a un conjunto de observaciones, podemos calcular un modelo aproximado, teniendo en cuenta que distintos conjuntos de datos pueden tender a generar rectas de regresión ligeramente distintas. Por tanto, en la práctica, β_0 y β_1 son desconocidos, por lo que para poder obtener una predicción de la variable respuesta, tenemos que obtener una estimación de los mismos utilizando los datos de entrenamiento:

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

El objetivo es obtener unos estimadores insesgados con los que el modelo lineal se ajuste bien a los datos disponibles. Para esto, la estrategia más comúnmente utilizada se basa en minimizar la **suma de residuos al cuadrado (RSS),** método conocido como **mínimos cuadrados**:

$$\mathrm{RSS} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

Los **residuos** no son más que la diferencia entre cada valor de la variable respuesta observada y la predicha por el modelo. Algunos residuos serán positivos (para observaciones por encima de la recta) y otros negativos (observaciones por debajo de la recta), siendo su promedio de 0. La recta que se ajuste bien a los datos tendrá residuos pequeños.

Las ecuaciones que minimizan el RSS son:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$
$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

siendo

$$\bar{y} \equiv \frac{1}{n}\sum_{i=1}^{n} y_i \qquad \bar{x} \equiv \frac{1}{n}\sum_{i=1}^{n} x_i$$

las medias muestrales.

Precisión de los coeficientes de regresión

ERROR ESTÁNDAR (SE)

Para estimar cómo de precisos son nuestros estimadores $\hat{\beta}_0$ y $\hat{\beta}_1$ y como, de media, difieren del valor de los verdaderos valores de los parámetros β_0 y β_1, calculamos el error estándar (SE) asociado con $\hat{\beta}_0$ y $\hat{\beta}_1$:

$$\mathrm{SE}(\hat{\beta}_0)^2 = \sigma^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right], \quad \mathrm{SE}(\hat{\beta}_1)^2 = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

donde

$$\sigma^2 = \mathrm{Var}(\epsilon)$$

La varianza del error **ϵ** es en general desconocida, pero se puede estimar a partir de los datos. Esta estimación es conocida como el **error estándar residual (RSE)**, que no es más que la raíz cuadrada de la media de la suma de los residuos al cuadrado:

$$\mathrm{RSE} = \sqrt{\mathrm{RSS}/(n-2)}$$

RSE nos dará una estimación sobre la desviación promedio de cualquier punto respecto a la verdadera recta de regresión, o lo que es lo mismo, estima la desviación estándar de **ϵ**. RSE se divide entre los grados de libertad del modelo *n* – 2 (perdemos dos grados de libertad porque estimamos dos parámetros) para hacer este estimador insesgado.

INTERVALO DE CONFIANZA (CI)

A partir del cálculo del error estándar, podemos obtener los **intervalos de confianza** para cada uno de los estimadores. Un intervalo de confianza del 95% se definiría como el rango de valores tales que con un 95% de probabilidad, dicho rango contendría el verdadero parámetro poblacional desconocido. En el caso de la regresión lineal, este intervalo para β_0 y β_1 toma la siguiente forma:

$$\hat{\beta}_0 \pm t_{df}^{\alpha/2} SE(\hat{\beta}_0)$$

$$\hat{\beta}_1 \pm t_{df}^{\alpha/2} SE(\hat{\beta}_1)$$

TEST DE HIPÓTESIS

El error estándar también puede usarse para llevar a cabo un **test de hipótesis** sobre los parámetros del modelo. El más común establece que

H_0 : $\beta_1 = 0$ (no existe relación entre X e Y)

H_a : $\beta_1 \neq 0$ (existe alguna relación entre X e Y)

Para comprobar la hipótesis nula, necesitamos determinar si $\hat{\beta}_1$ se aleja lo suficientemente de 0. La precisión con la que podemos determinar esto dependerá del $SE(\beta_1)$. Para ello llevamos a cabo un ***t*-test**, calculando el estadístico *t*, el cual mide el número de desviaciones estándar que el estimador $\hat{\beta}_1$ y $\hat{\beta}_0$ están del valor 0, y por último obtenemos el *p-value*:

$$t = \frac{\hat{\beta}_1 - 0}{\mathrm{SE}(\hat{\beta}_1)} \quad t = \frac{\hat{\beta}_0 - 0}{\mathrm{SE}(\hat{\beta}_0)}$$

Si el *p-value* es menor que el nivel de significancia establecido, podemos deducir que hay una relación entre el predictor y la variable respuesta.

BONDAD DE AJUSTE DEL MODELO

En el caso de que la hipótesis nula sea rechazada, podemos cuantificar el grado con el que el modelo se ajusta a los datos. La calidad de un ajuste de regresión lineal es normalmente estimada usando dos valores relacionados: RSE y el estadístico o coeficiente de determinación R^2, como fracción de la varianza explicada.

Error estándar residual (RSE)

El RSE, comentado anteriormente, se considera como una medida absoluta de la falta de ajuste del modelo a los datos, medido en las mismas unidades que *Y*. Cuanto más pequeño sea el valor del RSE, mejor se ajusta el modelo a los datos. Podemos calcular el % de error aproximado que el modelo comete en la predicción dividiendo el valor de RSE entre el valor promedio de la variable respuesta:

$$\%\ desviación = \frac{\text{RSE}}{\text{media de la variable respuesta}}$$

Coeficiente de determinación R^2

El coeficiente de determinación **R^2** constituye una alternativa al RSE. Toma un valor correspondiente a la **proporción** de variabilidad en Y explicada por el modelo en relación a la variabilidad total. Al corresponder a una proporción, tomará valores **entre 0 y 1**. Para calcular esta proporción, utilizamos la fórmula

$$R^2 = \frac{\text{TSS} - \text{RSS}}{\text{TSS}} = 1 - \frac{\text{RSS}}{\text{TSS}}$$

donde

$$\text{TSS} = \sum (y_i - \bar{y})^2$$

mide la varianza total inherente en la variable respuesta *Y* antes de ajustar el modelo lineal, a diferencia del RSS que mide la variabilidad que queda inexplicada tras llevar a cabo la regresión.

Cuanto más próximo sea a 1, mayor será la proporción de variabilidad en la variable respuesta explicada por el modelo. Determinar si su valor es lo suficientemente bueno dependerá de la aplicación en cuestión. Un valor bajo podría indicar que el modelo lineal no es adecuado o podría deberse a errores residuales debido a variables no tenidas en cuenta.

A diferencia del RSE, R^2 es independiente de la escala de medida de *Y*, es decir, **adimensional**, por lo que presenta la ventaja de ser más fácil de interpretar.

También podríamos decir que R^2 es una medida de la relación linear entre *X* e *Y*, al igual que el **coeficiente de correlación de Pearson (r)**, definida como

$$\mathrm{Cor}(X,Y) = \frac{\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\overline{x})^2}\sqrt{\sum_{i=1}^{n}(y_i-\overline{y})^2}}$$

Únicamente en el caso de regresión lineal simple, se cumple que

$$R^2 = r^2$$

CONDICIONES PARA LA REGRESIÓN LINEAL

1. **Linealidad**: la relación entre el predictor y la variable respuesta ha de ser lineal.

   ¿Cómo comprobarlo? **Diagrama de dispersión** de los datos**, graficar los residuos.**

2. **Distribución normal de los residuos**: los residuos deben distribuirse de forma normal, en torno a 0. Esta condición podría no cumplirse en presencia de observaciones atípicas o *outliers* que no siguen el patrón del conjunto de datos.

   ¿Cómo comprobarlo? **Histograma** de los datos**, distribución de quantiles** (normal Q-Q plot), **test de hipótesis de normalidad**

   En la página [https://gallery.shinyapps.io/slr_diag/] se muestran ejemplos visuales intuitivos.

3. **Variabilidad de los residuos constante** (homocedasticidad): la variabilidad de los datos en torno a la recta de regresión ha de ser aproximadamente constante, lo cual implica que la variabilidad de los residuos debería ser constante también en torno a 0.

   ¿Cómo comprobarlo? **Graficar los residuos**, **test de *Breusch-Pagan***

4. **Independencia**: las observaciones han de ser independientes unas de otras. Tener en cuenta en el caso de mediciones temporales.

   ¿Cómo comprobarlo? **Graficar los residuos** y estudiar si siguen un patrón o tendencia.

EJEMPLO EN R

En este ejemplo trabajaremos con el set de datos *trees*, que contiene datos sobre la circunferencia (en pulgadas), altura (en pies) y volumen (en pies cúbicos) del tronco de árboles de cerezos. Intentaremos ajustar un modelo de regresión lineal simple para predecir el volumen en función del diámetro.

1. Análisis exploratorio de los datos

- **head**(datos) -> Inspeccionar primeras observaciones del set de datos
- **glimpse**(datos) -> Estructura del set de datos (parecido a str())
- **summary**(datos) -> Resumen del set de datos (datos numéricos de las variables)
- **pairs**(datos) -> Gráficos dispersión del conjunto de variables
- **cor.test()** -> Test de correlación ("pearson", "kendall", "spearman")
- **ggpairs()** -> Combina en un único gráfico diagramas de dispersión, distribución de las variables y los valores de correlación.
- **ggplot()**
- **plot()**

Antes de generar el modelo, representaremos los datos para cuantificar la posible relación lineal entre variables y su magnitud. Si no detectáramos esta relación pasaríamos a plantearnos métodos de ajuste alternativos.

```
library(dplyr)
head(trees)
```

```
##   Girth Height Volume
## 1   8.3     70   10.3
## 2   8.6     65   10.3
## 3   8.8     63   10.2
## 4  10.5     72   16.4
## 5  10.7     81   18.8
## 6  10.8     83   19.7
```

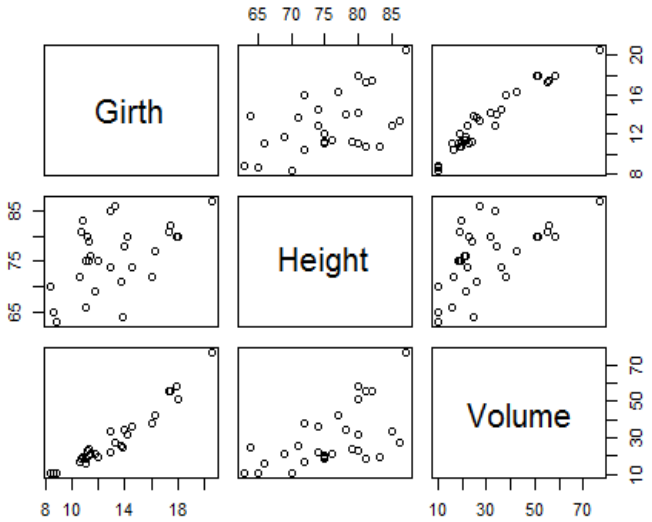
```
glimpse(trees)
```

```
## Observations: 31
## Variables: 3
## $ Girth  <dbl> 8.3, 8.6, 8.8, 10.5, 10.7, 10.8, 11.0, 11.0, 11.1, 11.2, 11....
## $ Height <dbl> 70, 65, 63, 72, 81, 83, 66, 75, 80, 75, 79, 76, 76, 69, 75, ...
## $ Volume <dbl> 10.3, 10.3, 10.2, 16.4, 18.8, 19.7, 15.6, 18.2, 22.6, 19.9, ...
```

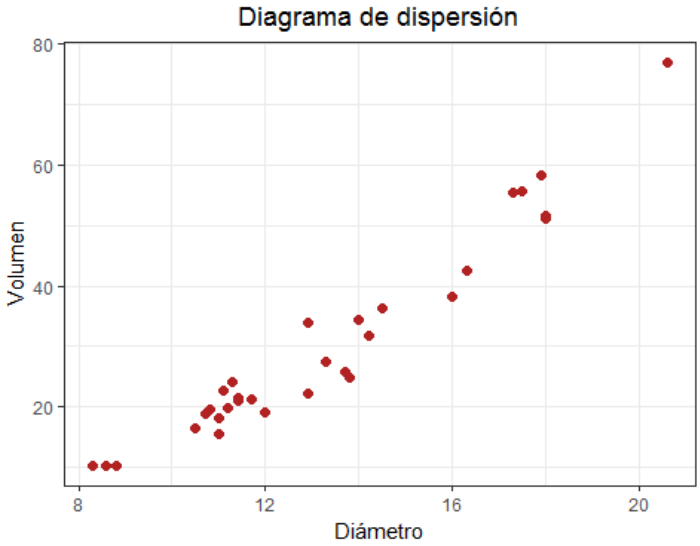
```r
summary(trees)
```

```
##      Girth           Height       Volume
##  Min.   : 8.30   Min.   :63   Min.   :10.20
##  1st Qu.:11.05   1st Qu.:72   1st Qu.:19.40
##  Median :12.90   Median :76   Median :24.20
##  Mean   :13.25   Mean   :76   Mean   :30.17
##  3rd Qu.:15.25   3rd Qu.:80   3rd Qu.:37.30
##  Max.   :20.60   Max.   :87   Max.   :77.00
```

```r
pairs(trees)
```



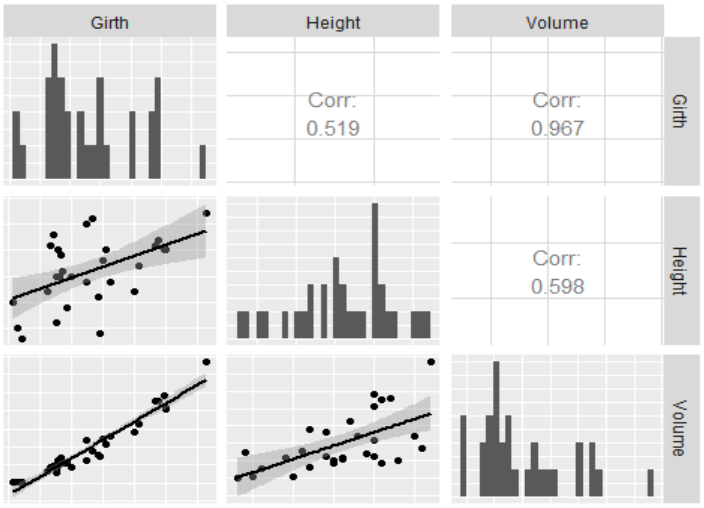

```r
library(ggplot2)

ggplot(data = trees, mapping = aes(x = Girth, y = Volume)) +
geom_point(color = "firebrick", size = 2) +
labs(title = "Diagrama de dispersión", x = "Diámetro", y = "Volumen") +
theme_bw() +
theme(plot.title = element_text(hjust = 0.5))
```



```r
cor.test(x = trees$Girth, y = trees$Volume, method = "pearson", digits = 3)
```

```
##
##  Pearson's product-moment correlation
##
## data:  trees$Girth and trees$Volume
## t = 20.478, df = 29, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.9322519 0.9841887
## sample estimates:
##       cor
## 0.9671194
```

```r
library(GGally)
ggpairs(trees, lower = list(continuous = "smooth"),
        diag = list(continuous = "bar"), axisLabels = "none")
```



De lo hasta ahora analizado podemos concluir que:

I. Observando los gráficos de dispersión podemos observar que la variable *Girth* (diámetro) está más linealmente asociada con la variable respuesta *Volume*, por lo que utilizaremos ésta para el modelo.

II. El coeficiente de correlación de Pearson es bastante alto (r = 0,967) y significativo (p-*value* = 2,2x10^{-16}). Ello indica una correlación entre ambas variables bastante intensa.

III. Tiene sentido generar el modelo de regresión lineal cumpliéndose estos primeros requisitos.

2. Cálculo del modelo de regresión lineal simple

- **lm(y ~ x, data)** -> Modelo de regresión lineal simple
- **summary**(modelo) -> Resumen del ajuste del modelo

```r
modelo.lineal <- lm(Volume ~ Girth, data = trees)


# Información del modelo
summary(modelo.lineal)
```

```
##
## Call:
## lm(formula = Volume ~ Girth, data = trees)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -8.065 -3.107  0.152  3.495  9.587
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -36.9435     3.3651  -10.98 7.62e-12 ***
## Girth         5.0659     0.2474   20.48  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.252 on 29 degrees of freedom
## Multiple R-squared:  0.9353, Adjusted R-squared:  0.9331
## F-statistic: 419.4 on 1 and 29 DF,  p-value: < 2.2e-16
```

Con la función *summary ( )* hemos obtenido los errores estándar de los coeficientes, los *p-values* así como el estadístico *F* y R^2. 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.

```r
names(modelo.lineal)
```

```
##  [1] "coefficients"  "residuals"     "effects"       "rank"
##  [5] "fitted.values" "assign"        "qr"            "df.residual"
##  [9] "xlevels"       "call"          "terms"         "model"
```

El *summary* del modelo contiene los errores estándar, el valor del estadístico *t* y el correspondiente *p-value* de ambos parámetros $\hat{\beta}_0$ y $\hat{\beta}_1$. El *p-value* nos permite determinar si los estimadores de los parámetros son significativamente distintos de 0, es decir, que contribuyen al modelo. El parámetro que suele ser más útil en estos modelos es la pendiente.

Conclusiones del modelo:

I. Tanto la ordenada en el origen como la pendiente son significativas (*p-value* = $2x10^{-16}$).
II. El coeficiente de determinación R^2 indica que el modelo es capaz de explicar el 93% de la variabilidad presente en la variable respuesta (volumen) mediante la variable independiente (diámetro).

III. El *p-value* obtenido en el test *F* (*p-value* = $2x10^{-16}$) determina que sí es significativamente superior la varianza explicada por el modelo en comparación con la varianza total, por lo que podemos aceptar nuestro modelo como válido y útil.

IV. Ecuación del modelo: *Volume* = - 36,94 + 5,065*Girth* -> por cada unidad que se incrementa el diámetro, el volumen aumenta en promedio 5,065 unidades.

3. Intervalo de confianza para los parámetros del modelo

- **confint(**modelo**)** -> Intervalos de confianza para los coeficientes del modelo

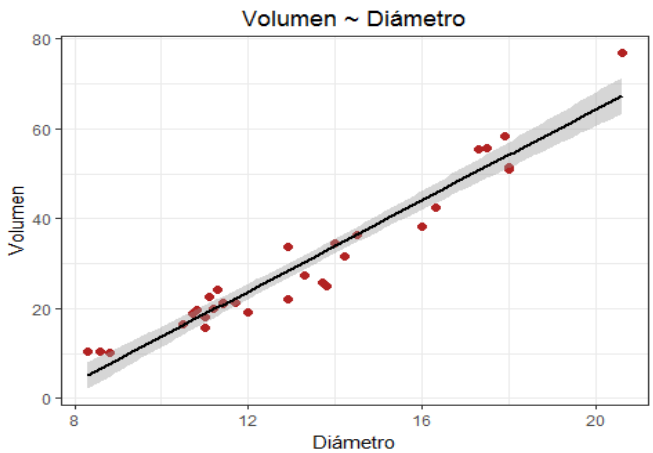
```
confint(modelo.lineal)
```

```
##                  2.5 %     97.5 %
## (Intercept) -43.825953 -30.060965
## Girth         4.559914   5.571799
```

4. Representación gráfica del modelo

Representamos la línea de mínimos cuadrados, representaremos el intervalo de confianza (límites superior e inferior) para cada predicción con la función *geom_smooth( )* del paquete *ggplot2*. Esto permite identificar la región en la que se encuentra el promedio de la variable respuesta según el modelo generado y para un determinado nivel de confianza:

```
ggplot(data = trees, mapping = aes(x = Girth, y = Volume)) +
geom_point(color = "firebrick", size = 2) +
geom_smooth(method = "lm", se = TRUE, color = "black") +
labs(title = "Volumen ~ Diámetro", x = "Diámetro", y = "Volumen") +
theme_bw() + theme(plot.title = element_text(hjust = 0.5))
```

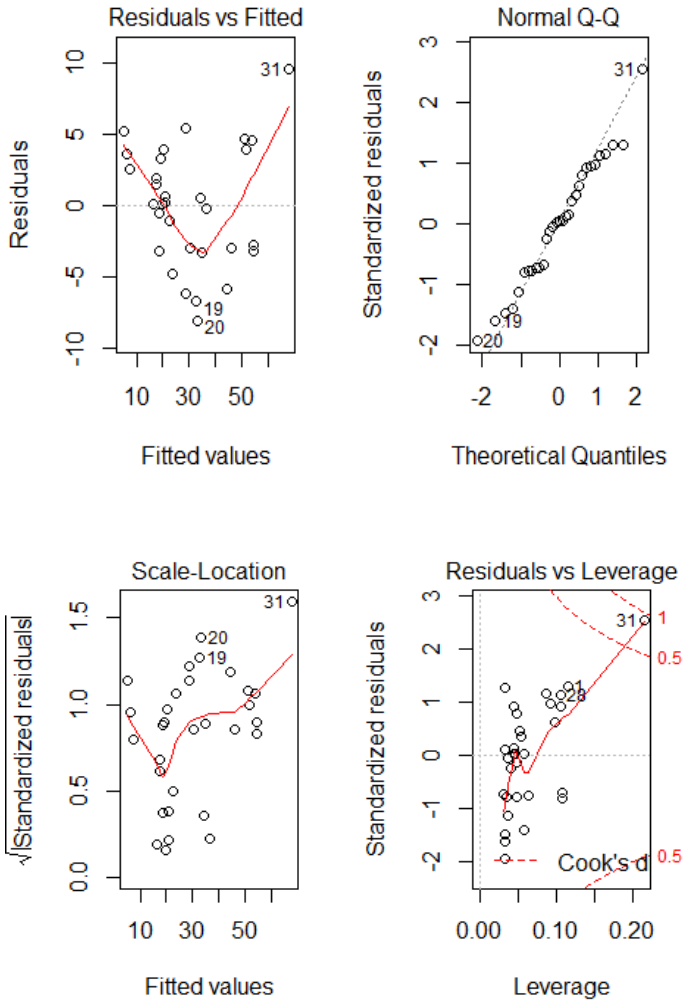


5. Verificar condiciones para aceptar el modelo

- **plot(**modelo**)** -> Análisis de los residuos **(**distribución, variabilidad...**)**
- **shapiro.test(**modelo$residuals**)** -> Test de hipótesis de Shapiro Wilk para el análisis de normalidad
- **bptest(**modelo**)** -> Test de contraste de homocedasticidad Breusch-Pagan
- **influence.measures**(modelo) -> Detección de observaciones influyentes
- **influencePlot(**modelo**)** -> Visualización de observaciones influyentes
- **outlierTest**(modelo) -> Test de detección de outliers
- **rstudent(modelo)** -> Cálculo de residuos estudentizados

Para evaluar las condiciones que permiten dar como válido el modelo lineal, haremos uso principalmente del análisis de los residuos:

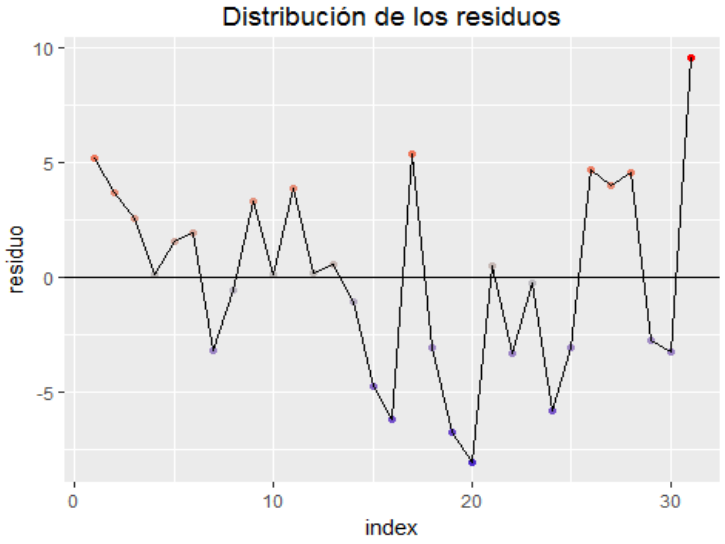
```
par(mfrow=c(1,2))
plot(modelo.lineal)
```



```r
# Contraste de hipótesis (normalidad de los residuos)
shapiro.test(modelo.lineal$residuals)

##
##  Shapiro-Wilk normality test
##
## data:  modelo.lineal$residuals
## W = 0.97889, p-value = 0.7811
```

```r
# Test de Breush-Pagan (homocedasticidad de los residuos)
library(lmtest)
bptest(modelo.lineal)

##
##  studentized Breusch-Pagan test
##
## data:  modelo.lineal
## BP = 5.6197, df = 1, p-value = 0.01776
```

```r
# Análisis gráfico autocorrelación de los residuos
ggplot(data = trees, aes(x = seq_along(modelo.lineal$residuals),
                         y = modelo.lineal$residuals)) +
geom_point(aes(color = modelo.lineal$residuals)) +
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(plot.title = element_text(hjust = 0.5), legend.position = "none")
```

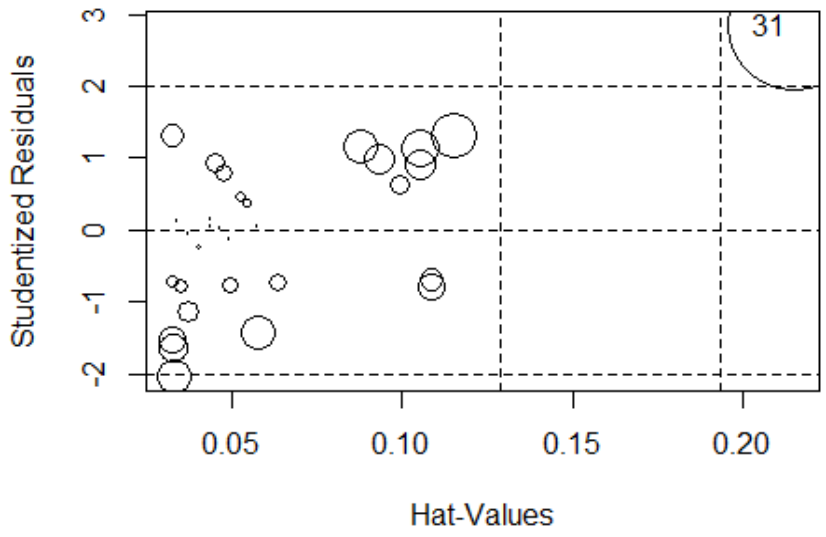


En este caso, la normalidad de los residuos podemos aceptarla, y tampoco parecen seguir una clara tendencia según el orden de registro de las observaciones, pero la condición de homocedasticidad parece

no cumplirse. Al observar los gráficos podríamos sospechar que la observación 31 podría estar influyendo al modelo. Para analizar en qué medida pueda estar influyendo esta u otras observaciones, reajustaremos el modelo excluyendo posibles observaciones sospechosas. Dependiendo de la finalidad del modelo, la exclusión de posibles *outliers* debe analizarse con detalles, ya que estas observaciones podrían ser errores de medida, pero también podrían representar casos interesantes.

```
# Residuos estudentizados
studentized_residual <- rstudent(modelo.lineal)
which(abs(studentized_residual) > 3)
## named integer(0)
```

```
library(car)
summary(influence.measures(model = modelo.lineal))

## Potentially influential observations of
##   lm(formula = Volume ~ Girth, data = trees) :
##
##    dfb.1_  dfb.Grth dffit   cov.r cook.d  hat
## 31 -1.20_*  1.37_*   1.49_*  0.82  0.89_*  0.22_*

influencePlot(model = modelo.lineal)
```

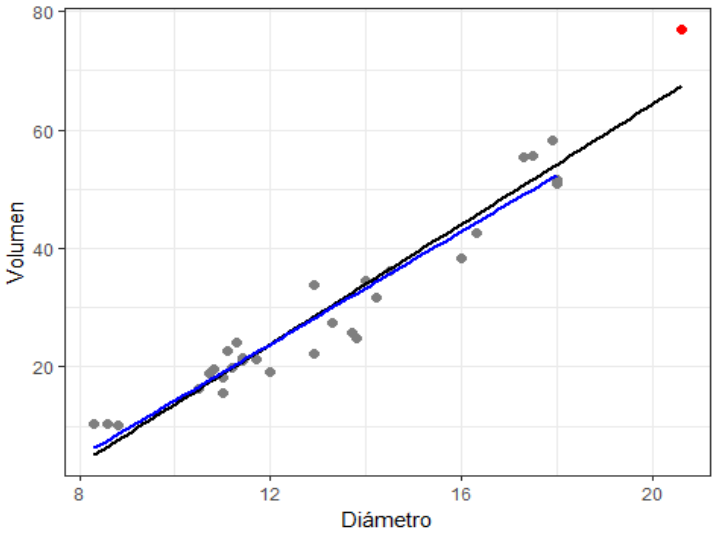



```
##     StudRes       Hat     CookD
## 31 2.837732 0.2151943 0.8880581
```

El análisis de los residuos estudentizados no detecta ninguna observación atípica. Sin embargo, tal y como sospechábamos, la observación 31 parece estar influenciando al modelo.

Procederemos a reajustar el modelo excluyendo la observación 31:

```r
ggplot(data = trees, mapping = aes(x = Girth, y = Volume)) +
geom_point(color = "grey50", size = 2) +
#recta de regresión con todas las observaciones
geom_smooth(method = "lm", se = FALSE, color = "black") +
#se resalta el valor excluido
geom_point(data = trees[31, ], color = "red", size = 2) +
#se añade la nueva recta de regresión
geom_smooth(data = trees[-31, ], method="lm", se =FALSE,color = "blue") +
labs(x = "Diámetro", y = "Volumen") +
theme_bw() + theme(plot.title = element_text(hjust = 0.5), legend.position = "none
")
```




```r
modelo.lineal2 <- lm(Volume ~ Girth, data = trees[-31,])


summary(modelo.lineal2)

##
## Call:
## lm(formula = Volume ~ Girth, data = trees[-31, ])
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.5036 -2.3834 -0.0489  2.3951  6.3726
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -33.3104     3.2784  -10.16 6.76e-11 ***
## Girth         4.7619     0.2464   19.33  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.813 on 28 degrees of freedom
## Multiple R-squared:  0.9303, Adjusted R-squared:  0.9278
## F-statistic: 373.6 on 1 and 28 DF,  p-value: < 2.2e-16
```

La eliminación de la observación identificada como influyente a penas cambia la recta de mínimos cuadrados.

6. Uso del modelo para predecir nuevas observaciones

- **predict()** -> Predicciones de nuevas observaciones a partir del modelo

Dando por válido el modelo podemos utilizar la función genérica *predict( )* para predecir el valor de nuevas observaciones. En nuestro ejemplo, podemos predecir el volumen a partir de mediciones del diámetro de nuevos árboles:

```
# Volumen PROMEDIO que esperaríamos de árboles de 15 pulgadas
predict(modelo.lineal, data.frame(Girth = 15), interval = "confidence")
```

```
##        fit      lwr      upr
## 1 39.04439 37.24858 40.84019
```

```
# Volumen esperado de UN árbol de 15 pulgadas
predict(modelo.lineal, data.frame(Girth = 15), interval = "prediction")
```

```
##        fit      lwr      upr
## 1 39.04439 30.16461 47.92416
```

Podemos interpretar la predicción de la siguiente manera: se espera que en promedio los árboles de diámetro 15 pulgadas tengan un volumen de 39,04 pies cúbicos. Podemos decir con un 95% de confianza que el verdadero valor de este promedio se encuentra entre (37,24 – 40,08), mientras que el intervalo de predicción para un solo árbol de este diámetro es de (30,16 – 47,92). Como es de esperar, el intervalo de predicción es mayor que el de confianza.

BIBLIOGRAFÍA

OpenIntro Statistics: Third Edition, David M Diez, Christopher D Barr, Mine Çetinkaya-Rundel

An Introduction to Statistical Learning: with Applications in R (Springer Texts in Statistics)