Regresión múltiple

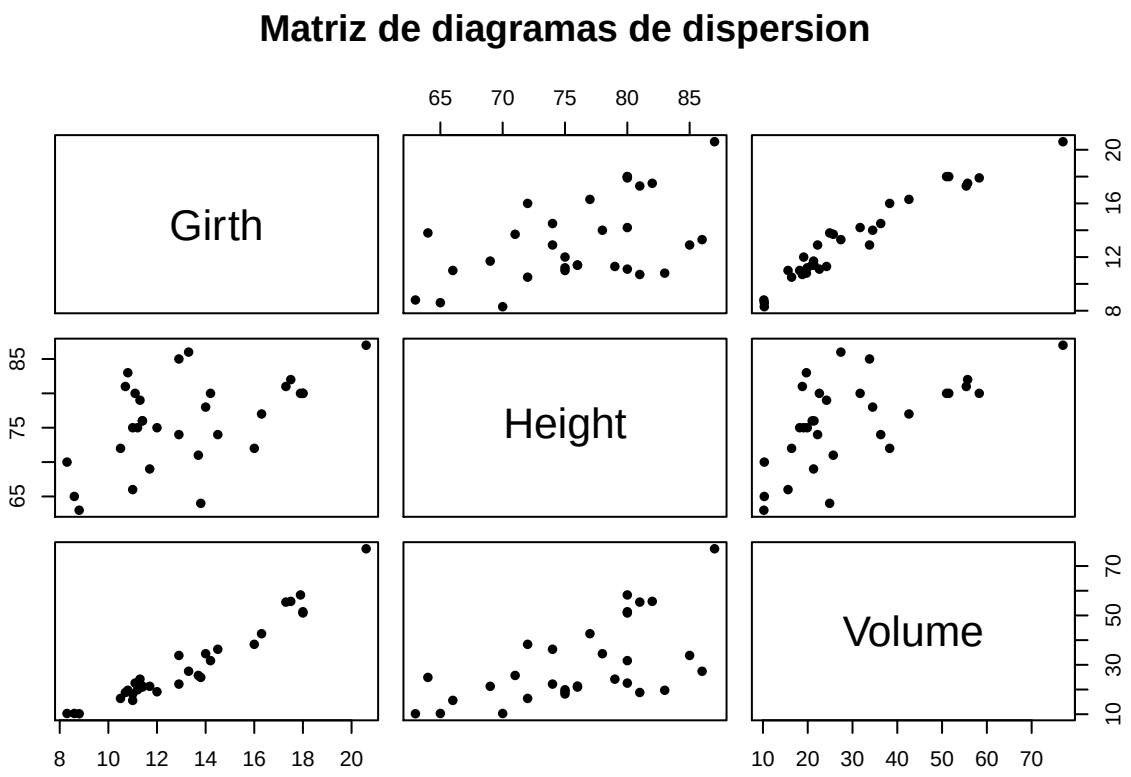
Conjunto de datos

Vamos a utilizar el conjunto de datos `trees` (ya incluido en R) que contiene medidas tomadas en 31 cerezos. Para cada uno de ellos se mide el volumen de la madera que se obtuvo de ellos, la altura y el diámetro (medido a cierta altura del suelo). Los nombres de las variables son:

- `Girth`: Diámetro en pulgadas.
- `Height`: Altura en pies.
- `Volume`: Volumen de madera en pies cúbicos.

El objetivo es explicar el volumen de madera obtenido en función de las otras dos variables. Para tener una primera idea de cómo son las relaciones entre las variables, podemos representar una matriz de diagramas de dispersión:

```
pairs(trees, pch=16, main = 'Matriz de diagramas de dispersion')
```



¿Cuáles de las hipótesis habituales del modelo de regresión lineal no parece que se vayan a cumplir? Compara el gráfico anterior con el que se obtiene cuando consideramos los logaritmos de las tres variables en lugar de las variables originales.

Ajuste del modelo

Si h es la altura, D el diámetro del árbol y V es el volumen, debemos esperar $V \approx \pi h D^2/4$. La relación anterior sugiere ajustar el modelo de regresión múltiple:

$$\log V = \beta_0 + \beta_1 \log h + \beta_2 \log D + \epsilon$$

Ajustamos el modelo y obtenemos un resumen de los resultados:

```
reg = lm(log(Volume) ~ log(Height) + log(Girth), data = trees)
summary(reg)
```

```
##
## Call:
## lm(formula = log(Volume) ~ log(Height) + log(Girth), data = trees)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.16856 -0.04849  0.00243  0.06364  0.12922
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -6.632      0.800   -8.29  5.1e-09 ***
## log(Height)    1.117      0.204    5.46  7.8e-06 ***
## log(Girth)     1.983      0.075   26.43  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.0814 on 28 degrees of freedom
## Multiple R-squared:  0.978,  Adjusted R-squared:  0.976
## F-statistic:  613 on 2 and 28 DF,  p-value: <2e-16
```

La tabla de análisis de la varianza del modelo se obtiene de la siguiente forma:

```
anova(reg)
```

```
## Analysis of Variance Table
##
## Response: log(Volume)
##             Df Sum Sq Mean Sq F value Pr(>F)
## log(Height)  1   3.50    3.50     528 <2e-16 ***
## log(Girth)   1   4.63    4.63     699 <2e-16 ***
## Residuals   28   0.19    0.01
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

La matriz de covarianzas estimadas del vector $\hat{\beta}$ (es decir, $s_R^2(X'X)^{-1}$, donde X es la matriz de diseño y s_R^2 es la vaianza residual) se obtiene con:

```
vcov(reg)
```

```
##             (Intercept) log(Height) log(Girth)
## (Intercept)     0.63966   -0.160062   0.020794
## log(Height)    -0.16006    0.041795  -0.008131
## log(Girth)      0.02079   -0.008131   0.005627
```

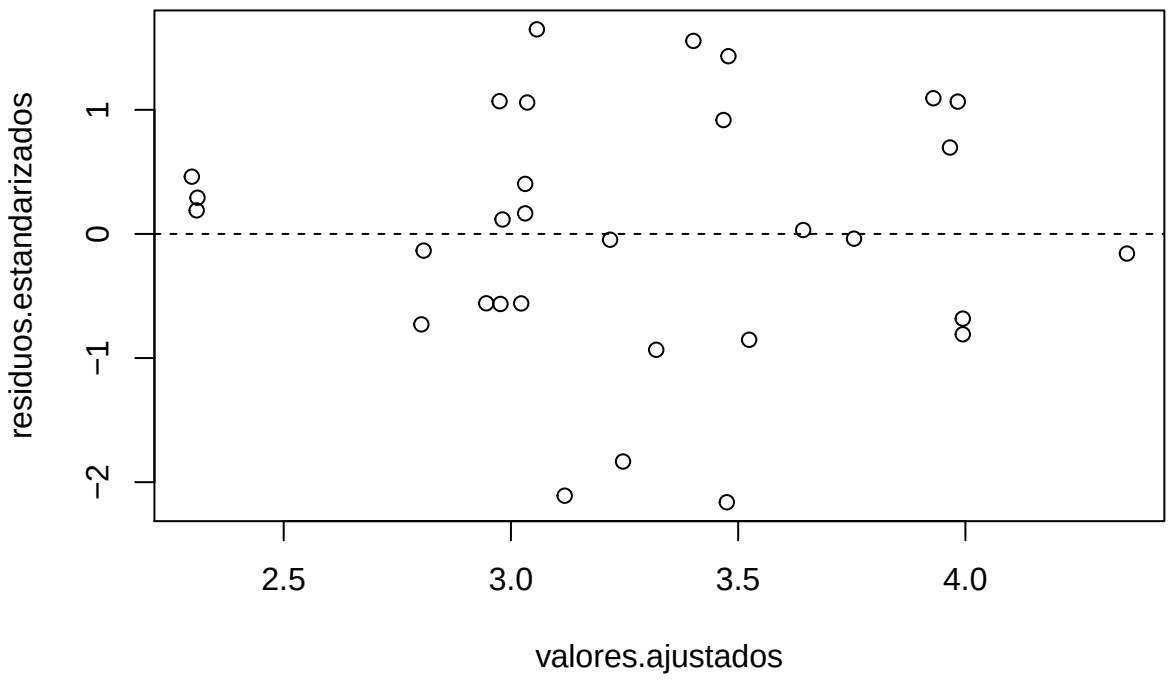
Ejercicios

1. Contrasta la hipótesis $H_0 : \beta_1 = \beta_2 = 0$ a nivel 0.01.
2. ¿Qué valores es de esperar que tomen β_1 y β_2? ¿Se parecen los estimadores obtenidos a estos valores?
3. Calcula intervalos de confianza de nivel 0.9 para β_1 y β_2 (recuerda el comando `confint`).
4. Contrasta, a nivel 0.9, que el valor de β_2 coincide con su valor esperado según el ejercicio 2.
5. Calcula la suma de cuadrados explicada por la regresión y la correspondiente media de cuadrados. ¿Cuántos grados de libertad le corresponden?
6. Calcula la matriz de correlaciones del vector $\hat{\beta}$.
7. Contrasta $H_0 : \beta_0 = 0, \beta_2 = 2\beta_1$ (sumultáneamente) mediante el método de *incremento relativo de la variabilidad*.

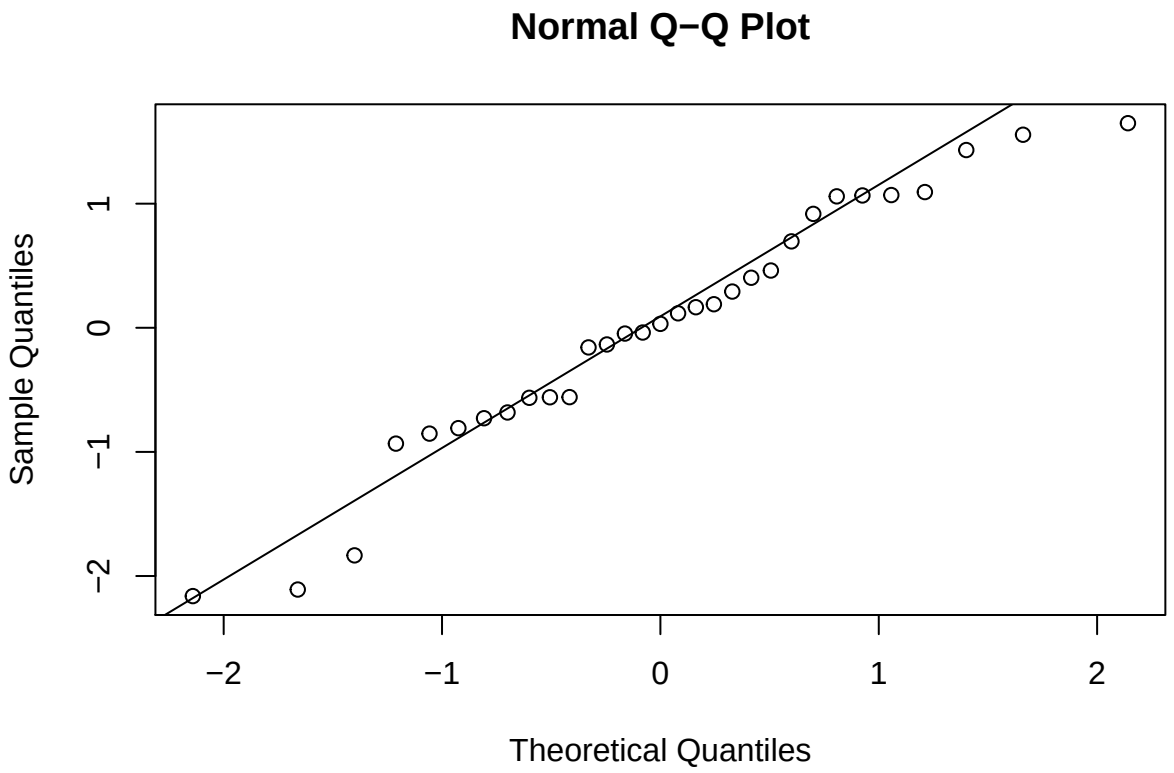
Diagnóstico del modelo

Los comandos siguientes producen los principales gráficos de residuos (estandarizados):

```
# residuos frente a valores ajustados
residuos.estandarizados <- rstandard(reg)
valores.ajustados <- fitted(reg)
plot(valores.ajustados, residuos.estandarizados)
abline(h=0, lty=2)  # añade recta y=0
```

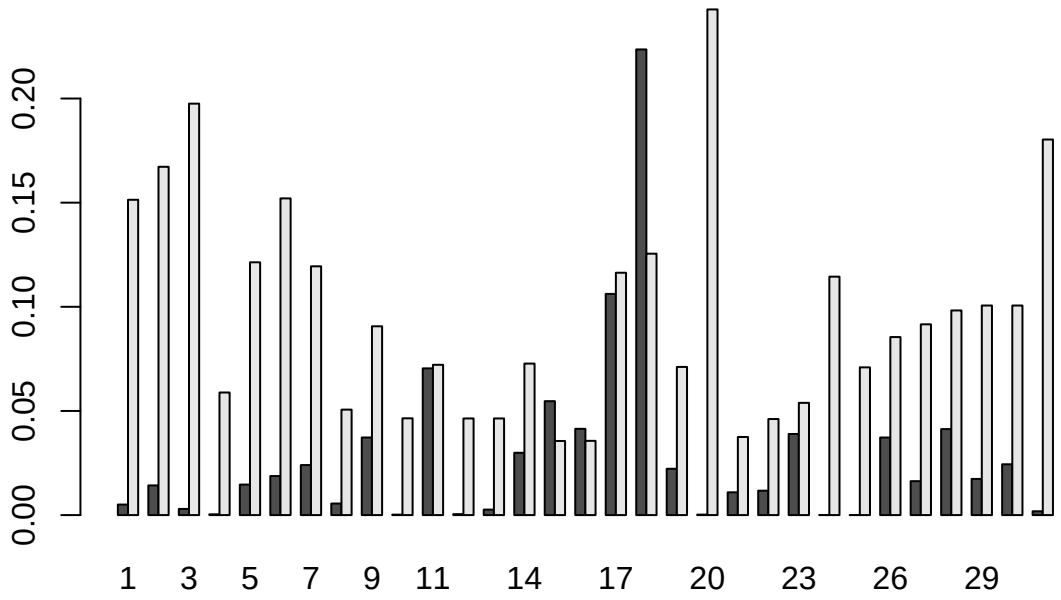


```r
# gráficos de probabilidad
qqnorm(residuos.estandarizados)
qqline(residuos.estandarizados)
```



Para analizar la influencia de cada punto en el ajuste se pueden calcular las distancias de Cook, o los elementos de la diagonal de la matriz H:

```r
cook <- cooks.distance(reg)
hii <- hatvalues(reg)
barplot(rbind(cook, hii), beside = TRUE)
```



Ejercicios

1. ¿Qué se puede decir sobre si se cumplen o no las hipótesis habituales del modelo de regresión?
2. Elimina la observación cuya distancia de Cook es máxima y ajusta de nuevo el modelo. ¿Hay mucha disferencia con los resultados obtenidos anteriormente?
3. ¿Cómo interpretas la existencia de puntos para los que h_{ii} toma valores altos y, simultáneamente, la distancia de Cook es pequeña?