Modelos de regresión lineal

David García Callejas

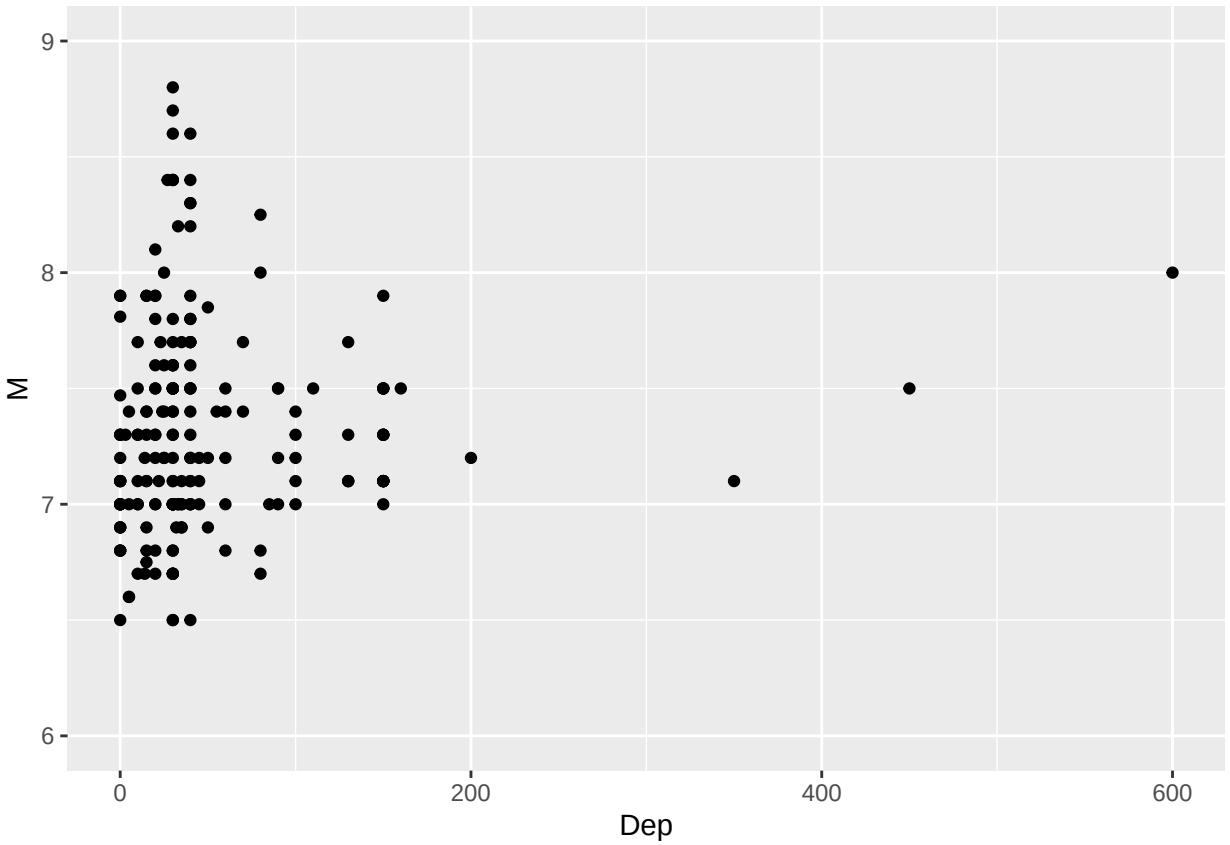
R es un lenguaje pensado para el tratamiento de datos (lo que hemos venido aprendiendo hasta ahora) y para realizar análisis estadísticos. En este curso no vamos a aprender estadística, pero es importante que sepáis los fundamentos de los modelos estadísticos en R.

El modelo estadístico más sencillo, que veremos en esta sección, es la regresión lineal entre dos variables. Este análisis nos dice si la relación entre dos variables es significativa o no. Es decir, si es una relación que se puede distinguir de lo que esperaríamos por azar. Podéis imaginar infinitas preguntas que se pueden responder con análisis estadísticos de este tipo: ¿está relacionada la riqueza per cápita de un país con su esperanza de vida? ¿hay relación entre la cantidad de territorio protegido y la biodiversidad de una región?

Aunque no veremos los fundamentos estadísticos de estos análisis, tened cuidado... estos tests lo que muestran es si hay relación estadística o no entre dos variables. Pero una frase muy usada en estadística ya nos advierte: "correlación no implica causalidad"... https://www.tylervigen.com/spurious-correlations

```
# para el primer ejemplo, veremos la relación entre la
# profundidad y la magnitud de nuestros datos de terremotos
eq <- read.csv2("../data/Earthquake_data.csv",
                header = TRUE,dec = ".",stringsAsFactors = FALSE)
head(eq)

# en la siguiente sesión aprenderemos a visualizar datos.
# Por ahora, fijaos en el resultado de esto
library(tidyverse)
eq %>% ggplot(aes(x = Dep, y = M)) +
  geom_point()
```



Los modelos de regresión lineal (linear regression models) relacionan una variable Y con una o más variables X1, X2, etc. Estas relaciones se codifican en R de la siguiente manera:

$$Y \sim X1 + X2$$

Lo que viene a decir que nuestro modelo explica Y en función de X1 y en función de X2. La función para calcular este tipo de modelos en R es `lm`. Esta función tiene como argumentos principales una fórmula, como la que he mostrado, y el conjunto de datos con el que queremos trabajar.

```
# ¿cómo funciona la función lm?
?lm

# nuestra primera regresión lineal!
# ¿viene la magnitud de un terremoto explicada por su profundidad?
m1 <- lm(formula = M ~ Dep, data = eq)

# hemos guardado los resultados del modelo en una variable, m1. ¿Y ahora?
m1
# ¿qué tipo de objeto es m1?
str(m1)
# mucha información, pero la clave es la primera línea...
# el resultado de llamar a la función lm es una lista!
# podemos obtener un resumen del modelo con la función "summary"
summary(m1)
```

El resumen de un modelo (summary) nos da información básica sobre el mismo. Primero, la fórmula que hemos usado. Segundo, una distribución de los residuos (es decir, el error de nuestro modelo). Tercero, los coeficientes del modelo lineal. Estos coeficientes se corresponden a la ecuación del modelo:

$$y_i = \beta_0 + \beta_1 x_1$$

Donde y_i es nuestra variable respuesta (la magnitud de un terremoto determinado), y β_0 es el término independiente (intercept en inglés), el término del modelo que no depende de ninguna variable independiente. En términos matemáticos, es el punto de corte de la recta con el eje vertical. Siguiendo con nuestro ejemplo, β_1 es el coeficiente que indica el efecto de x_1, la profundidad ("Dep", nuestra variable explicativa) sobre la magnitud (M, nuestra variable respuesta).

Al calcular un modelo estadístico, R calcula los coeficientes que mejor ajustan una recta que relaciona los valores de profundidad y magnitud. Estos coeficientes son los "Estimate" de la tabla resumen. Ya que el resultado del modelo lineal viene en una lista, podemos recuperar estos coeficientes fácilmente

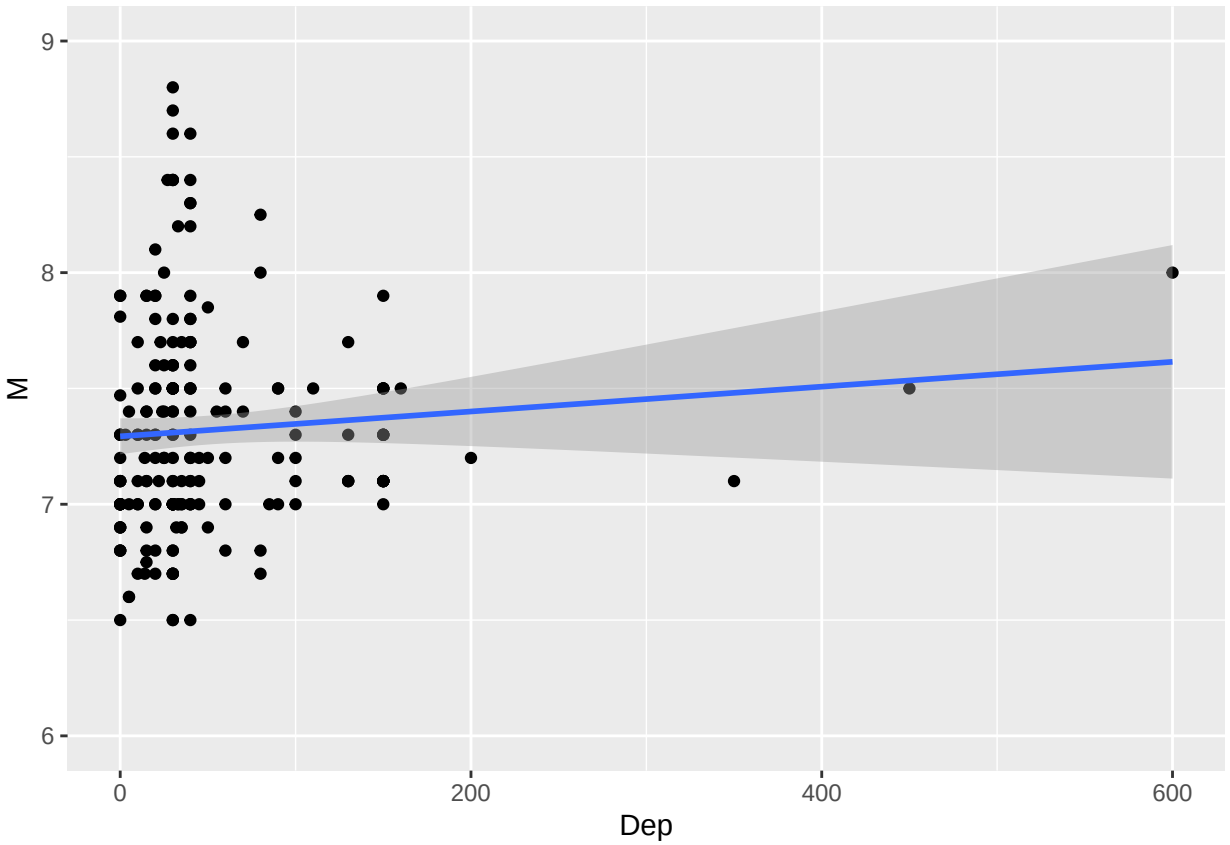
```
coefs <- m1$coefficients
coefs
```

Estos coeficientes tienen un error asociado (segunda columna), y un grado de significación. El grado de significación nos indica, simplificando mucho, si la relación es diferente o no de lo que esperaríamos por azar. Esto viene dado por la última columna. Cuanto menor sea el número de la última columna (el conocido como "p-valor"), más fuerte es la relación. De manera general, para valores de 0.05 o menores aceptamos que una relación entre dos variables puede tener sentido estadístico.

Nuestro modelo nos dice que la relación entre profundidad y magnitud en los terremotos registrados no es significativa (tiene un p-valor de 0.247, bastante mayor de 0.05).

Podemos ver cómo es la recta que predice el modelo (profundizaremos en esto en la sesión sobre visualización). Una imagen vale más que mil palabras: realmente la relación casi no existe.

```
ggplot(eq, aes(x = Dep, y = M)) +
  geom_point() +
  geom_smooth(method = "lm")
```



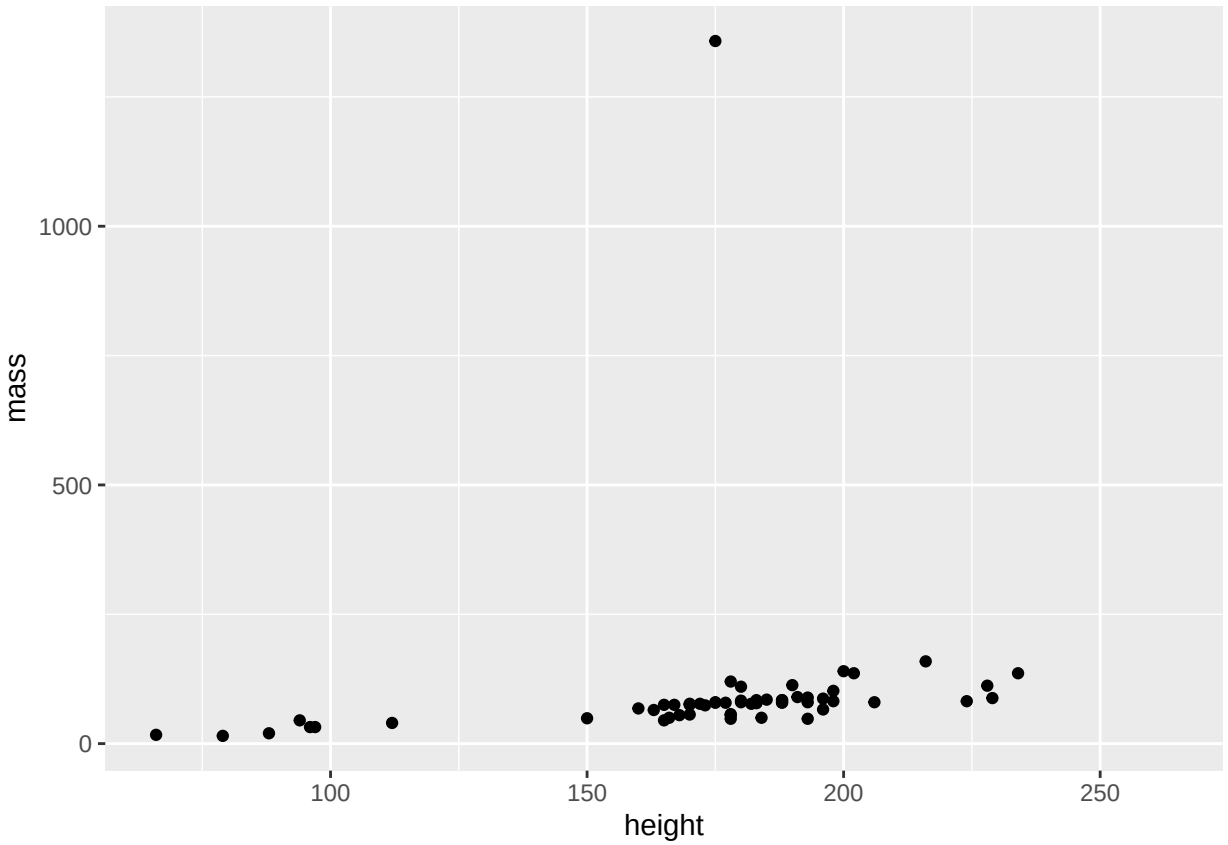
Veamos más ejemplos. ¿Tiene relación la altura de una persona con su peso? ... ¿y si esas personas no son necesariamente humanas?

```
# información sobre personajes de Star Wars
personajes_SW <- read.csv2(file = "../data/starwars_info_personajes.csv",
                           header = TRUE,
                           stringsAsFactors = FALSE)
head(personajes_SW)

# tenemos altura (height) y peso (mass) de cada personaje
sw1 <- lm(formula = mass ~ height, data = personajes_SW)
summary(sw1)
```

La relación tampoco es significativa. ¿Podemos entender algo más viendo los datos?

```
ggplot(personajes_SW, aes(x = height, y = mass)) +
  geom_point()
```



Parece que hay un único punto que rompe la tendencia. ¿Qué personaje es, que tiene una altura más o menos normal, pero un peso... enorme?

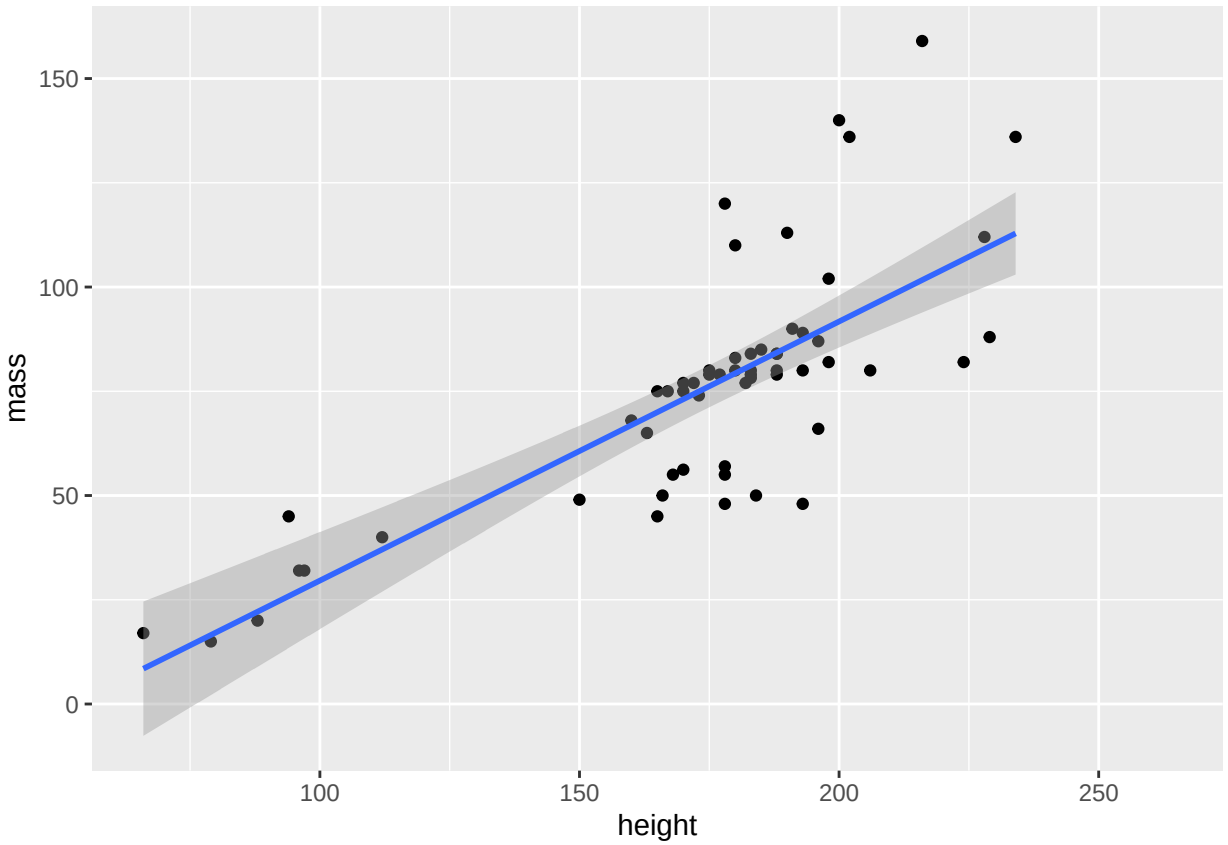
```
# ¿cuál es el peso máximo de los datos?
# literalmente: cuál es el nombre (personajes_SW$name)
# de la fila cuyo peso (mass) es igual al peso máximo
# de cualquiera de las filas (función "max").
peso.max <- which(personajes_SW$mass == max(personajes_SW$mass,na.rm = TRUE))
peso.max # el peso máximo es el del personaje de la fila 16
personajes_SW$name[peso.max] # ¿cuál es su nombre?
```

Pues claro, no podía ser otro... ¿qué sucede si eliminamos ese punto?

```
# eliminamos sólo la fila que corresponde al peso máximo
personajes_2 <- personajes_SW[-peso.max,]

# y repetimos el modelo con los datos nuevos
sw2 <- lm(formula = mass ~ height, data = personajes_2)
summary(sw2)

# lo visualizamos
ggplot(personajes_2, aes(x = height, y = mass)) +
  geom_point() +
  geom_smooth(method = "lm")
```



En este segundo modelo la relación sigue sin ser perfecta pero es estadísticamente significativa. Con este modelo podríamos predecir razonablemente bien el peso de un personaje nuevo en función unicamente de la constante y de su altura, usando los coeficientes de nuestro modelo:

$$peso = -32.54 + 0.621 * altura$$

¿Podemos mejorar nuestro modelo añadiendo otra variable explicativa? Aunque no tiene mucho sentido biológico, podemos comprobar si añadir el año de nacimiento de cada personaje nos ayuda

```
sw3 <- lm(formula = mass ~ height + birth_year,data = personajes_2)
summary(sw3)
```

Como esperábamos, el año no es significativo, no nos ayuda a explicar el peso. El modelo más convincente es el que tiene la altura como única variable explicativa, después de eliminar el punto extremo que era Jabba. Un último valor de interés, que nos sirve para evaluar el ajuste general de nuestro modelo, en vez de cada variable por separado, es el coeficiente de correlación, o Rš. Este varía entre 0 (ninguna correlación) y 1 (correlación perfecta), y aparece al final del "summary" del modelo. En general, es recomendable fijarse en el Rš ajustado, que en nuestro mejor modelo es de 0.572.

Estos valores se pueden extraer con la orden

```
summary(sw3)$r.squared
summary(sw3)$adj.r.squared
```

En estos puntos sólo hemos tratado de manera muy básica el modelo estadístico más sencillo, la regresión lineal. Incluso dentro de un modelo tan sencillo, hay mucho que nos hemos dejado, y como siempre, en internet hay muchísima información que podemos consultar. Un tutorial muy sencillo y más completo es, por ejemplo, el de DataCamp: https://www.datacamp.com/community/tutorials/linear-regression-R