

Análisis de Datos

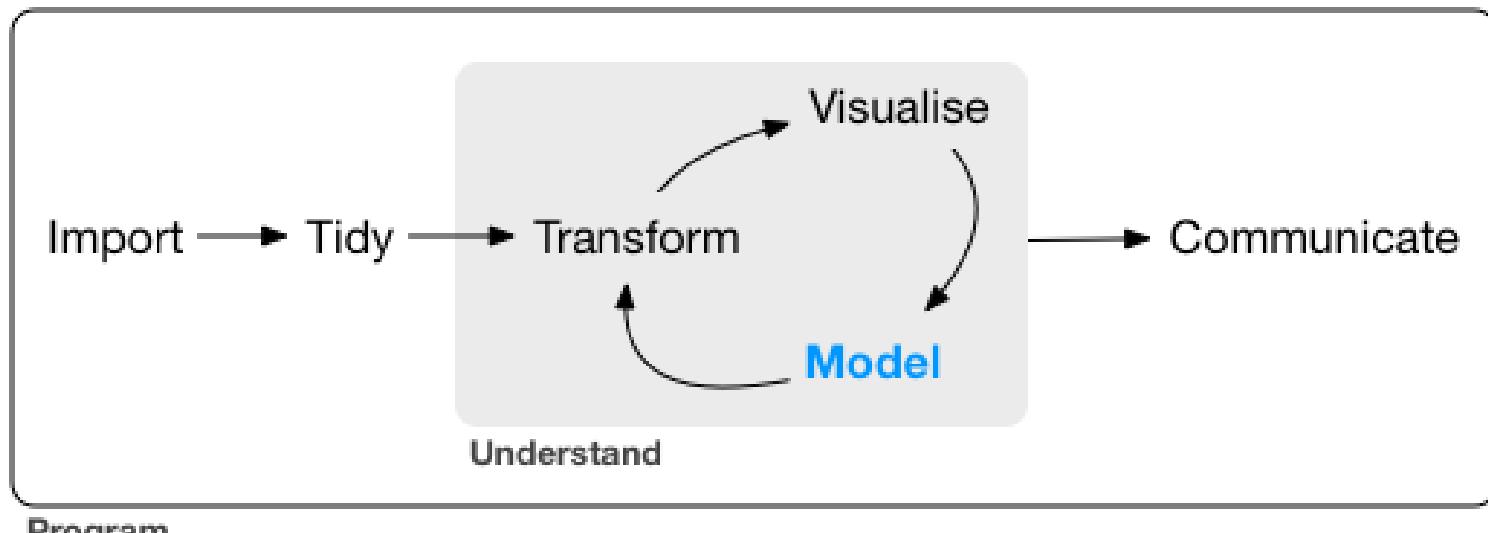
Tema 4 - Modelización

Roi Naveiro

Modelización

Las herramientas de modelización tienen cuatro objetivos fundamentales:

- Explorar los datos: los modelos a veces revelan patrones que no son evidentes en las visualizaciones (>3D)
- Generalizar hallazgos de una muestra a la población (inferencia)
- Determinar **relaciones causa-efecto** (inferencia causal)



Modelización

Los patrones descubiertos por las herramientas de modelización pueden ser:

- Patrones de asociación
- Relaciones causa-efecto

Además, estos patrones pueden:

- Darse únicamente en los datos observados
- Generalizarse a la población

Modelización

La forma de recolectar los datos afecta al tipo de generalización de las conclusiones:

- Si queremos que las conclusiones extraídas a partir de una muestra de datos sean generalizables a la población, debemos muestrear los sujetos **al azar**.

INTERNACIONAL [BARCOS](#) [ESPAÑA](#) [INTERNACIONAL](#)

'Troleo' en la red para bautizar 'Blas de Lezo' a un buque británico

'Troleo' en la red para bautizar 'Blas de Lezo' a un buque británico

[Fuente](#)

Modelización

¿Muestrear datos al azar nos garantiza que los patrones detectados sean relaciones causa-efecto reales?

Study: Cereal Keeps Girls Slim

Se observó que las mujeres que desayunaban tenían un **índice de masa corporal promedio más bajo**, un indicador común de obesidad, que las que no desayunaban. El índice fue aún más bajo para las que **desayunaban cereales**, según los hallazgos del estudio realizado por el Instituto de Investigación Médica de Maryland con fondos de NIH y el fabricante de cereales General Mills.

[...]

Los resultados se obtuvieron de una encuesta de NIH de 2379 mujeres en California, Ohio y Maryland.

[...]

Como parte de la encuesta, se preguntaba a las niñas una vez al año qué habían comido durante los tres días anteriores...

Fuente

Study: Cereal Keeps Girls Slim

Existen tres posibles explicaciones de este hallazgo:

- Comer cereales es la causa de que las mujeres estén delgadas
- Que las mujeres estén delgadas es la causa de que coman cereales
- Existe una tercera variable que es causa de estas dos, **variable de confusión**

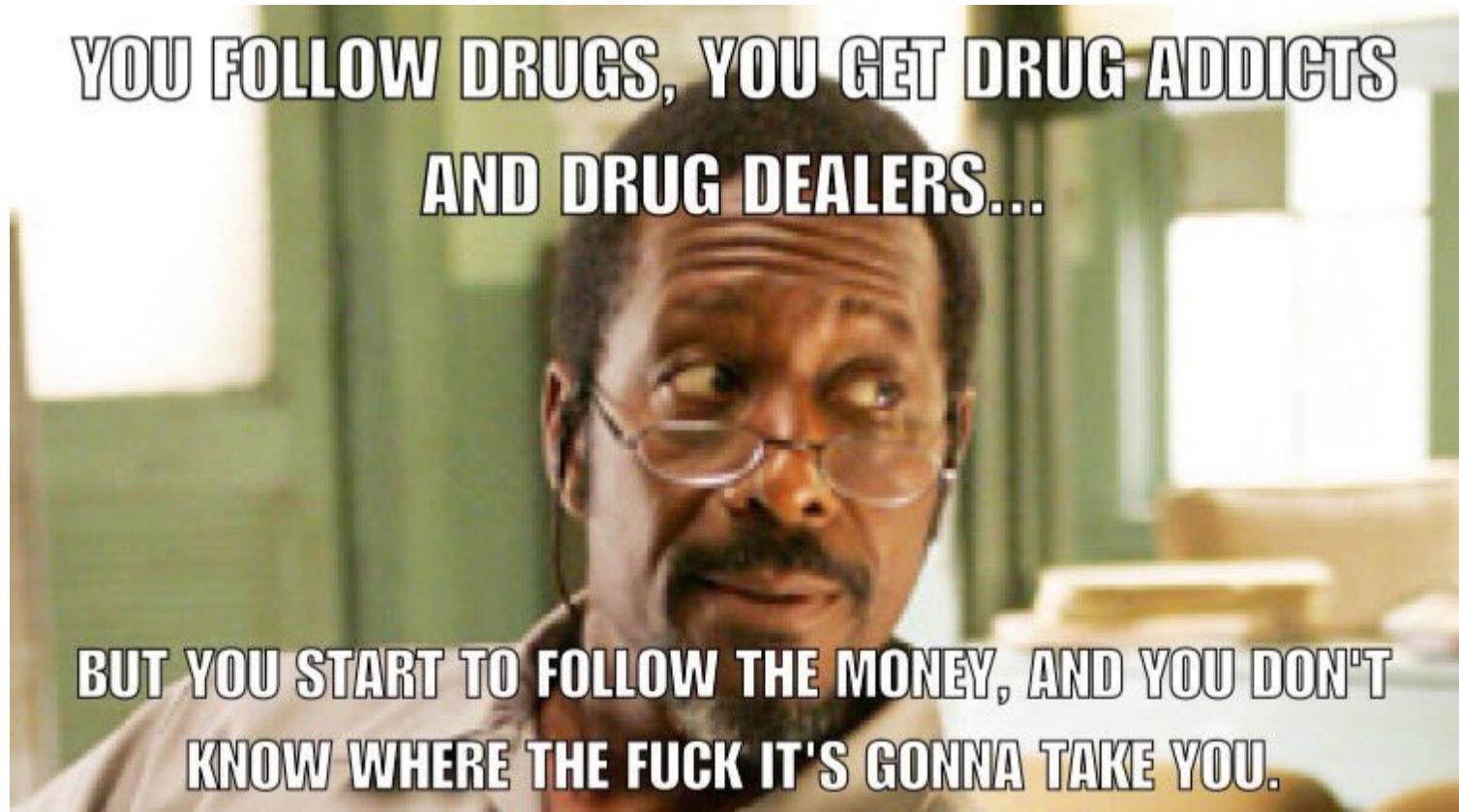
Una **variable de confusión** es una variable exógena que es causa tanto a la variable explicativa como a la de respuesta, y que hace que parezca que existe una relación entre ellas.

Study: Cereal Keeps Girls Slim

"Aquellos que desayunan regularmente tienen más probabilidades de tener un plan de alimentación estructurado a lo largo del día y, en consecuencia, es menos probable que coman entre comidas y consuman calorías vacías".

Study: Cereal Keeps Girls Slim

¿Por qué se publica esto?



Estudios científicos

Según el proceso de recolección de los datos, distinguimos estudios:

- **Observacionales**

- Se recogen datos de forma que no se altera el proceso de generación de los mismos
- Sirven para determinar **asociación**

- **Experimentales**

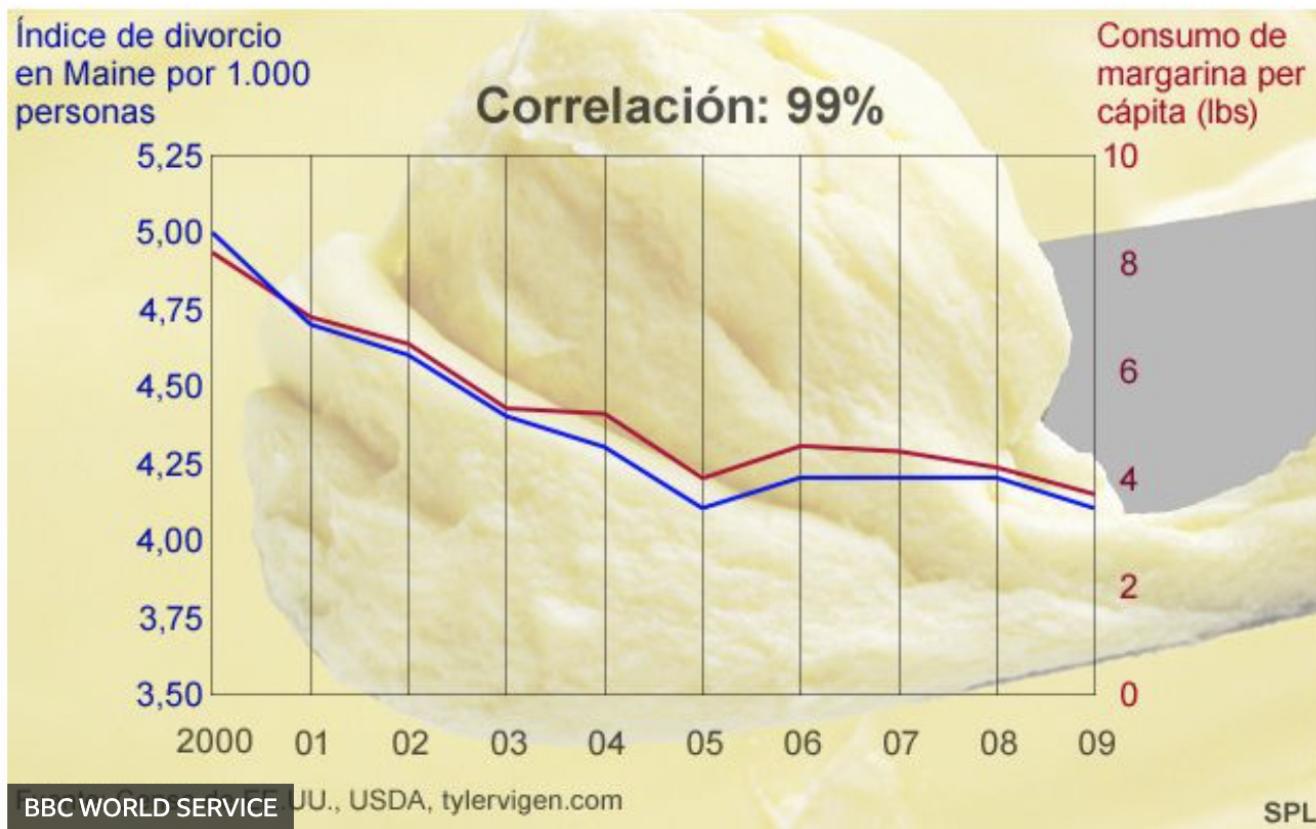
- Se asignan diferentes tratamientos a distintos individuos
- Establecer relaciones **causa-efecto**

Estudios científicos

	Random assignment	No random assignment	
Random sampling	Causal and generalizable	Not causal, but generalizable	Generalizable
No random sampling	Causal, but not generalizable	Neither causal nor generalizable	Not generalizable
	Causal	Not causal	

Correlación no implica necesariamente causalidad

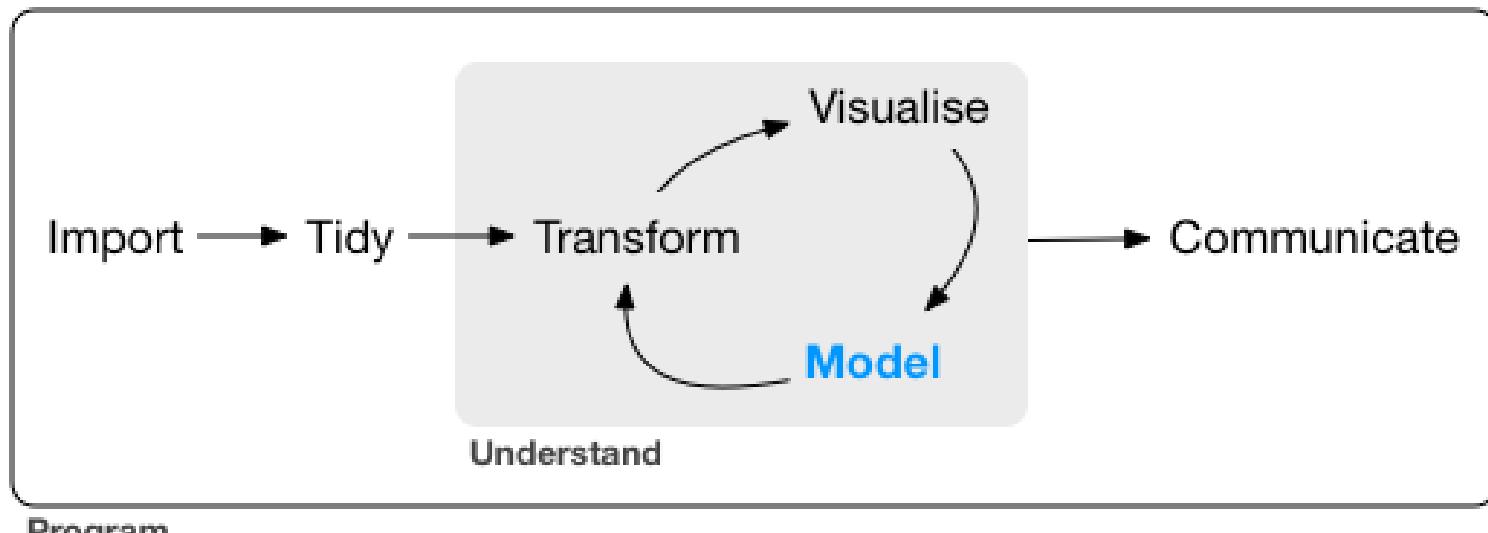
Índice de divorcio en Maine - Consumo de margarina per cápita



Modelización

Las herramientas de modelización tienen cuatro objetivos fundamentales:

- Explorar los datos: los modelos a veces revelan patrones que no son evidentes en las visualizaciones (>3D)
- Generalizar hallazgos de una muestra a la población (inferencia)
- Determinar **relaciones causa-efecto** (inferencia causal)



Modelización

- En lo que resta de curso, introduciremos herramientas básicas de modelización, poniendo el foco en su uso como **técnicas de exploración** de datos.
- **No** vamos a estudiar cómo usar estas herramientas para generalizar resultados a poblaciones, ni para determinar causalidad.
- No hay nada malo con la exploración, pero nunca debes vender un análisis exploratorio como un análisis confirmatorio porque es **engañoso**.

Elementos básicos de los modelos

¿Qué son los modelos?

- Herramientas que nos permiten extraer patrones de los datos.
- Patrones vs residuos
- Estudiaremos modelos que relacionan un serie de variables (variables predictoras) con una variable de interés (variable respuesta)

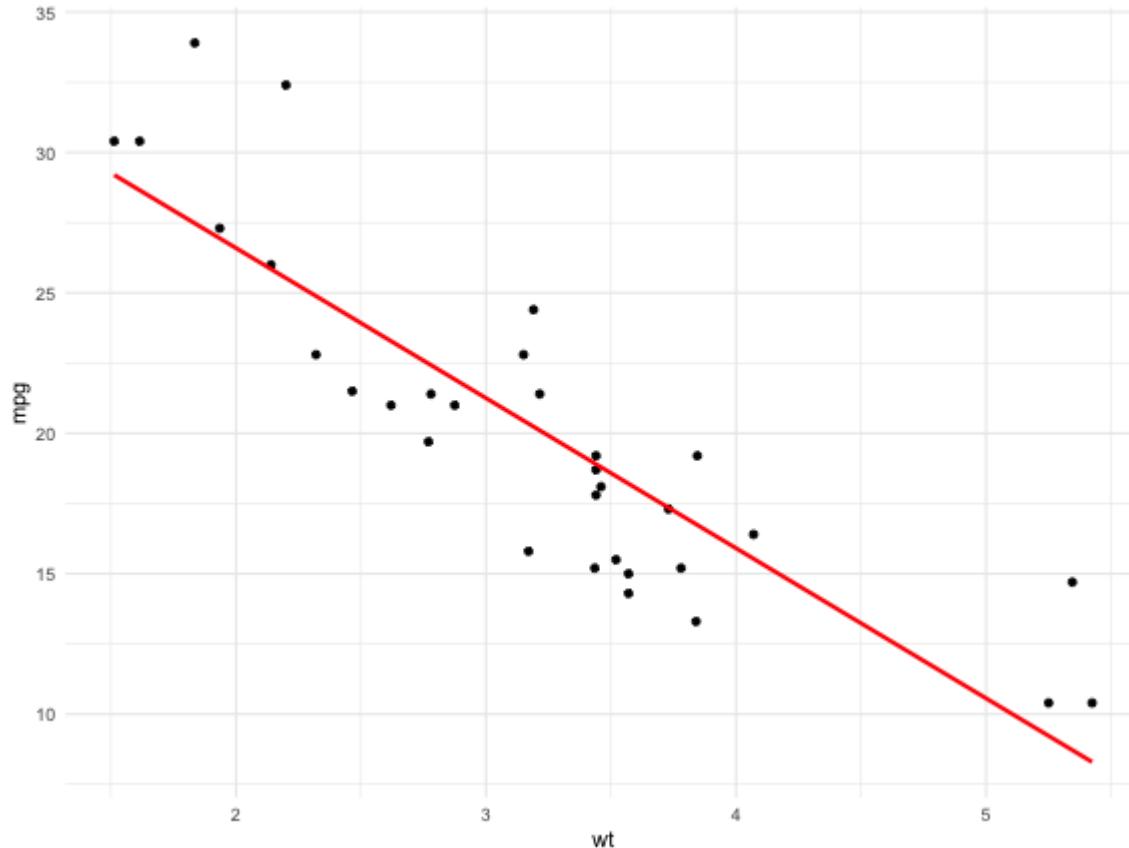
¿Qué son los modelos?

- Representamos estas relaciones como **funciones**
- Una función mapea unos inputs a un output
- Esta función tiene input x y output y

$$y = 3x + 7$$

¿Qué son los modelos?

```
ggplot(mtcars, aes(x=wt, y=mpg)) + geom_point() + theme_minimal() +  
  geom_smooth(method="lm", se=FALSE, color='red')
```



¿Qué son los modelos?

Distinguimos elementos importantes

- **Familia de modelos:** definen el patrón que queremos capturar. Por ejemplo: la familia de modelos lineales que relacionan x con y es:

$$y = \beta_0 + \beta_1 x$$

- **Modelo ajustado:** aquel dentro de la familia que mejor reproduce los datos observados

OJO: el mejor modelo de la familia no tiene por qué ser la realidad. Los modelos son simplificaciones de la realidad que nos sirven de algún propósito

¿Qué son los modelos?

*All models are wrong
but some are useful*



George E.P. Box

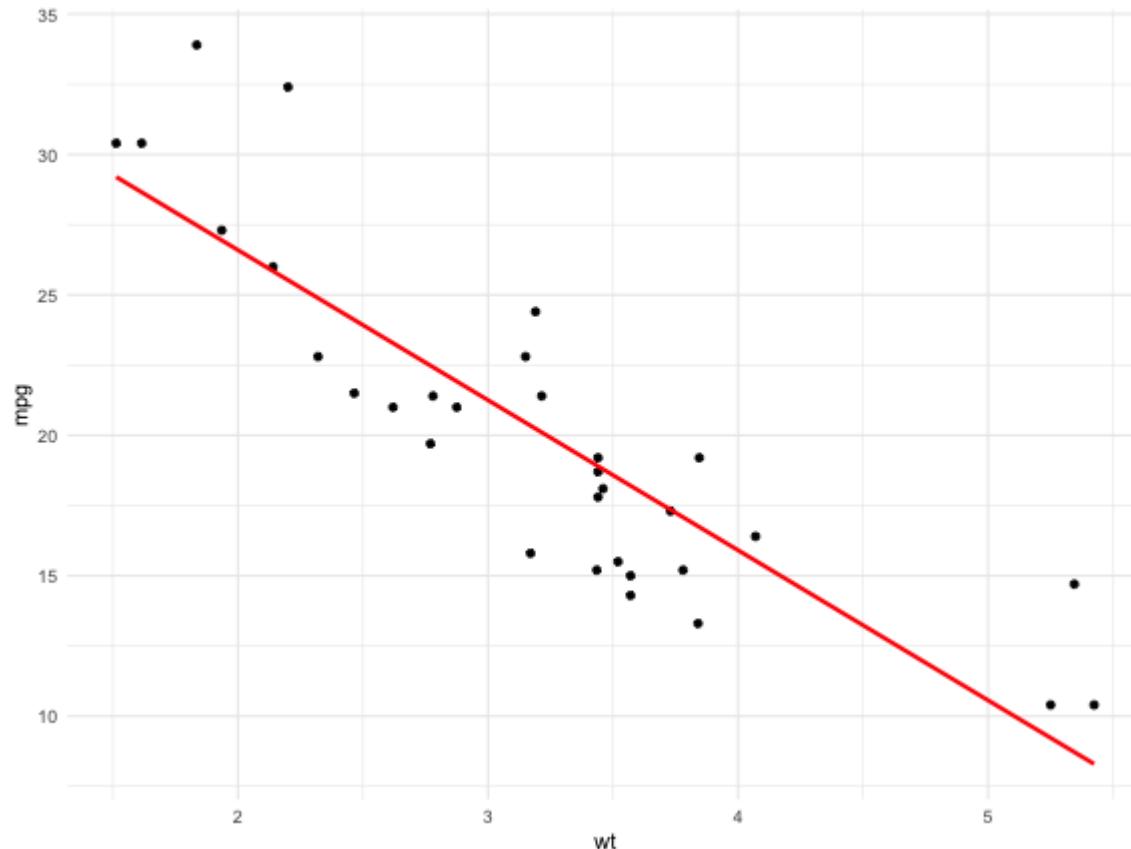
Vocabulario

- **Variable respuesta:** Variable cuyo comportamiento o variación se está tratando de entender. También llamada variable dependiente. Eje y.
- **Variables explicativas:** otras variables que desea utilizar para explicar la variación en la respuesta. También llamadas variables independientes, covariables, predictores o *features*. Eje x.
- **Valor predicho:** output del modelo para cierto valor de las covariables.

Vocabulario

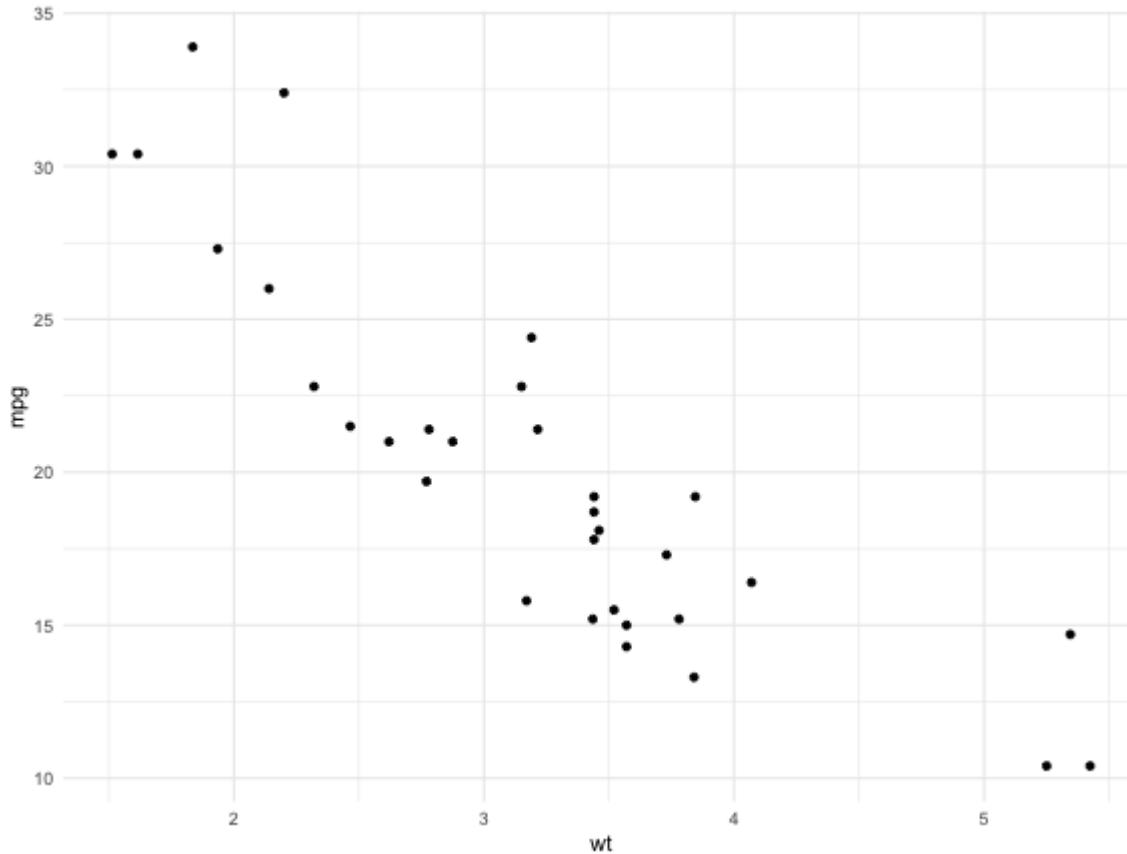
Discute los elementos anteriores en este ejemplo.

```
ggplot(mtcars, aes(x=wt, y=mpg)) + geom_point() + theme_minimal() +  
  geom_smooth(method="lm", se=FALSE, color='red')
```



Ajustando modelos

```
ggplot(mtcars, aes(x=wt, y=mpg)) + geom_point() + theme_minimal()
```

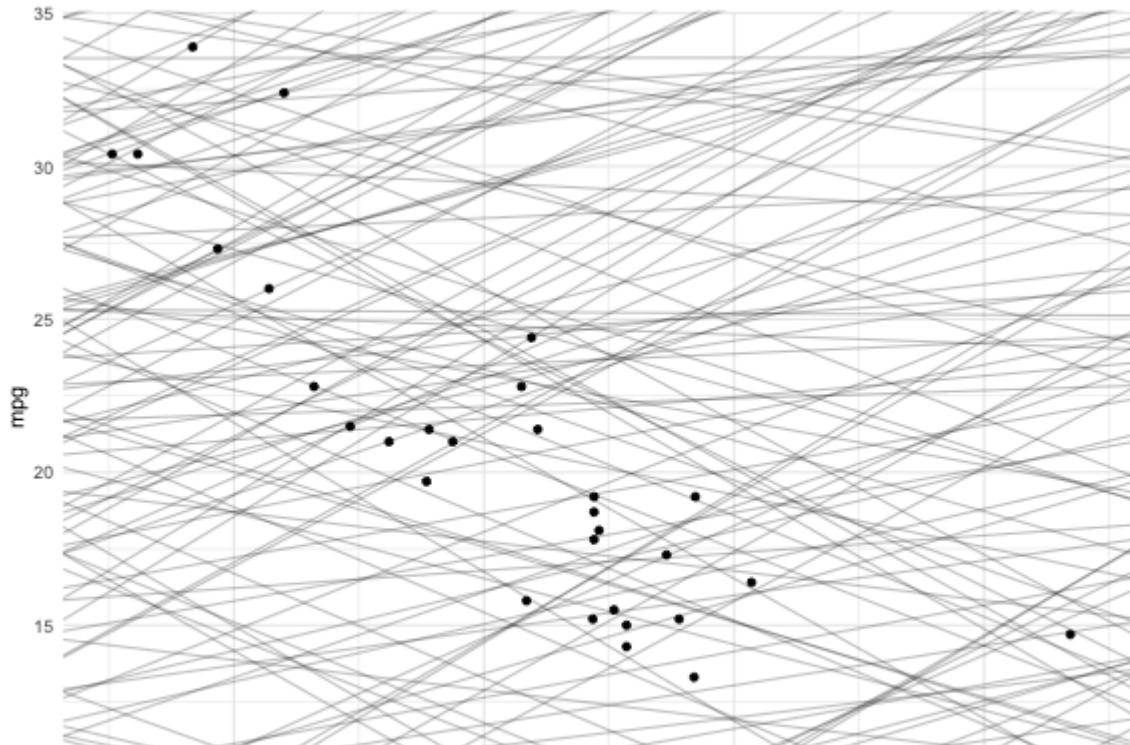


Ajustando modelos

- Los datos anteriores presentan un patrón claro
- Usaremos un modelo para capturar ese patrón y hacerlo explícito
- Un modelo lineal parece razonable $y = \beta_0 + \beta_1 x$
- Existen infinitos modelos en esta familia

Ajustando modelos

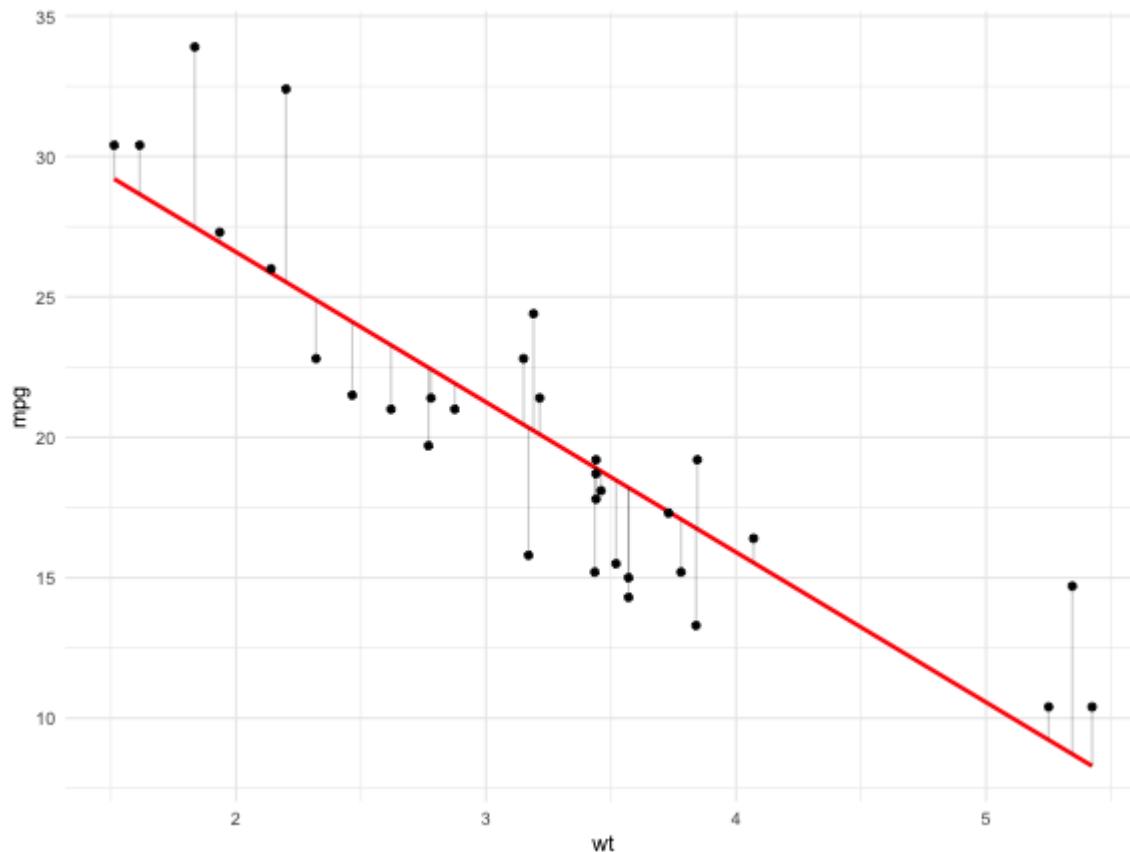
```
models <- tibble(  
  a1 = runif(250, -20, 40),  
  a2 = runif(250, -5, 5)  
)  
  
ggplot(mtcars, aes(wt, mpg)) +  
  geom_abline(aes(intercept = a1, slope = a2), data = models, alpha = 1/4) +  
  geom_point() + theme_minimal()
```



Ajustando modelos

- La mayoría de estos modelos son **malos**, no capturan el patrón
- Necesitamos determinar qué modelos son **más cercanos** a los datos
- Una opción, usar el modelo que minimice suma de las distancias verticales de cada punto a la recta del modelo

Ajustando modelos



Regresión lineal

Este modelo se conoce como regresión lineal. Para obtener el valor de los coeficientes usando R, hacemos

```
reg_mod <- lm(mpg ~ wt, data = mtcars)  
coef(reg_mod)
```

```
## (Intercept)           wt  
##   37.285126    -5.344472
```

- El objeto **mpg ~ wt** es una fórmula. Equivale a $\text{mpg} = \beta_0 + \beta_1 \cdot \text{wt}$
- Intercept es la estimación del coeficiente β_0 y el otro número es la estimación de β_1

Regresión lineal

La regresión lineal es un modelo estadístico de la relación entre un predictor x y una variable respuesta cuantitativa y cuando esta relación es lineal con cierto error ϵ

$$y = \beta_0 + \beta_1 x + \epsilon$$

En este curso no hablaremos mucho del error ϵ pero recuerda que siempre existe.

Regresión lineal

- Para estimar los valores de β_0 y β_1 , usamos los datos.
- Llamamos a las estimaciones b_0 y b_1 y $\hat{y} = b_0 + b_1x$.
- b_0 y b_1 son aquellos valores que hacen que las distancias verticales vistas anteriormente sean mínimas.

Regresión lineal - Residuos

Los residuos nos dicen cómo de lejos está cada valor predicho de su valor observado

Residuo = Valor observado - valor predicho: $y - \hat{y}$

Regresión lineal

- La recta de regresión es aquella que minimiza la suma de distancias verticales.
- Esto es equivalente a minimizar la suma de los residuos al cuadrado

$$\sum_{i=1}^n [y_i - \hat{y}_i]^2 =$$
$$\sum_{i=1}^n [y_i - (\beta_0 + \beta_1 x)]^2$$

Visualización de modelos

Podemos visualizar:

- Las predicciones de los modelos
- Los residuos de los modelos

Para esto, utilizamos **modelr** de **tidyverse**

Visualización de modelos

- Primero creamos una red de puntos de la variable predictora

```
library(modelr)
grid <- mtcars %>%
  data_grid(wt)

grid
```

```
## # A tibble: 29 × 1
##       wt
##   <dbl>
## 1 1.51
## 2 1.62
## 3 1.84
## 4 1.94
## 5 2.14
## 6 2.2 
## 7 2.32
## 8 2.46
## 9 2.62
## 10 2.77
## # ... with 19 more rows
```

Visualización de modelos

- Con el modelo, podemos añadir predicciones

```
reg_mod <- lm(mpg ~ wt, data = mtcars)

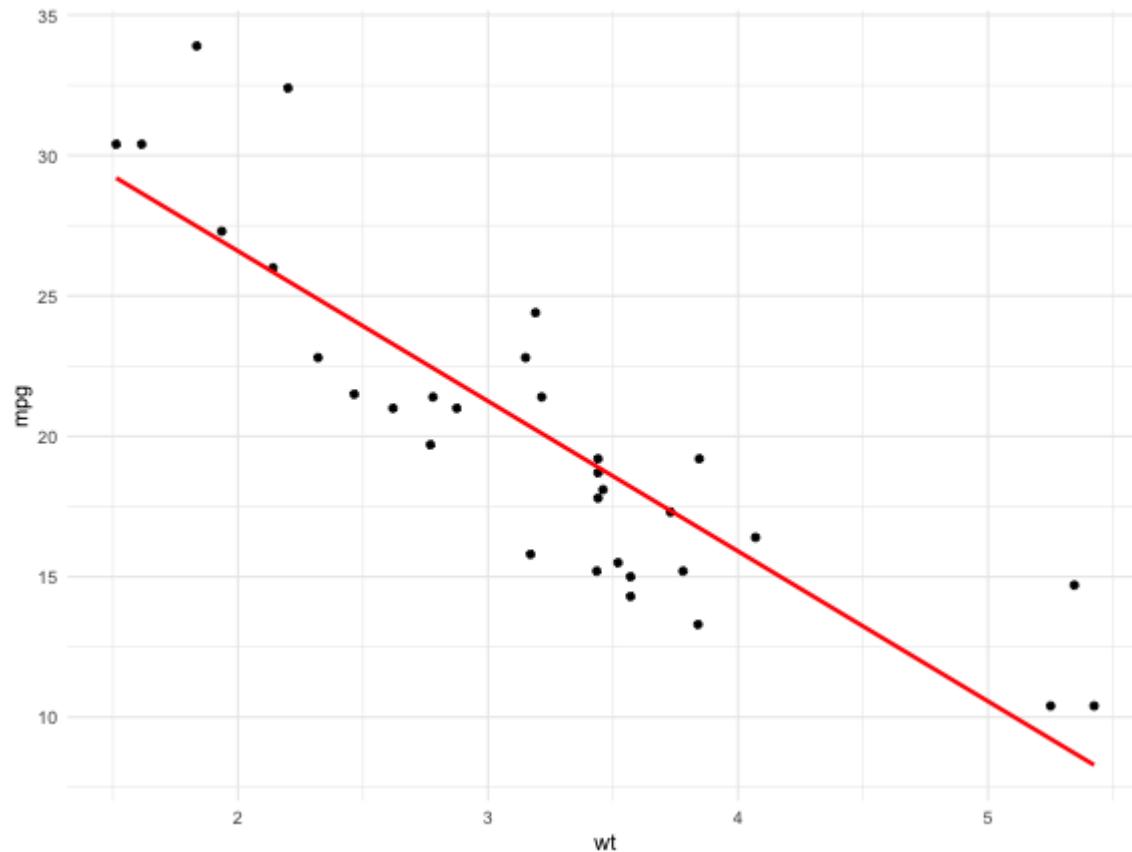
grid <- grid %>%
  add_predictions(reg_mod)

grid
```

```
## # A tibble: 29 × 2
##       wt   pred
##   <dbl> <dbl>
## 1  1.51  29.2
## 2  1.62  28.7
## 3  1.84  27.5
## 4  1.94  26.9
## 5  2.14  25.8
## 6  2.2   25.5
## 7  2.32  24.9
## 8  2.46  24.1
## 9  2.62  23.3
## 10 2.77  22.5
## # ... with 19 more rows
```

Visualización de modelos

```
ggplot(mtcars, aes(wt)) +  
  geom_point(aes(y = mpg)) + theme_minimal() +  
  geom_line(aes(y = pred), data = grid, colour = "red", size = 1)
```



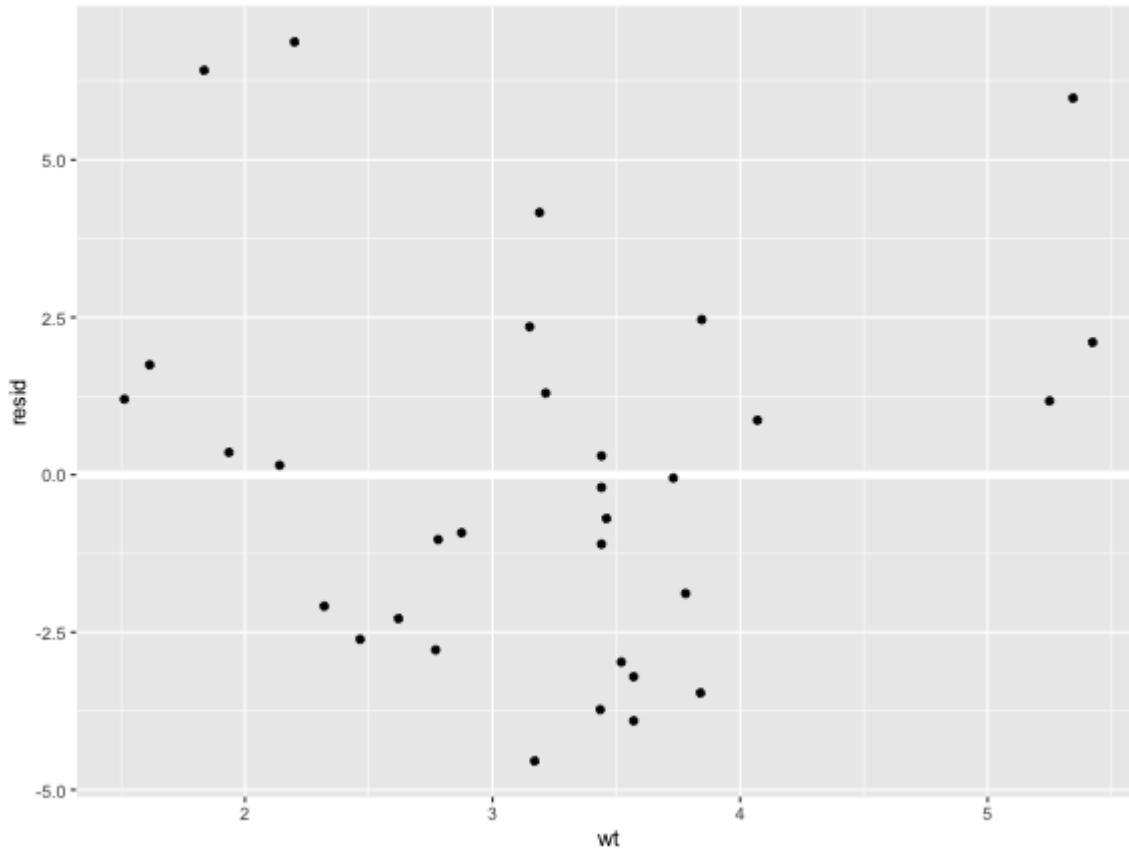
Visualización de modelos

Añadimos residuos usando

```
reg_mod <- lm(mpg ~ wt, data = mtcars)  
mtcars <- mtcars %>% add_residuals(reg_mod)
```

Visualización de modelos

```
ggplot(mtcars, aes(wt, resid)) +  
  geom_ref_line(h = 0) +  
  geom_point()
```



Visualización de modelos

- Las predicciones nos dicen qué patrón hemos capturado
- Los residuos indican qué patrón queda sin capturar

¿Qué hacer con los modelos?

- **Explicación:** caracterizar la relación entre y y x a través de los parámetros β_0 y β_1
- **Predicción:** para un nuevo valor de x , obtener su valor y

Interpretación de la regresión lineal

Volvamos a la regresión anterior. La librería **broom** de tidyverse, sirve para ordenar los resultados de **lm**. OJO: no se carga automáticamente al cargar **tidyverse**.

```
library(broom)
mod_reg <- lm(mpg ~ wt, data=mtcars)
tidy(mod_reg)
```



```
## # A tibble: 2 × 5
##   term     estimate std.error statistic p.value
##   <chr>     <dbl>     <dbl>     <dbl>     <dbl>
## 1 (Intercept) 37.3      1.88     19.9  8.24e-19
## 2 wt        -5.34     0.559    -9.56  1.29e-10
```

Interpretación de la regresión lineal

- Pendiente

El modelo de regresión lineal es

$$\widehat{\text{mpg}} = \beta_0 + \beta_1 \text{wt}$$

Aumentemos wt en una unidad

$$\begin{aligned}\beta_0 + \beta_1(\text{wt} + 1) &= \\ \beta_0 + \beta_1 \text{wt} + \beta_1 &= \\ \widehat{\text{mpg}} + \beta_1\end{aligned}$$

¿Cómo se interpreta β_1 ?

Interpretación de la regresión lineal

- Ordenada en el origen

El modelo de regresión lineal es

$$\widehat{\text{mpg}} = \beta_0 + \beta_1 \text{wt}$$

Sustituímos wt por cero

$$\widehat{\text{mpg}} = \beta_0 + \beta_1 \times 0 = \beta_0$$

¿Cómo se interpreta β_0 ?

Interpretación de la Regresión Lineal

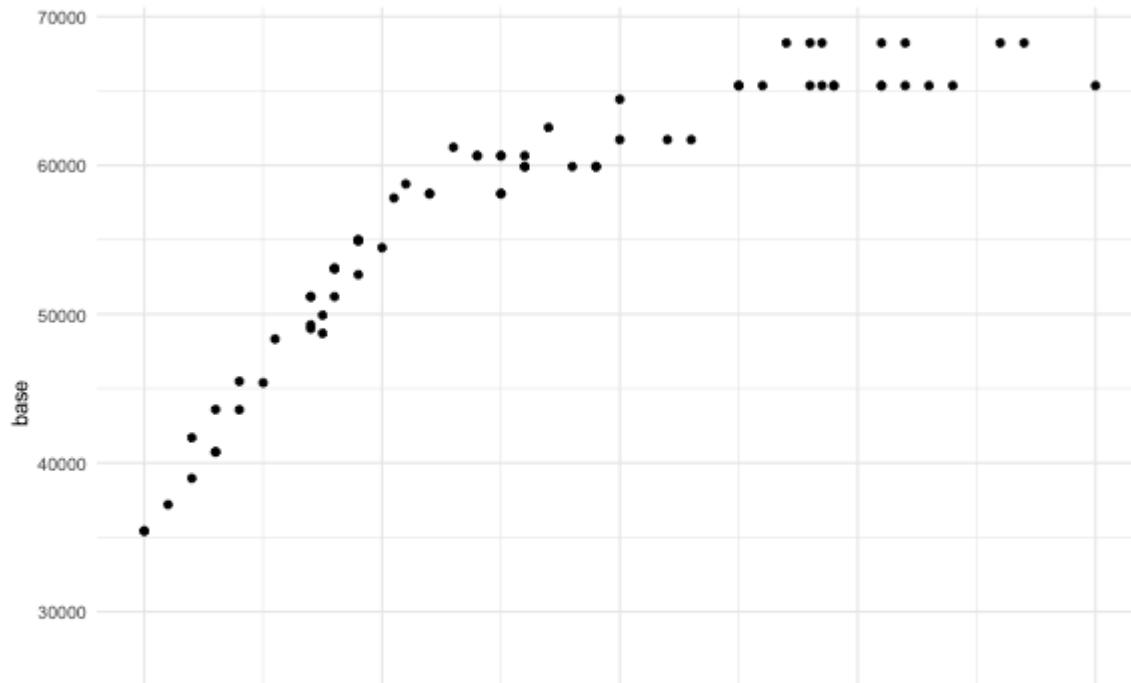
Vamos a modelizar cómo afectan los años de experiencia en el salario de un profesor

```
library(openintro)
data("teacher")
set.seed(1)
grade <- sample(c(rep("elementary", 20), rep("middle", 25), rep("high", 26)))
teacher <- teacher %>%
  mutate(degree = factor(degree, c("MA", "BA")),
        grade = grade)
```

Interpretación de la Regresión Lineal

Vamos a modelizar cómo afectan los años de experiencia en el salario de un profesor

```
ggplot(teacher, aes(x=years, y=base)) + geom_point() + theme_minimal()
```



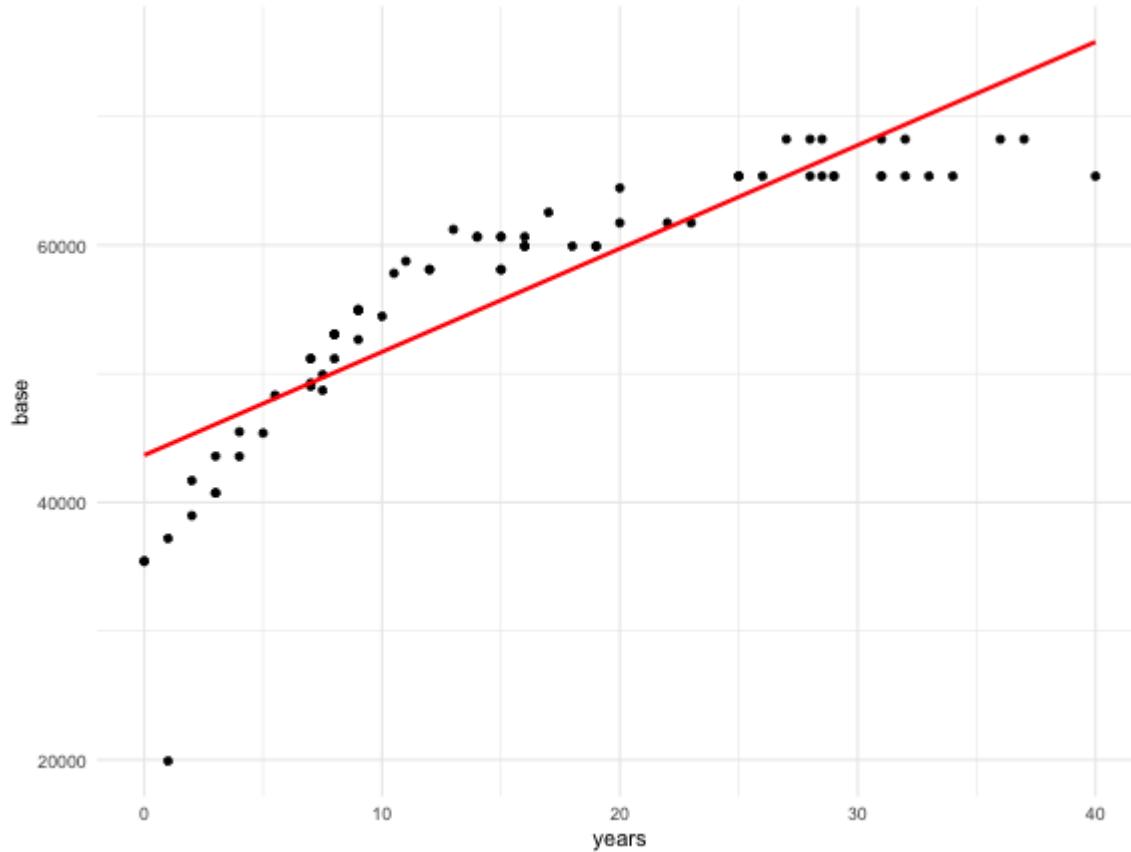
Interpretación de la Regresión Lineal

Ajusta un modelo lineal, obtén los coeficientes e interprétalos.

Determina la predicción del salario de un profesor con 15 años de experiencia.

Interpretación de la Regresión Lineal

```
ggplot(teacher, aes(x=years, y=base)) + geom_point() +  
  geom_smooth(method="lm", color='red', se=FALSE) + theme_minimal()
```



Regresión Lineal: Predictores Categóricos

- Hasta ahora hemos considerado la x continua. ¿Qué sucede si es categórica?
- Imaginemos que la x se refiere a género y toma dos valores: masculino y femenino.
- $y = \beta_0 + \beta_1 x$ no tendría sentido, x no es un número!
- Podemos hacer $y = \beta_0 + \beta_1 \text{gen_masc}$, donde gen_masc toma valor 1 para hombres y cero para mujeres

Regresión Lineal: Predictores Categóricos

- Salario frente a grado. ¿Qué niveles tiene la variable degree?
- Ajustamos un modelo

```
mod_base_degree <- lm(base ~ degree, data=teacher)
tidy(mod_base_degree)
```

```
## # A tibble: 2 × 5
##   term      estimate std.error statistic p.value
##   <chr>      <dbl>     <dbl>      <dbl>    <dbl>
## 1 (Intercept) 56610.     1777.     31.8  5.44e-43
## 2 degreeBA     -352.     2398.    -0.147 8.84e- 1
```

Regresión Lineal: Predictores Categóricos

- R ha creado una **variable indicatriz** degreeBA: si el grado es BA toma valor 1, sino 0.
- Si la variable tuviese tres niveles A, B y C, R crearía dos variables indicatrices, e.g. para A y B.
- El nivel base es el nivel que toma la variable cuando todas las indicatrices son 0.
- Los coeficientes se interpretan respecto al nivel base.

Regresión Lineal: Predictores Categóricos

```
mod_base_degree <- lm(base ~ degree, data=teacher)
tidy(mod_base_degree)
```

```
## # A tibble: 2 × 5
##   term      estimate std.error statistic p.value
##   <chr>      <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept) 56610.     1777.     31.8  5.44e-43
## 2 degreeBA     -352.     2398.    -0.147 8.84e- 1
```

Interpreta cada coeficiente

Regresión Lineal: Visualizamos las predicciones

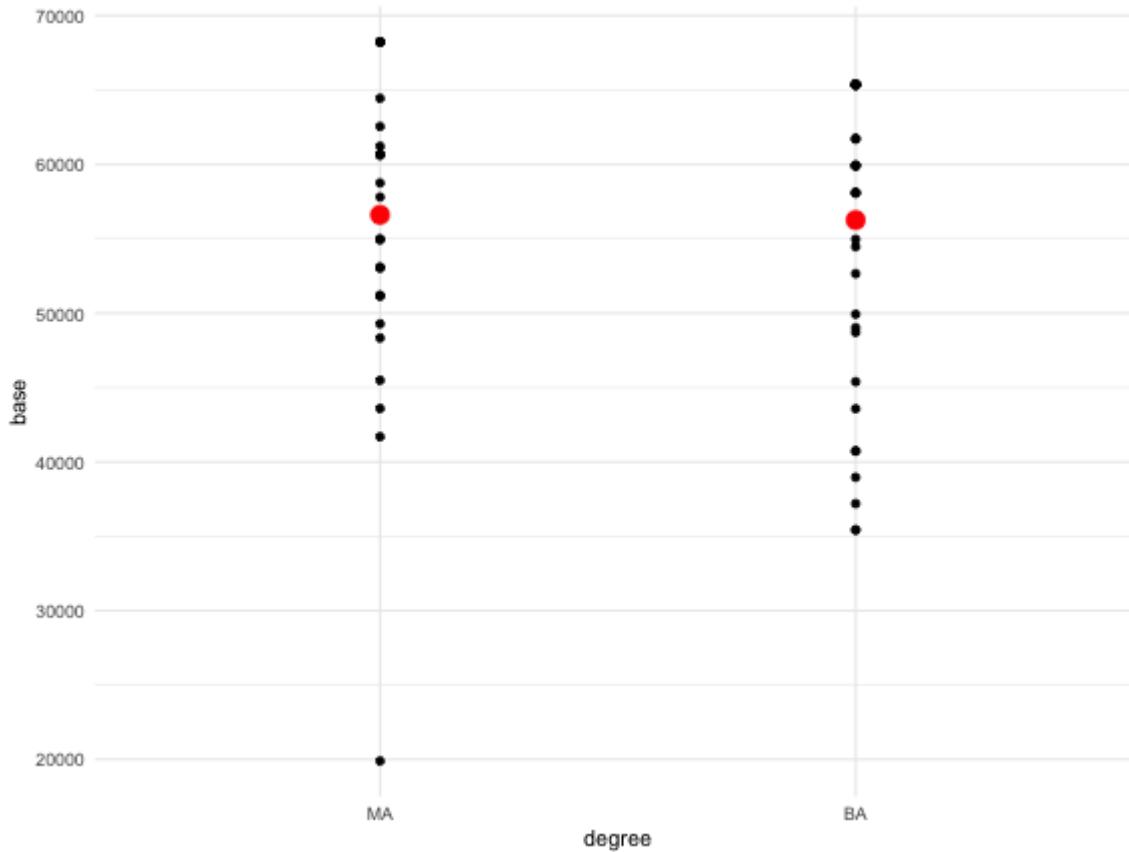
```
mod_base_degree <- lm(base ~ degree, data=teacher)

grid <- teacher %>%
  data_grid(degree) %>%
  add_predictions(mod_base_degree)
grid
```

```
## # A tibble: 2 × 2
##   degree     pred
##   <fct>     <dbl>
## 1 MA       56610.
## 2 BA       56257.
```

Regresión Lineal: Visualizamos las predicciones

```
ggplot(teacher, aes(x=degree)) + geom_point(aes(y=base)) + theme_minimal() +  
  geom_point(data = grid, aes(y = pred), colour = "red", size = 4)
```



Regresión Lineal: Visualizamos las predicciones

Predecimos la media para cada grupo!

Regresión Lineal Múltiple

- En múltiples ocasiones, tenemos más de una variable predictora.
- El modelo lineal más sencillo para este caso, es la **regresión lineal múltiple**

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$$

- El efecto de cada variable se estima independientemente del resto de variables.

Regresión Lineal Múltiple - Interpretación

- ¿Cómo se interpreta β_0 ?
- ¿Cómo se interpreta β_1 ?

Regresión Lineal Múltiple

En R

```
mod_reg <- lm(base ~ years + degree, data = teacher)
tidy(mod_reg)
```

```
## # A tibble: 3 × 5
##   term      estimate std.error statistic p.value
##   <chr>     <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept) 44838.    1156.     38.8  4.11e-48
## 2 years        818.     54.0     15.2  1.75e-23
## 3 degreeBA     -2560.    1164.    -2.20  3.12e- 2
```

Regresión Lineal Múltiple

Visualizamos las predicciones

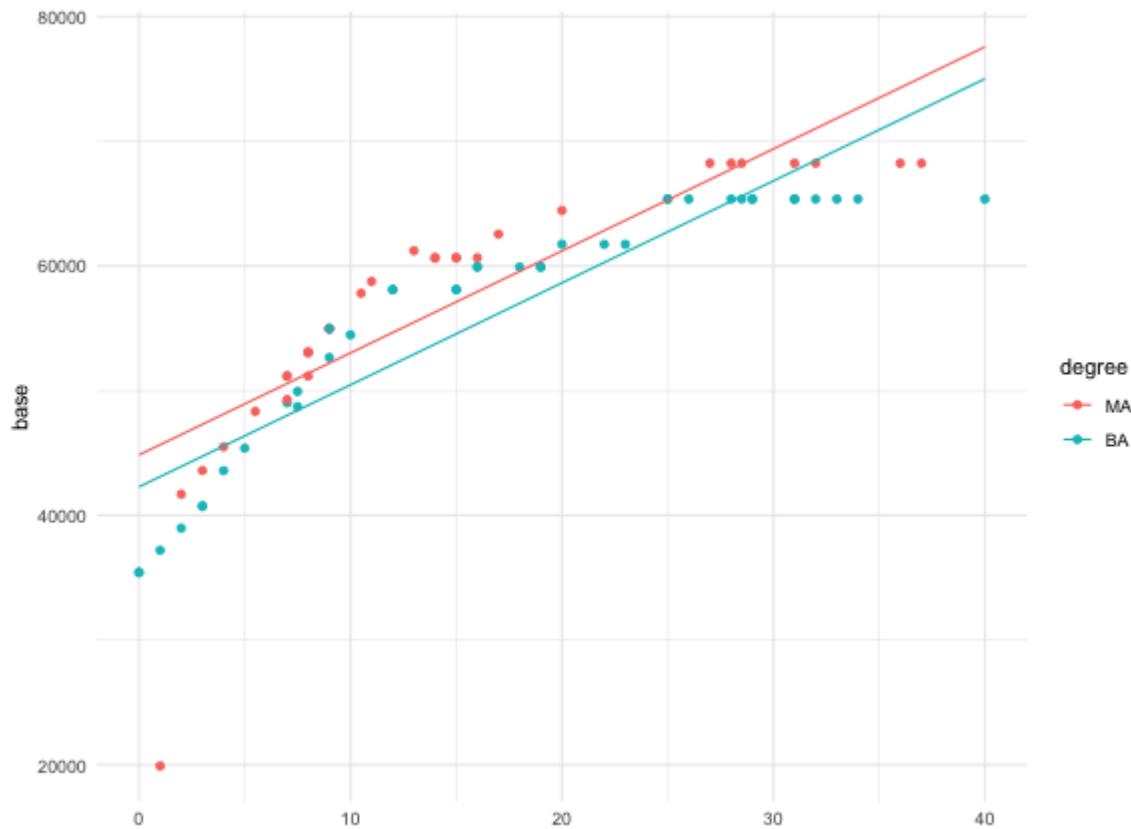
```
grid <- teacher %>%
  data_grid(years, degree) %>%
  add_predictions(mod_reg)
grid

## # A tibble: 76 × 3
##   years degree   pred
##   <dbl> <fct>   <dbl>
## 1     0 MA     44838.
## 2     0 BA     42278.
## 3     1 MA     45656.
## 4     1 BA     43096.
## 5     2 MA     46474.
## 6     2 BA     43914.
## 7     3 MA     47292.
## 8     3 BA     44732.
## 9     4 MA     48110.
## 10    4 BA     45550.
## # ... with 66 more rows
```

Regresión Lineal Múltiple

Visualizamos las predicciones

```
ggplot(teacher, aes(years, base, colour = degree)) +  
  geom_point() +  
  geom_line(data = grid, aes(y = pred)) + theme_minimal()
```



Bibliografía

Este tema está fundamentalmente basado en [R for Data Science](#), Wickham and Grolemund (2016)