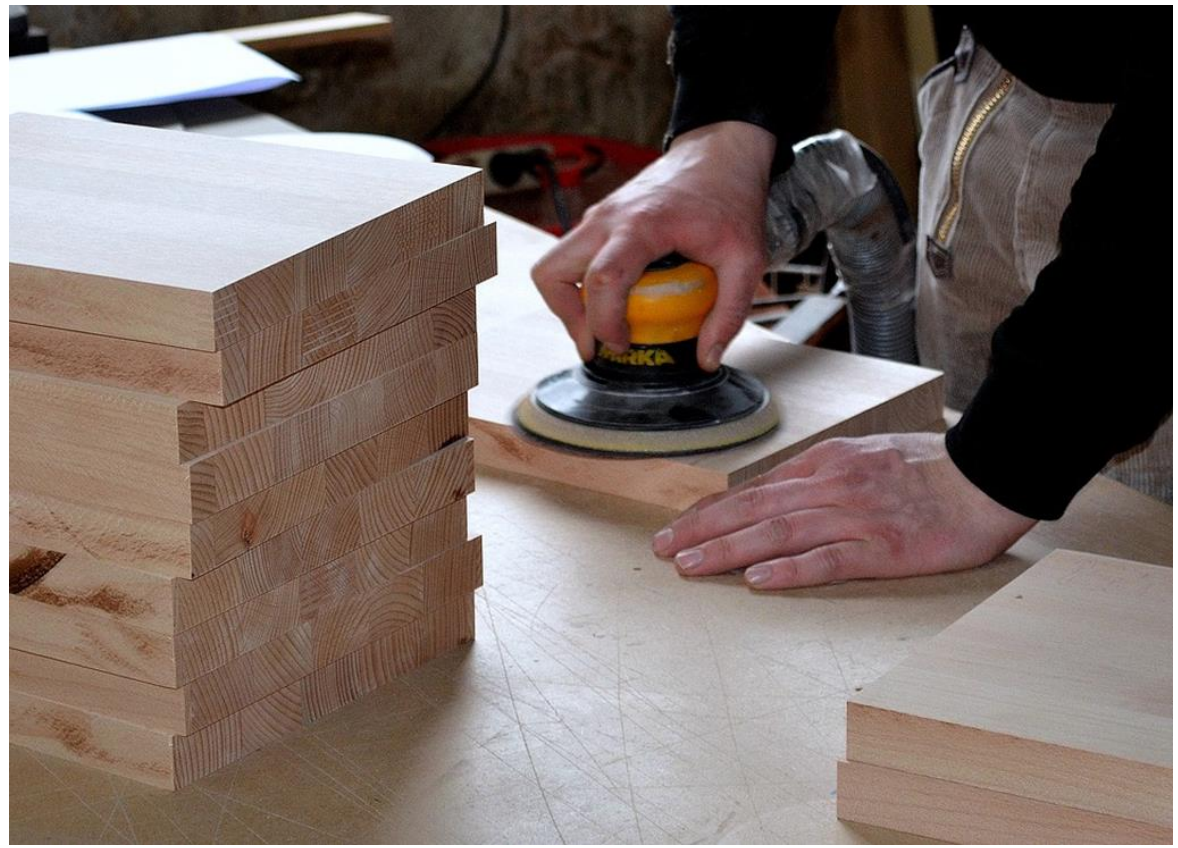
Modelos predictivos

El paso a paso de los modelos predictivos

¿Qué vas a ver en este bloque?

- Qué es un modelo estadístico y los dos grandes súper poderes
- El mapa de los modelos:
  - Las ANOVA también son modelos
  - Las regresiones y su generalización
  - Los dos grandes retos:
    - Saber si un modelo es coherente con los datos
    - Selección del modelo y comparación de modelos
- Cómo interpretar los resultados de un modelo lineal



Qué es un modelo estadístico y sus dos grandes súper poderes

El objetivo de un modelo estadístico

Inferir un modelo

MONTH	Nº CANDIDATES	CAPACITY	% ASSISTANCE	Nº CANDIDATES	CAPACITY	% ASSISTANCE	Nº CANDIDATES	CAPACITY	% ASSISTANCE	WTI	BRENT
1	20	28	71%	14	16	88%	6	12	50%	94.62	108.12
2	24	28	86%	16	16	100%	8	12	67%	100.82	108.90
3	18	28	64%	10	16	63%	8	12	67%	100.80	107.48
4	19	27	70%	15	15	93%	6	6	100%	107.42	104.76
5	28	36	78%	20	24	83%	8	12	67%	102.18	100.54
6	20	28	71%	9	16	56%	11	12	92%	105.79	111.80
7	25	28	89%	16	16	100%	9	12	75%	100.59	106.72
8	18	28	64%	8	16	50%	10	12	83%	95.54	101.61
9	21	28	75%	11	16	69%	10	12	83%	98.21	97.09
10	25	36	69%	16	24	67%	9	12	75%	84.68	87.45
11	16	22	73%	10	16	63%	6	6	100%	75.79	73.44
12	15	20	75%	5	8	63%	10	12	83%	50.29	62.34
13	17	24	71%	9	16	56%	8	8	100%	47.22	42.76
14	9	12	75%	5	8	63%	4	4	100%	50.58	58.10
15	18	24	75%	10	16	63%	8	8	100%	47.82	55.80
16	17	32	53%	11	24	46%	6	8	75%	54.45	59.52
17	9	12	75%	7	8	88%	3	4	75%	56.17	64.06
18	14	24	58%	11	16	69%	3	8	38%	50.62	61.48
19	11	28	39%	3	16	19%	8	12	67%	50.00	50.36
20	4	12	33%	2	8	25%	2	4	50%	43.87	48.13
21	7	16	44%	3	8	38%	4	8	50%	45.48	47.62
22	4	28	14%	3	24	13%	1	4	25%	46.22	48.48
23	4	12	33%	2	8	25%	2	4	50%	43.44	44.37
24	7	16	44%	1	8	13%	6	8	75%	37.79	38.03
25	4	4	100%	0	0	0%	4	4	0%	33.68	30.70
26	1	4	25%	0	0	0%	1	4	0%	30.52	32.18
27	5	12	42%	1	8	13%	2	4	50%	37.55	38.31
28	1	4	25%	0	0	0%	1	4	0%	40.76	41.38
29	3	12	25%	2	8	25%	1	4	25%	40.71	40.74
30	2	12	17%	1	8	13%	1	4	25%	48.76	48.35

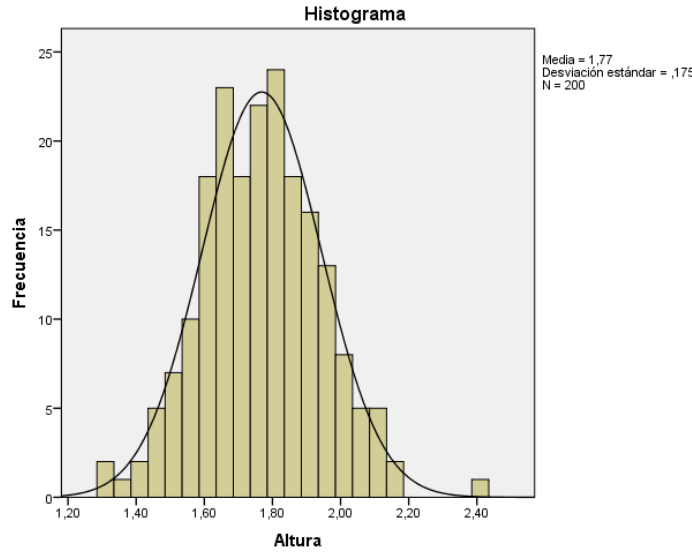
Población

$$y = f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Densidad de Frecuencia

Variables Numérica (Altura en cm)

Muestra

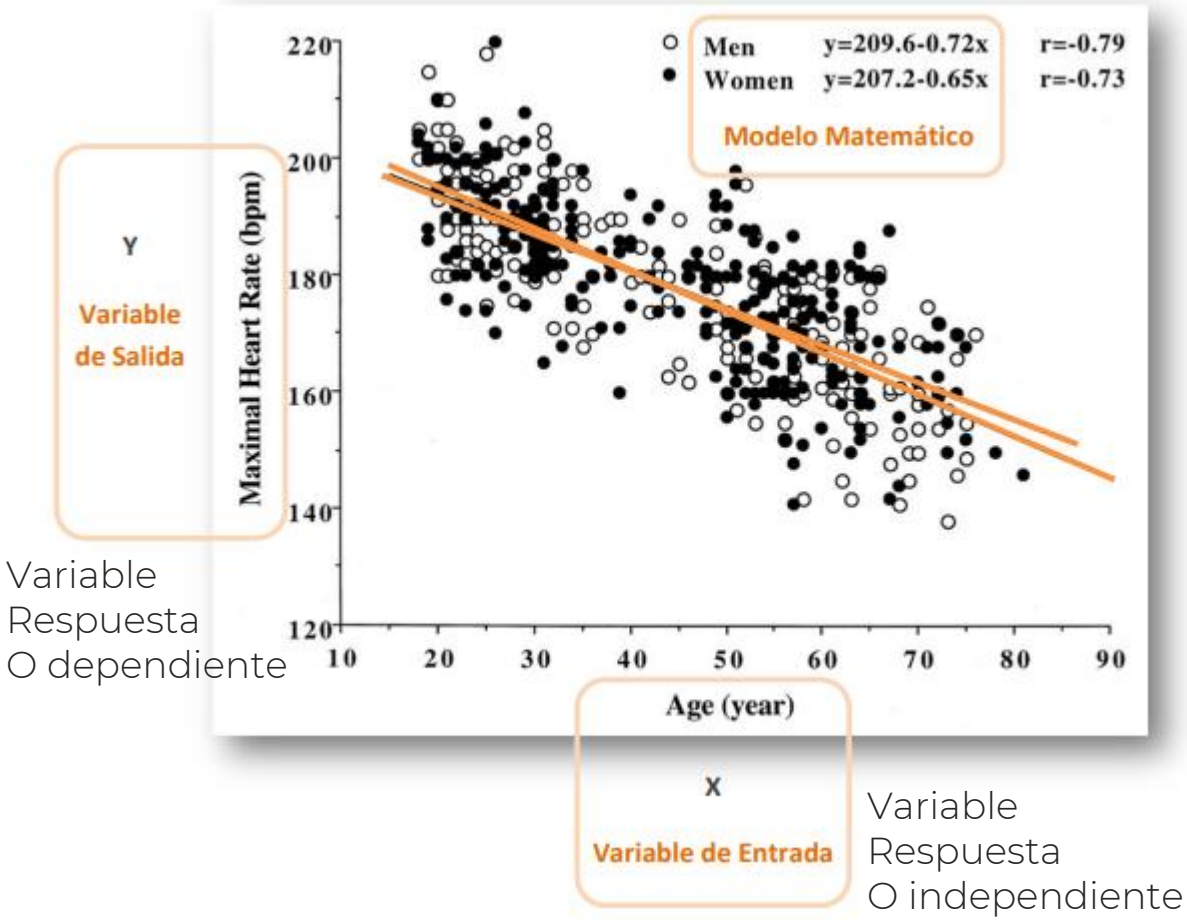


¿Qué es un modelo predictivo?

- Un modelo es **una función matemática** que a partir de unos datos de entrada (variables de entrada) obtienes los datos de la variables de salida (o respuesta)

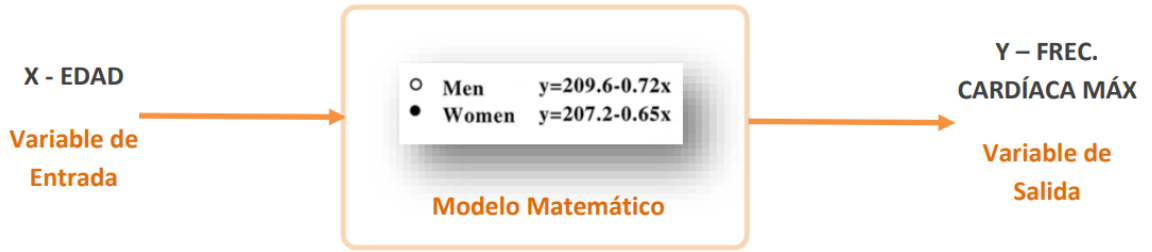
- Un modelo tiene tres partes bien diferenciadas:

Variable Salida – Y	Frecuencia Cardíaca Máxima
Variable Entrada – X	Edad
Función matemática	○ Men y=209.6-0.72x ● Women y=207.2-0.65x



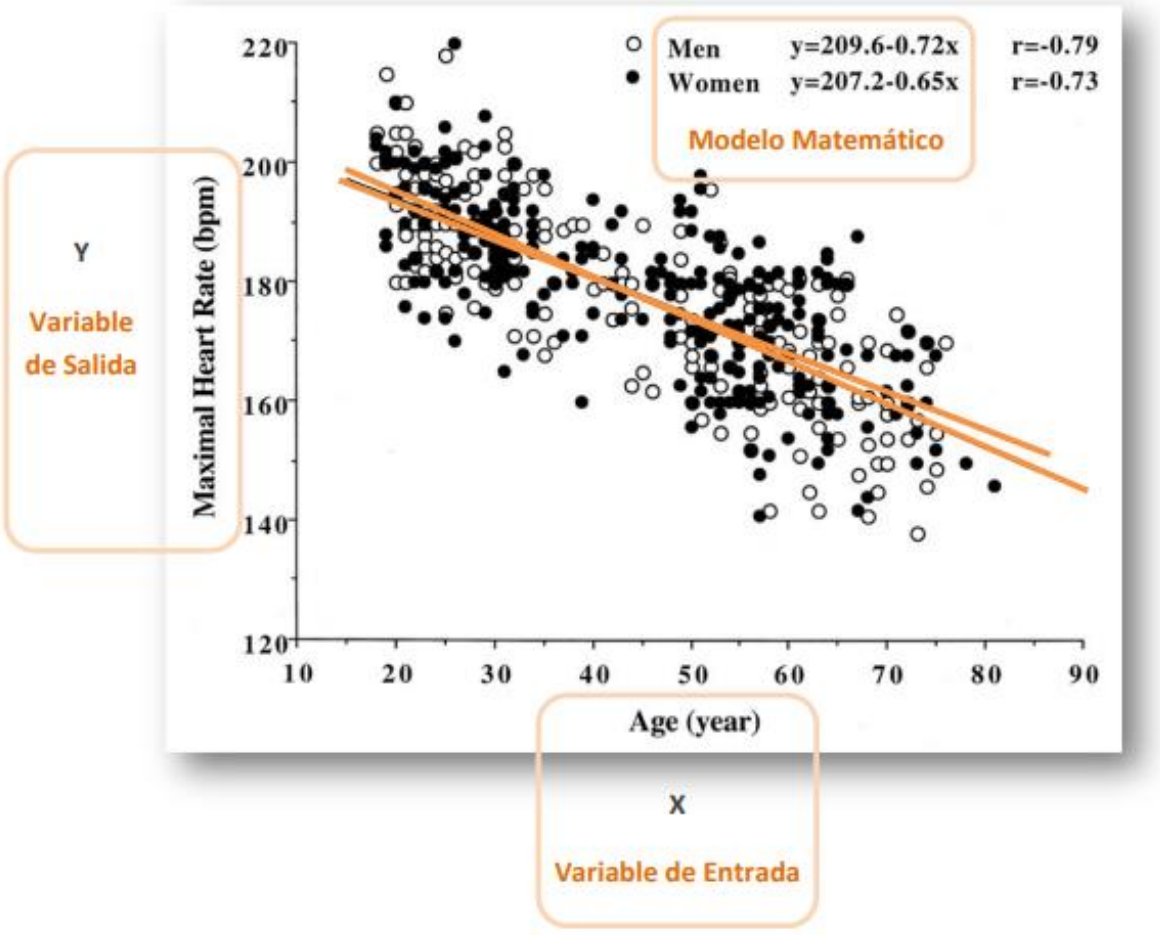
Variable Respuesta O dependiente

Variable Respuesta O independiente



Conceptosclaros²

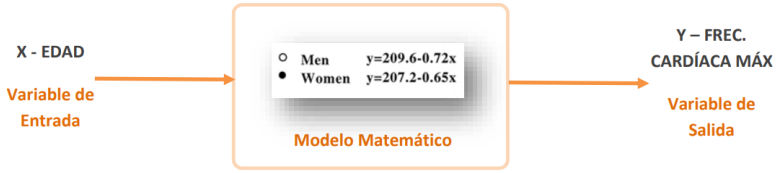
¿Qué es un modelo predictivo?



- El **modelo es una recta** en este caso (es el modelo más sencillo)

$$y = a \cdot x + b$$

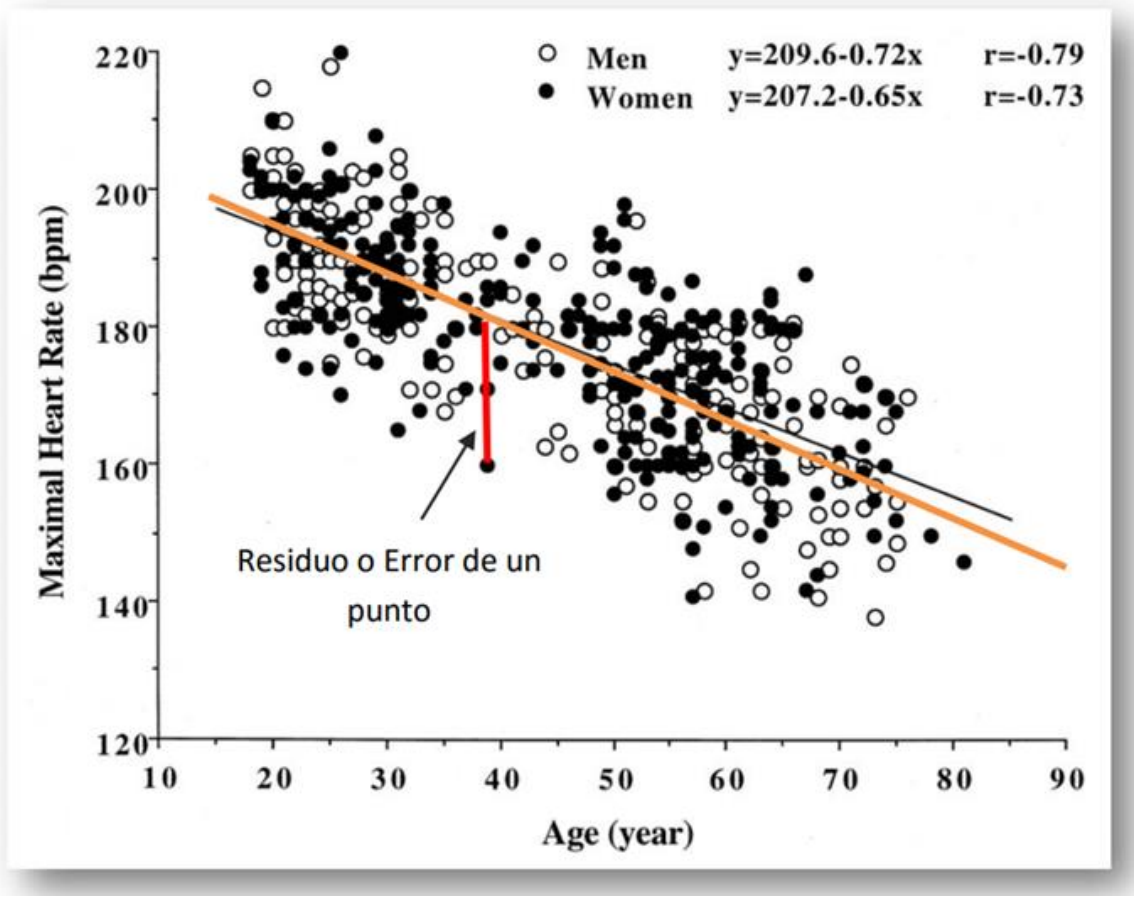
  - **a**: es la pendiente de la recta
  - **b**: es dónde cruza la recta en el eje vertical
- A partir de un cálculo de inferencia podemos obtener el valor de los parámetros del modelo **a** y **b** que mejor ajustan los datos de hombre y en mujeres
- **a** = 209,6 y **b** = -0,72 para los datos de **hombres**
- **a** = 207,2 y **b** = -0,65 para los datos de **mujeres**



$$Frec.Cardíaca\ Máxima = -0{,}72 \cdot Edad + 209{,}6 + Error$$

Variable de Salida — Parámetros del modelo — Variable de Entrada

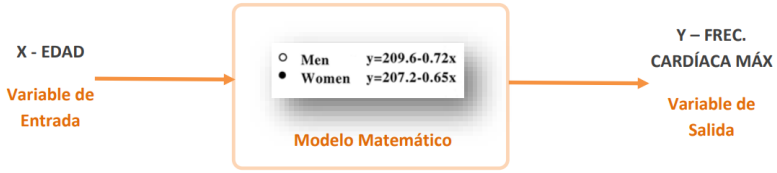
¿Para qué sirve? Utilidad nº1: **La Estimación**



- Por ejemplo si pongo **la edad de 32**. La frecuencia cardíaca Máxima es:

$$Frec.\,Cardíaca\,Máxima = -0{,}72 \cdot \mathbf{32} + 209{,}6 = 186{,}56$$

- Un modelo siempre tiene **un error de estimación o error del modelo**



$$Frec.\,Cardíaca\,Máxima = -0{,}72 \cdot Edad + 209{,}6 + Error$$

Variable de Salida

Parámetros del modelo

Variable de Entrada

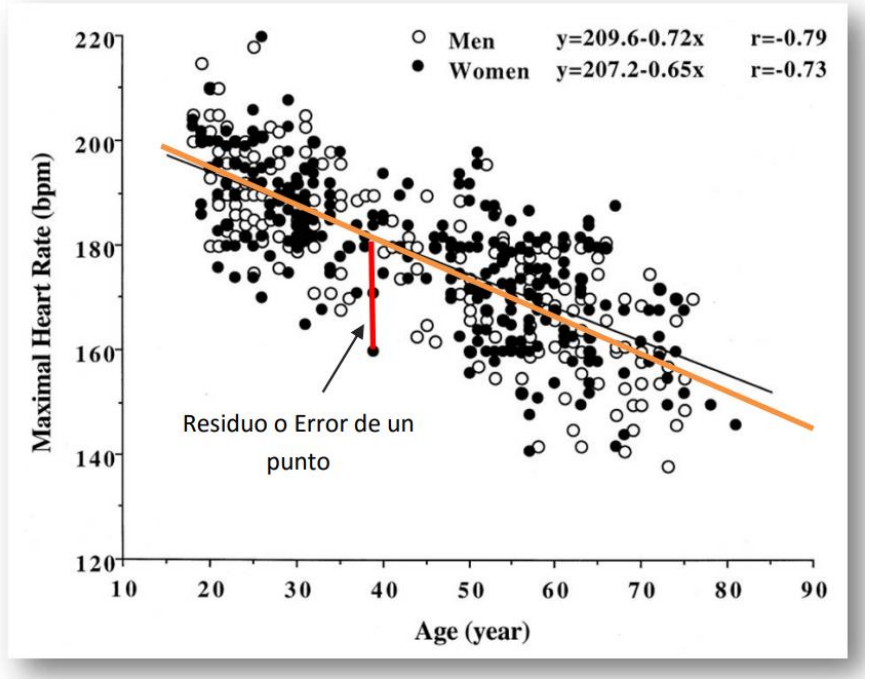
¿Para qué sirve?
Utilidad nº1: **La Estimación**

- Imagina que para uno de los pacientes de 32 años la frecuencia cardíaca máxima es de 145 pulsaciones por minuto
- Para el modelo que hemos calculado la frecuencia cardíaca máxima es la siguiente:

$$Frec.\,Cardíaca\,Máxima = -0{,}72 \cdot \mathbf{32} + 209{,}6 = 186{,}56$$

- El error o residuo del modelo es el valor observado menos el valor estimado por el modelo

$$observado - esperado = y_{obs} - \hat{y} = 145 - 186{,}56 = -41{,}56$$



$$Frec.\,Cardíaca\,Máxima = -0{,}72 \cdot Edad + 209{,}6 + Error$$

Variable de Salida

Parámetros del modelo

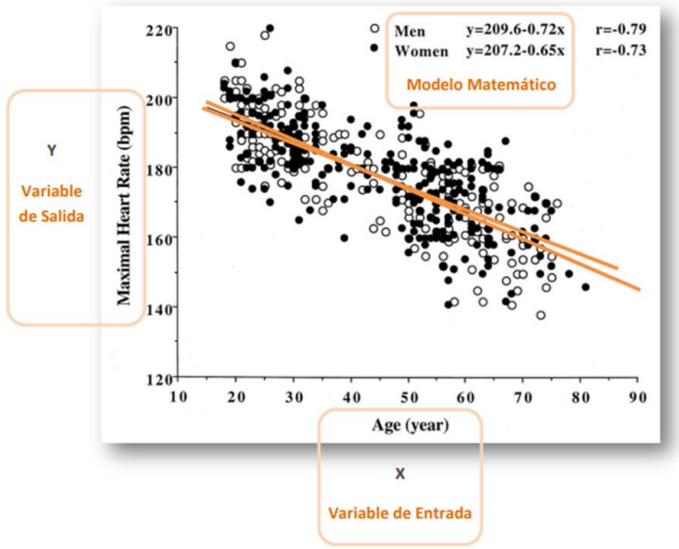
Variable de Entrada

¿Para qué sirve?
Utilidad nº2: **Relacionar**

- Podemos utilizar el modelo para explicar lo que está pasando con la frecuencia cardíaca máxima y la edad
- Un modelo puede explicar relaciones causa-efecto
- El modelo lineal del ejemplo es la línea naranja del gráfico
- Es una recta con pendiente negativa ya que el valor del coeficiente a es de -0,72
- Eso significa que cuanta más edad tienen los pacientes menor es la frecuencia cardíaca máxima
- A mayor edad → corazón más débil → frecuencia cardíaca máxima menor
- Además en el estudio se han calculado dos modelos, uno para los datos de hombre y otro para mujeres. También podemos comparar los modelos y ver la influencia de la edad con la frecuencia cardíaca máxima:

$\boldsymbol{Frec.Cardíaca\ Máxima} = -0{,}72 \cdot \boldsymbol{Edad} + 209{,}6$ → Para Hombres

$\boldsymbol{Frec.Cardíaca\ Máxima} = -0{,}65 \cdot \boldsymbol{Edad} + 207{,}2$ → Para Mujeres



$$Frec.Cardíaca\ Máxima = -0{,}72 \cdot Edad + 209{,}6 + Error$$

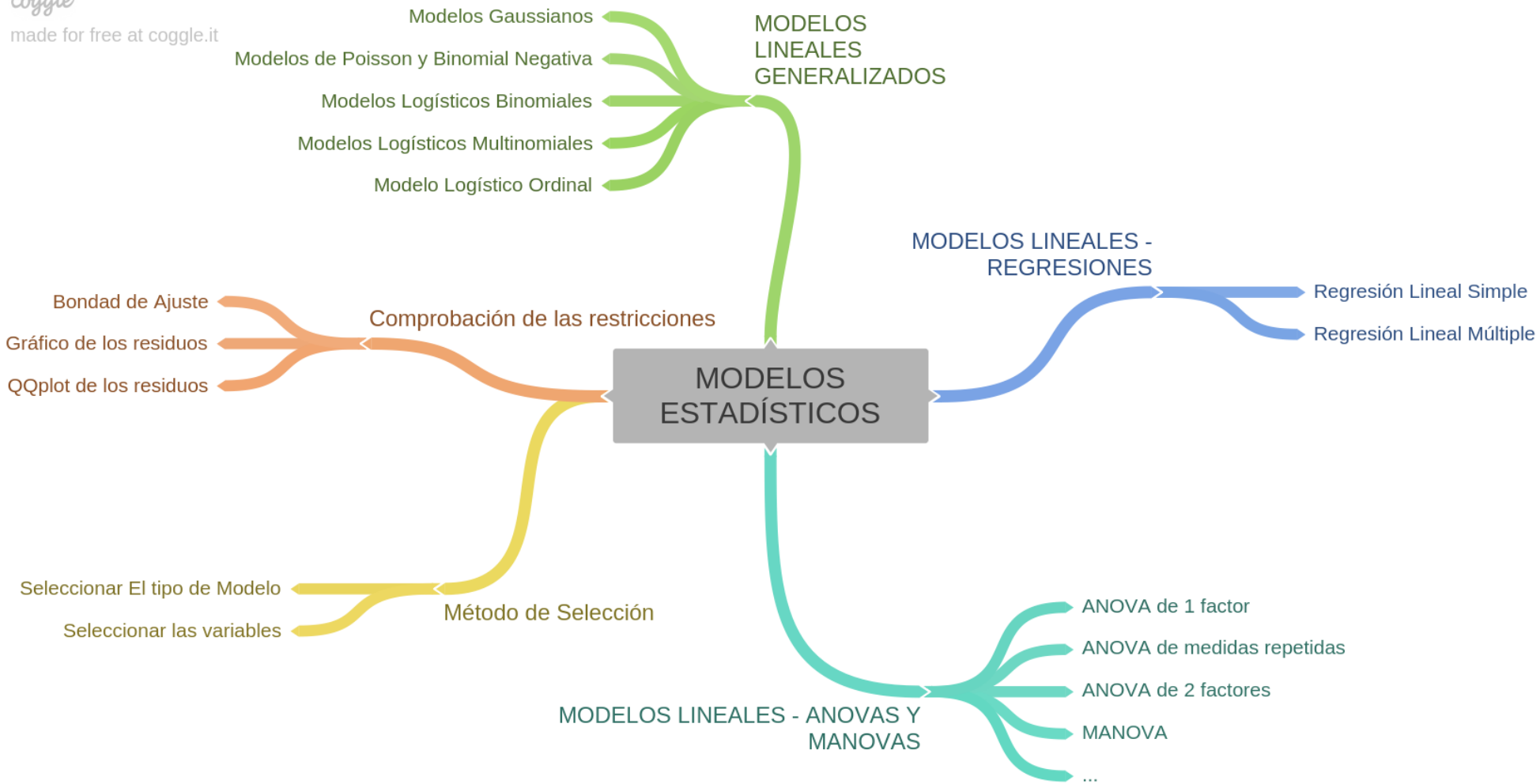
Variable de Salida

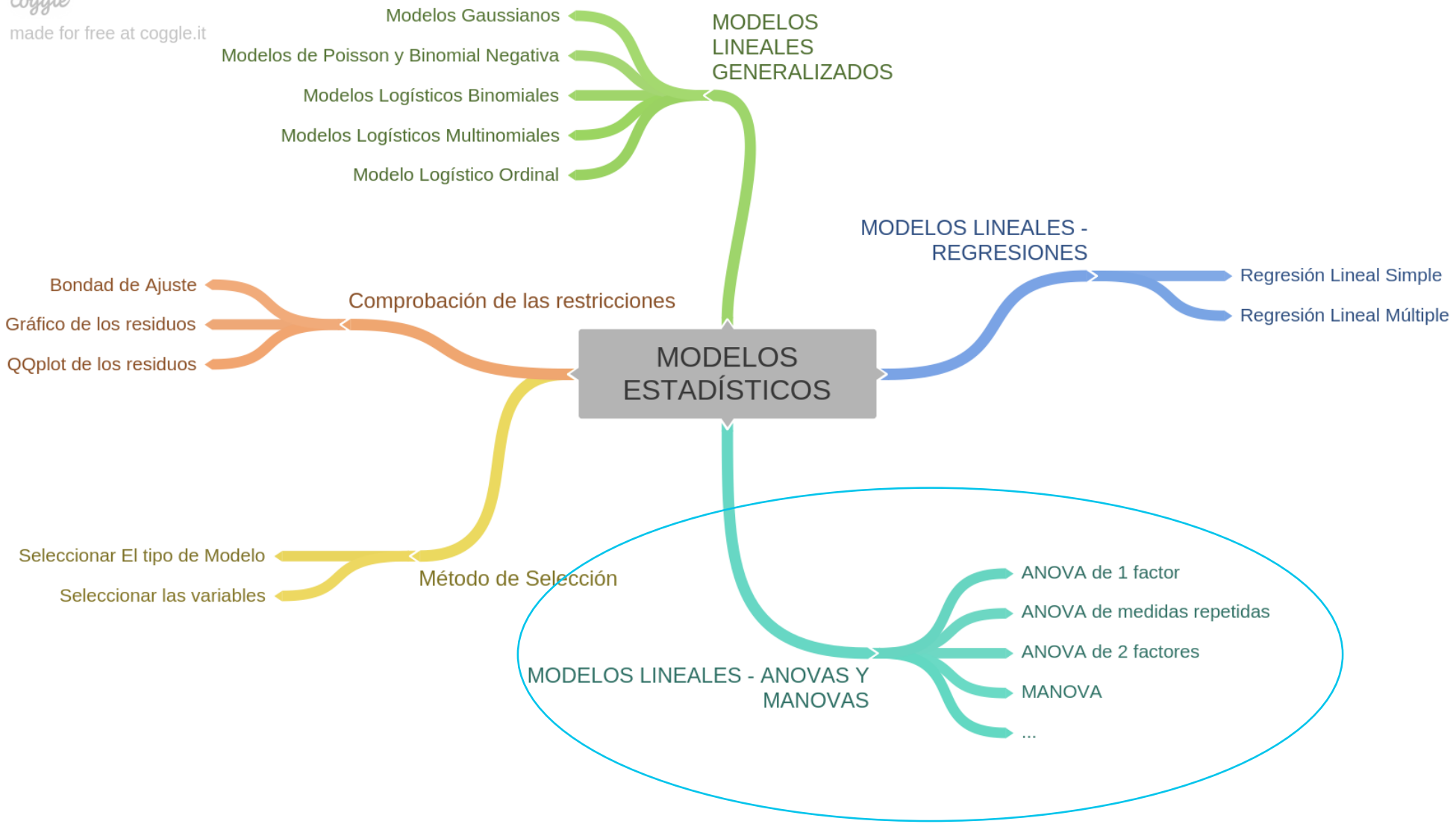
Parámetros del modelo

Variable de Entrada

El mapa de los modelos

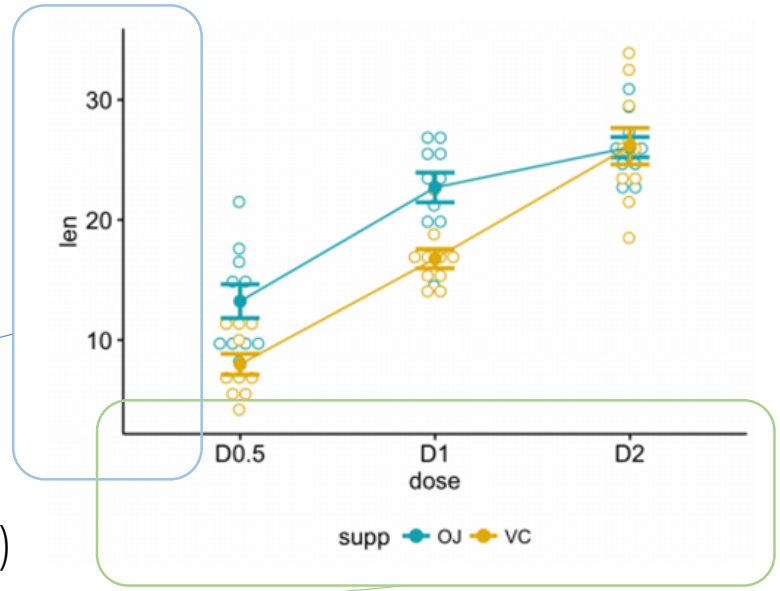
La clasificación de los modelos

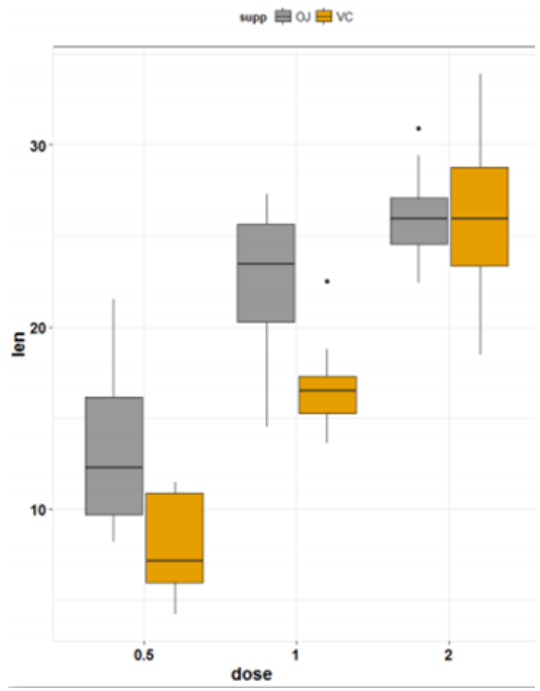




La ANOVA des del punto de vista de modelo

- Queremos saber cómo influye la cantidad suministrada (“Dose”) y el tipo de sustancia (“Supp”) en el crecimiento de los dientes:
  - medida → Log. del diente “len”
  - 2 factores →
    - Tipo de Suplemente “Supp”
    - Dosis suministrada “Dose”





Variable Dependiente o de Salida (Respuesta)

Variable Independiente o de entrada (Estudio)

La ANOVA des del punto de vista de modelo

Modelo 1 – ANOVA de dos factores

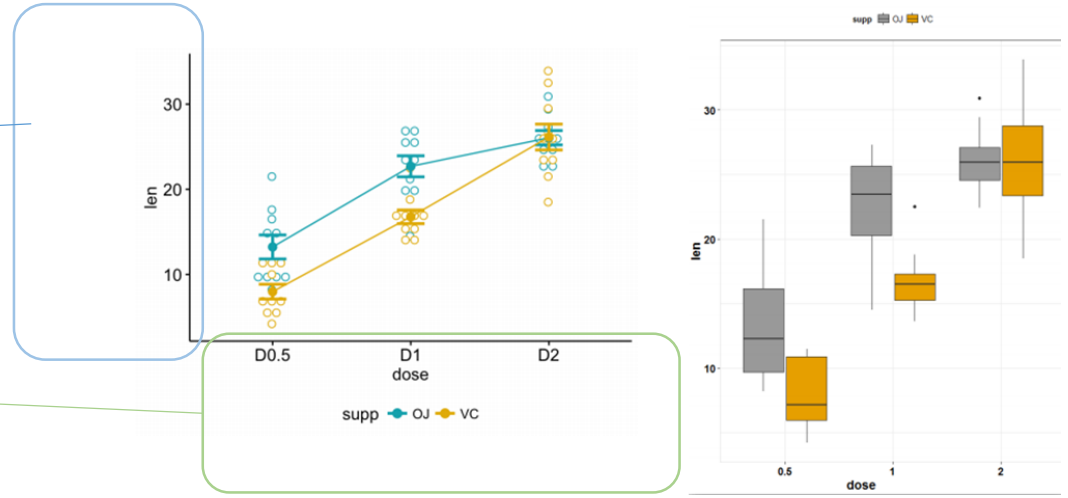
```
            Df Sum Sq Mean Sq F value   Pr(>F)
supp         1  205.4   205.4   14.02 0.000429 ***
dose         2 2426.4  1213.2   82.81  < 2e-16 ***
Residuals   56  820.4    14.7
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

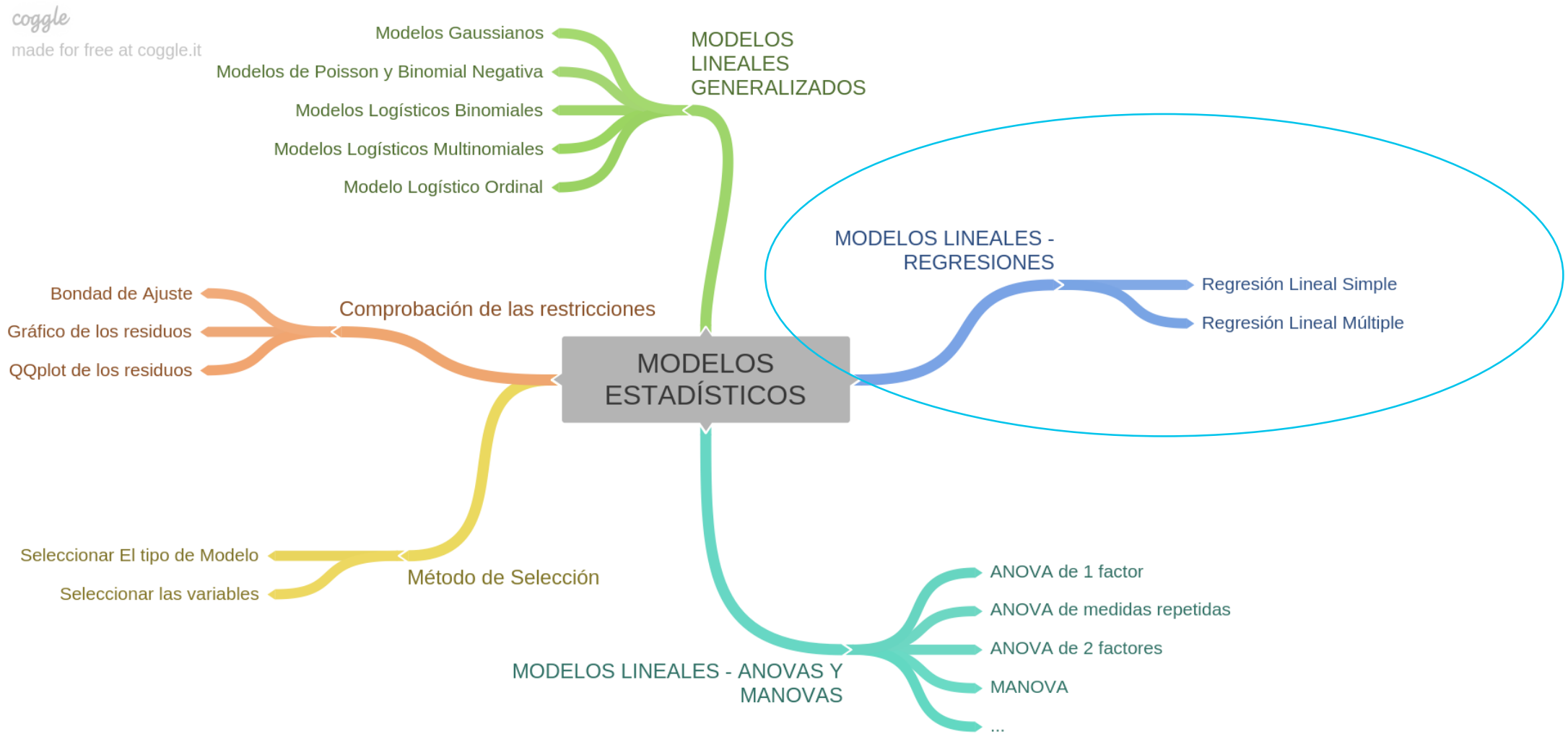
Modelo 2 – ANOVA de dos factores con interacción

```
            Df Sum Sq Mean Sq F value   Pr(>F)
supp         1  205.4   205.4  15.572 0.000231 ***
dose         2 2426.4  1213.2  92.000  < 2e-16 ***
supp:dose    2  108.3    54.2   4.107 0.021860 *
Residuals   54  712.1    13.2
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

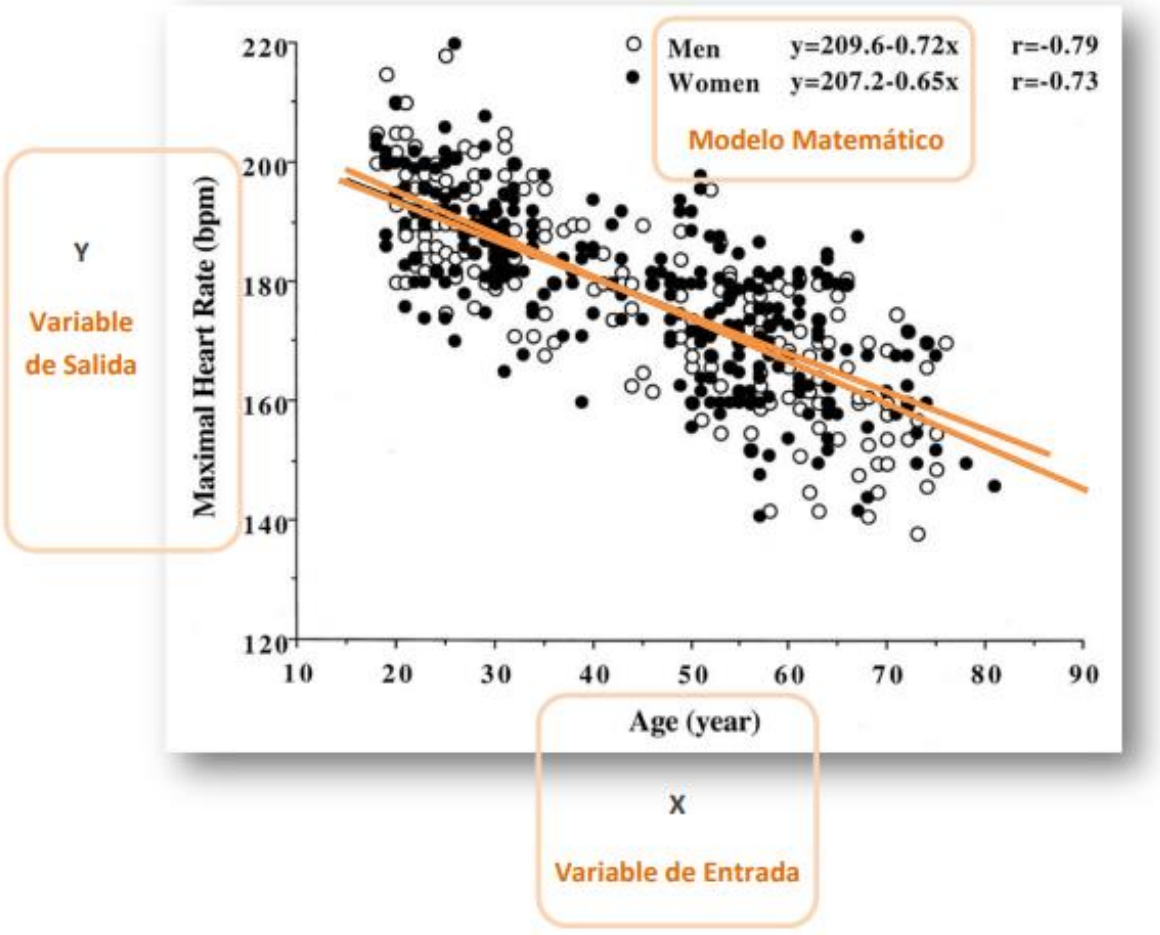
Variable Dependiente o de Salida (Respuesta)

Variable Independiente o de entrada (Estudio)




coggle
made for free at coggle.it
MODELOS ESTADÍSTICOS
MODELOS LINEALES GENERALIZADOS
Modelos Gaussianos
Modelos de Poisson y Binomial Negativa
Modelos Logísticos Binomiales
Modelos Logísticos Multinomiales
Modelo Logístico Ordinal
MODELOS LINEALES - REGRESIONES
Regresión Lineal Simple
Regresión Lineal Múltiple
Comprobación de las restricciones
Bondad de Ajuste
Gráfico de los residuos
QQplot de los residuos
Método de Selección
Seleccionar El tipo de Modelo
Seleccionar las variables
MODELOS LINEALES - ANOVAS Y MANOVAS
ANOVA de 1 factor
ANOVA de medidas repetidas
ANOVA de 2 factores
MANOVA
...

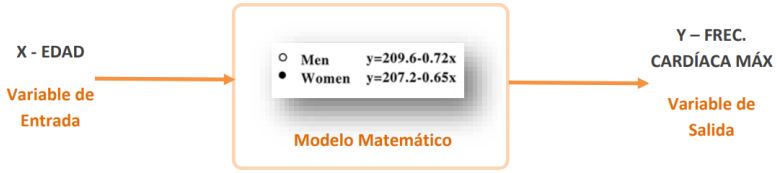
El modelo más sencillo la regresión lineal



- El **modelo es una recta** en este caso (es el modelo más sencillo)

$$y = a \cdot x + b$$

  - **a**: es la pendiente de la recta
  - **b**: es dónde cruza la recta en el eje vertical
- A partir de un cálculo de inferencia podemos obtener el valor de los parámetros del modelo **a** y **b** que mejor ajustan los datos de hombre y en mujeres
- **a** = 209,6 y **b** = -0,72 para los datos de **hombres**
- **a** = 207,2 y **b** = -0,65 para los datos de **mujeres**



$$Frec.Cardíaca\ Máxima = -0{,}72 \cdot Edad + 209{,}6 + Error$$

Variable de Salida

Parámetros del modelo

Variable de Entrada

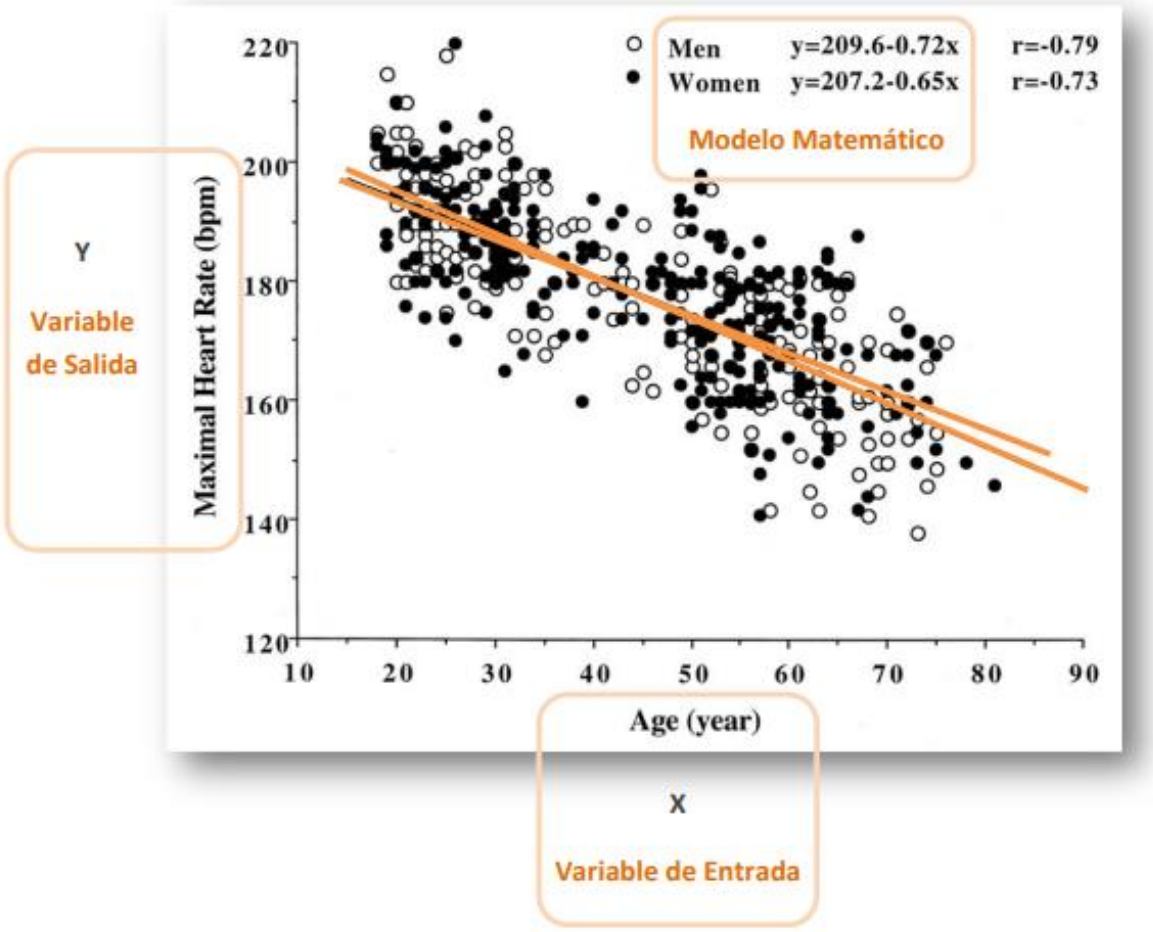
El modelo más sencillo
La regresión lineal

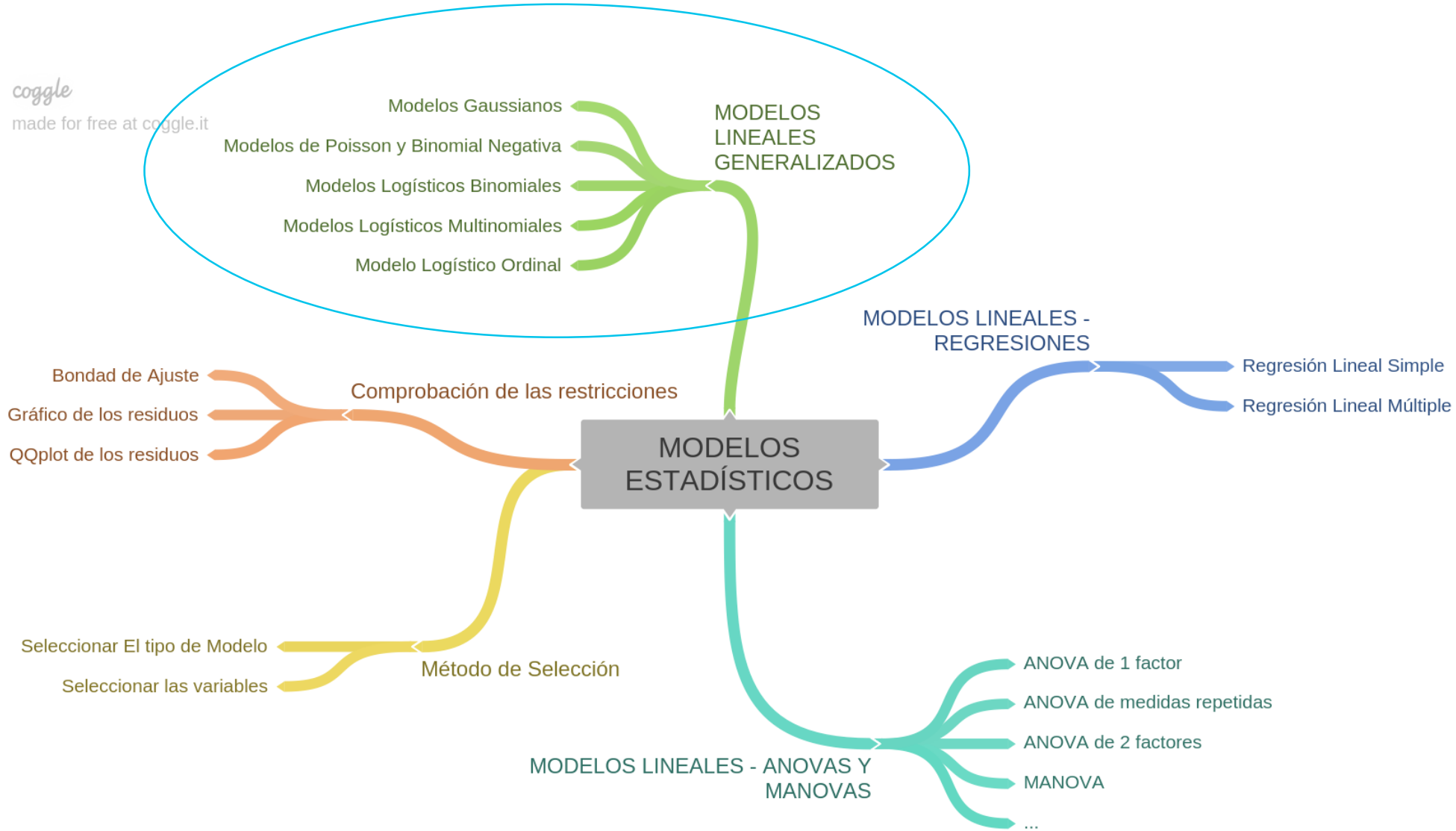
Restricciones:

- Variables continuas de entrada y salida
- Variables Gaussianas
- Minimizando el error se obtienen los parámetros

Problema:

- Poco flexibles





Generalizando las regresiones lineales: GLM

Modelos lineales Generalizados - GLM

Engloban todos los modelos lineales: ANOVA y todo tipo de regresiones

Las restricciones:

Los residuos normales

Igualdad de varianzas en los residuos

La gran ventaja

Flexibilidad y posibilidades de aplicación

$$logit\,(Grado\ de\ Abnea\ Sueño) = -6.16228 + 0.04681 * Presión\ Diastólica$$

Función de enlace — Variable Salida Y — Modelo Lineal

Variable Respuesta	Nivel de Apnea del Sueño 0 – No grave 1 – Grave
Distribución de la variable Respuesta	Bernoulli o binomial
Función de Enlace	Logística
Modelo lineal	$-6.16228 + 0.04681 * Presión\ Diastólica$ ▪ Los Coeficientes son -6.16228 y 0.04681 ▪ Variable de entrada es Presión Diastólica

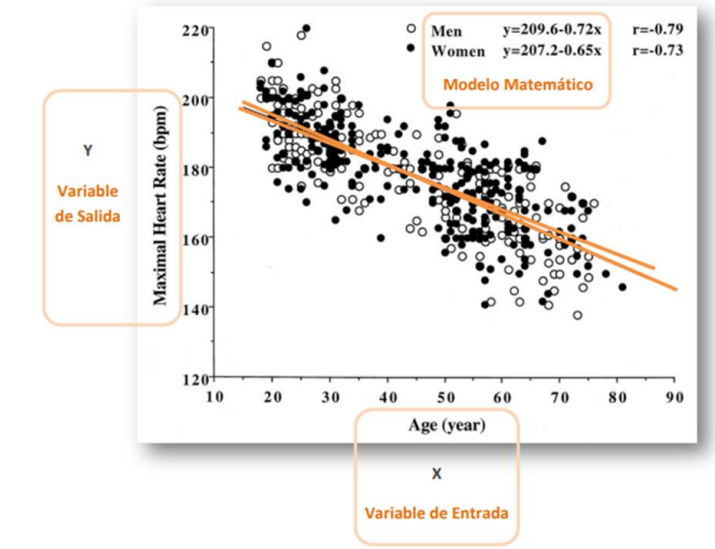
Los dos grandes retos de los modelos

RETO 1

¿QUÉ MODELO ES EL QUE MEJOR EXPLICA LOS DATOS?

¿Qué función matemática es la más adecuada?

- Tipo de función



RETO 2

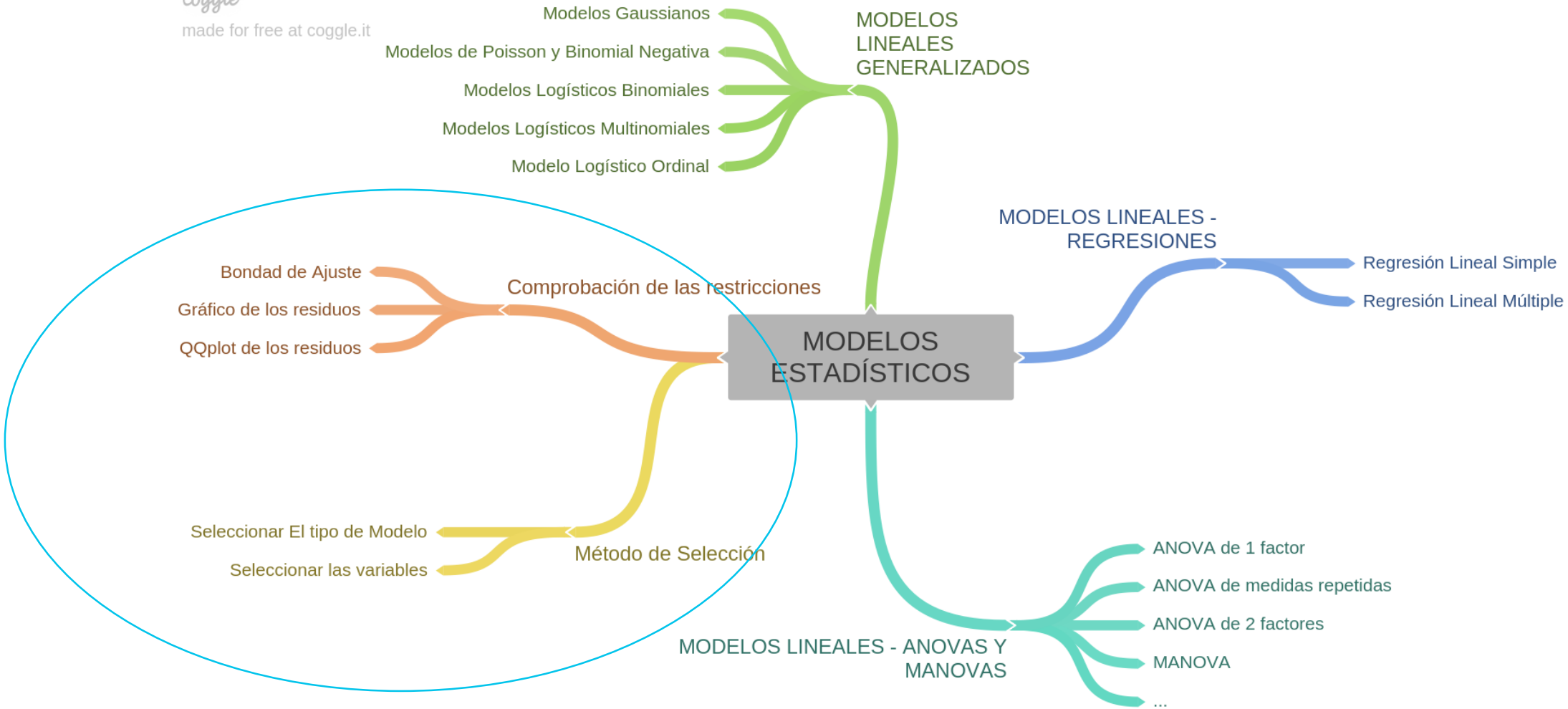
¿QUÉ VARIABLES SON LAS MÁS REPRESENTATIVAS?

¿Qué variables introduzco en el modelo?

Criterio para decidir si el modelo es válido

Probar varios modelos y escoger el mejor

Criterio para comparar los modelos



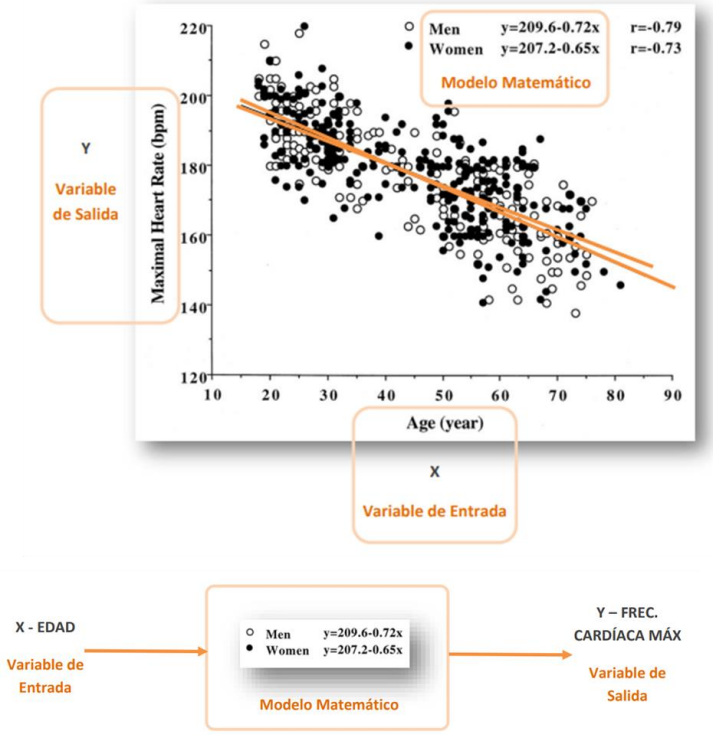
Los dos grandes retos de los modelos

RETO 1

¿QUÉ MODELO ES EL QUE MEJOR EXPLICA LOS DATOS?

Tendrás una clasificación de los modelos – Excel

Aprenderás criterios para decidir si el modelo explica bien los datos



RETO 2

¿QUÉ VARIABLES SON LAS MÁS REPRESENTATIVAS?

Puedes comparar modelos con indicadores como verás en los ejemplos (BIC, AIC)

Criterio para decidir si el modelo es válido

Probar varios modelos y escoger el mejor

Criterio para comparar los modelos

Interpretando los resultados de un modelo

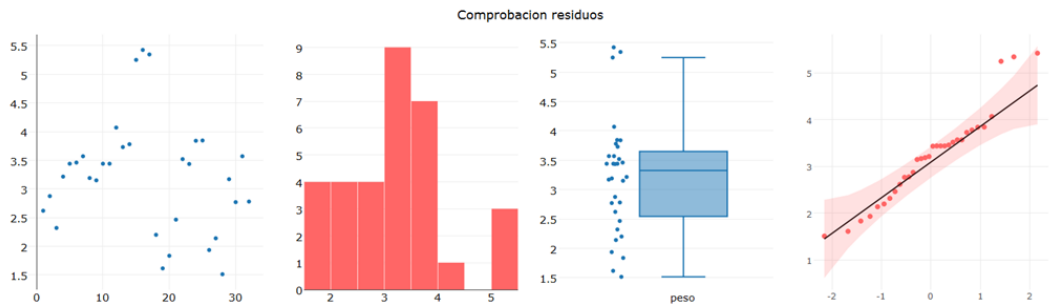
Ejemplos para entender la interpretación de los resultados

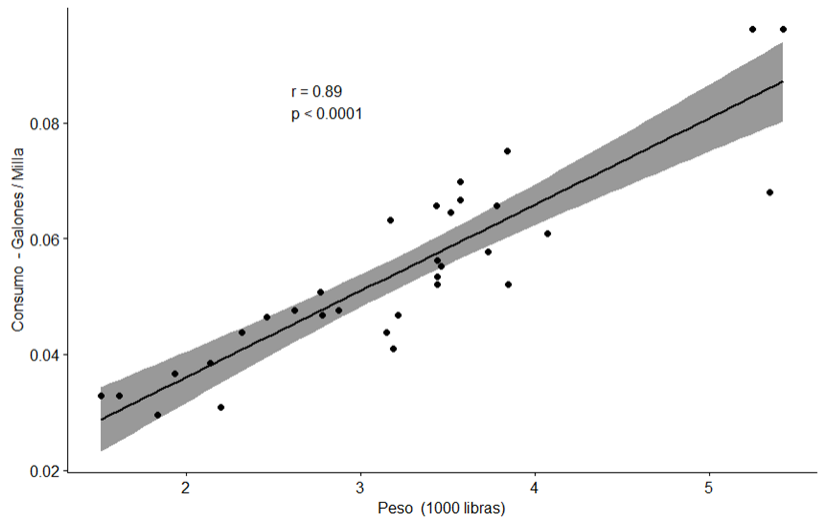
Ejemplo 1 – Regresión Lineal Simple

```
Call:
lm(formula = consumo ~ peso, data = varEstudio)

Residuals:
      Min        1Q    Median        3Q       Max
-0.0179837 -0.0044301  0.0009413  0.0045354  0.0116583

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.006169   0.004695   1.314    0.199
peso        0.014938   0.001398  10.685 9.57e-12 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.007616 on 30 degrees of freedom
Multiple R-squared:  0.7919,    Adjusted R-squared:  0.785
F-statistic: 114.2 on 1 and 30 DF,  p-value: 9.566e-12
```

Consumo = 0,006169 + 0,014938·peso + error

Variable de Entrada Variable Independientes Variable de Estudio -Eje horizontal (son sinónimos)	PESO
Variable de Salida Variable Dependiente Variable respuesta -Eje vertical (son sinónimos)	CONSUMO





consumo	peso
0.04761905	2.620
0.04761905	2.875
0.04385965	2.320
0.04672897	3.215
0.05347594	3.440
0.05524862	3.460
0.06993007	3.570
0.04098361	3.190
0.04385965	3.150
0.05208333	3.440
0.05617978	3.440
0.06097561	4.070
0.05780347	3.730
0.06578947	3.780
0.09615385	5.250
0.09615385	5.424
0.06802721	5.345
0.03086420	2.200
0.03289474	1.615

Ejemplo 2 – Regresión Logística Binomial

```
Call:
glm(formula = IAH30 ~ TAS_m, family = "binomial", data = varEstudio)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.9949  -1.0487   0.6106   1.0163   1.7737

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -6.16228    1.47788  -4.170 3.05e-05 ***
TAS_m        0.04681    0.01093   4.284 1.84e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 218.12  on 157  degrees of freedom
Residual deviance: 196.42  on 156  degrees of freedom
  (1 observation deleted due to missingness)
AIC: 200.42

Number of Fisher Scoring iterations: 4
```

Logit (IAH30) = -6.16228 + 0.04681*Presión + error

Variable de Entrada Variable Independientes Variable de Estudio -Eje horizontal (son sinónimos)	Presión – TAS_m
Variable de Salida Variable Dependiente Variable respuesta -Eje vertical (son sinónimos)	Gravedad del paciente - IAH30

En este caso las variables son las siguientes:

- IAH30: pacientes con valores altos de apnea. Gravedad del paciente.
  - 0 = No grave
  - 1 = grave
- TAS_m: presión sistólica

TAS_m	IAH30
160	1
130	1
150	1
132	0
166	1
147	0
138	1
143	0
165	1
165	0

Take away

El resumen de la lección

Lo más importante de la lección

- Los modelos estadísticos explican los datos con un modelo matemático o función
- Las tres partes fundamentales de un modelo son:
  - Variable de salida – independientes o respuesta (eje vertical)
  - Variable de entrada – dependiente o de estudio (eje horizontal)
- Las dos poderes de un modelo:
  - Predicción (con un error)
  - Relación causa-efecto entre variables
- Todos los modelos lineales se engloban en GLM – son más flexibles
- Importante aprender a:
  - Decidir la mejor estructura de un modelo
  - Comprobar si un modelo es correcto y está explicando bien los datos
  - Decidir qué modelo es el mejor
- Para tu tranquilidad, la interpretación de estos modelos es bastante similar

Tú turno

A por los conceptos claros

A poner en práctica lo que has visto

- Descarga la hoja de trabajo
- Empieza a pensar en modelos y su estructura

- ¡Te ayudará mucho a familiarizarte!