

# **Curso-Taller de R para investigadores**

**UMSA, La Paz, Bolivia 23 - 25 Feb 2023**

**UMSS, Cochabamba, Bolivia 6 - 8 Mar 2023**

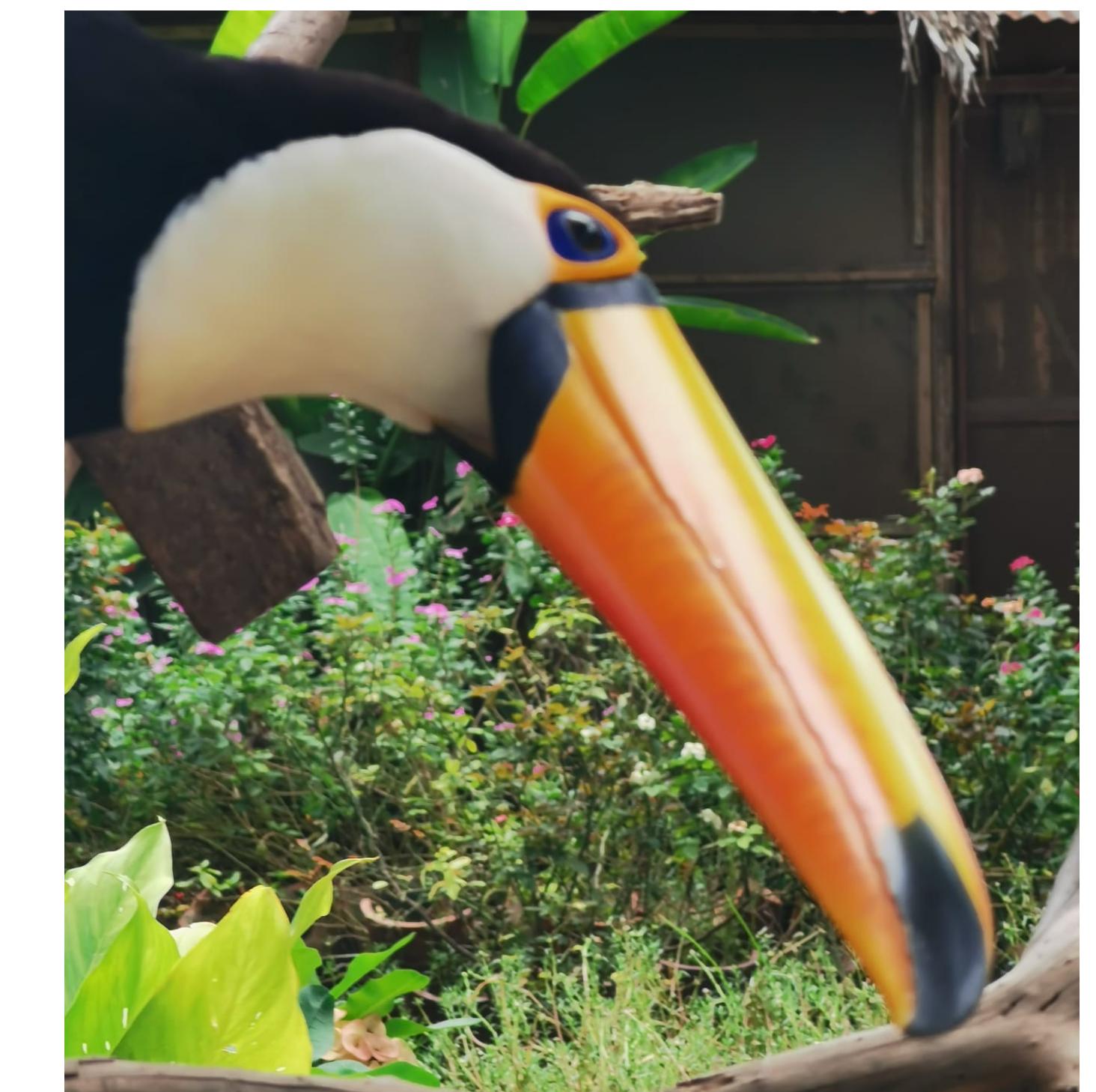
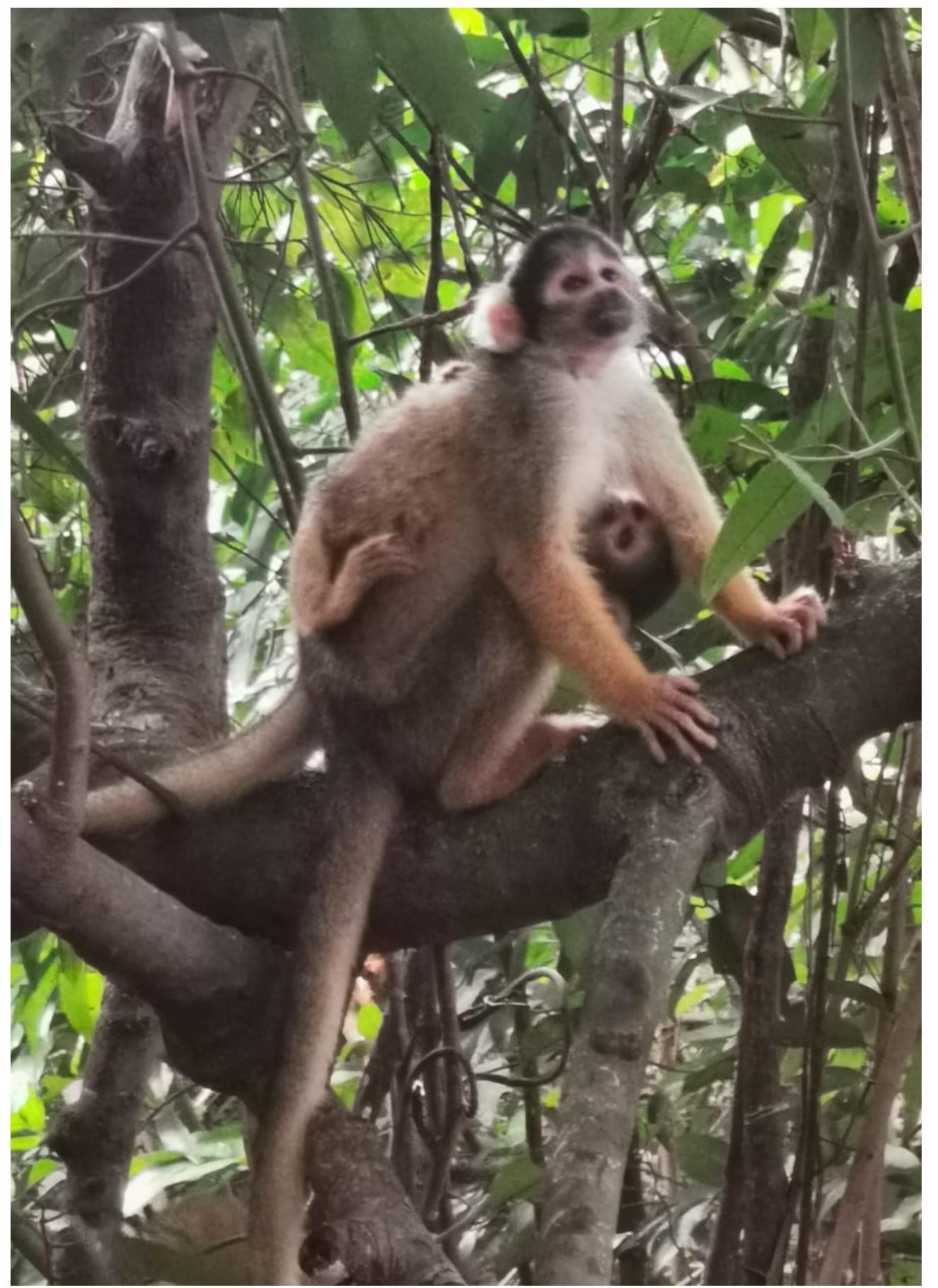
---

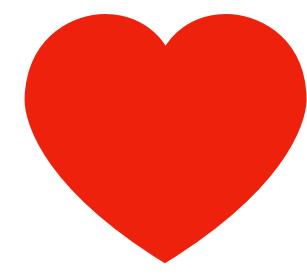
**Saras Windecker & David Uribe**

# Código de Conducta

<https://github.com/smwindecker/R-para-ecologia>







Bolivia



---

**Qué esperan aprender en este curso?**

**Con qué tipos de datos trabajan?**

## Introducción

---

aprender nueva terminología  
identificar modelos apropiados para sus datos  
entender supuestos de modelos comunes  
saber dónde buscar ayuda

## Introducción

---

- Como funciona R y Rstudio (uso de funciones, paquetes, ambiente)

## Introducción

---

- Como funciona R y Rstudio (uso de funciones, paquetes, ambiente)
- Análisis de varianza y regresiones lineales

## Introducción

---

- Como funciona R y Rstudio (uso de funciones, paquetes, ambiente)
- Análisis de varianza y regresiones lineales
- Supuestos y generalizaciones de los modelos lineales

## Introducción

---

- Como funciona R y Rstudio (uso de funciones, paquetes, ambiente)
- Análisis de varianza y regresiones lineales
- Supuestos y generalizaciones de los modelos lineales
- Distribuciones de datos y funciones de enlace

- Como funciona R y Rstudio (uso de funciones, paquetes, ambiente)
- Análisis de varianza y regresiones lineales
- Supuestos y generalizaciones de los modelos lineales
- Distribuciones de datos y funciones de enlace
- Efectos fijos y efectos aleatorios

- Como funciona R y Rstudio (uso de funciones, paquetes, ambiente)
- Análisis de varianza y regresiones lineales
- Supuestos y generalizaciones de los modelos lineales
- Distribuciones de datos y funciones de enlace
- Efectos fijos y efectos aleatorios
- Diagnosticando y evaluando modelos lineales

- Como funciona R y Rstudio (uso de funciones, paquetes, ambiente)
- Análisis de varianza y regresiones lineales
- Supuestos y generalizaciones de los modelos lineales
- Distribuciones de datos y funciones de enlace
- Efectos fijos y efectos aleatorios
- Diagnosticando y evaluando modelos lineales
- Interpretación y presentación de resultados

- Como funciona R y Rstudio (uso de funciones, paquetes, ambiente)
- Análisis de varianza y regresiones lineales
- Supuestos y generalizaciones de los modelos lineales
- Distribuciones de datos y funciones de enlace
- Efectos fijos y efectos aleatorios
- Diagnosticando y evaluando modelos lineales
- Interpretación y presentación de resultados

\* Análisis estadístico reproducible

# I. Entendiendo los modelos

**Cómo empezamos un proyecto de investigación?**

**Modelo conceptual**

**Formular la pregunta**

**Diseño experimental**

**Recolección de datos**

**Escribir y ajustar el modelo**

**Presentar resultados**

**Modelo conceptual**

**Formular la pregunta**

**Diseño experimental**

**Recolección de datos**

**Escribir y ajustar el modelo**

**Presentar resultados**

**Modelo conceptual**

**Formular la pregunta / escribir el modelo**

**Diseño experimental**

**Recolección de datos**

**Ajustar el modelo**

**Presentar resultados**

## **Entendiendo los modelos**

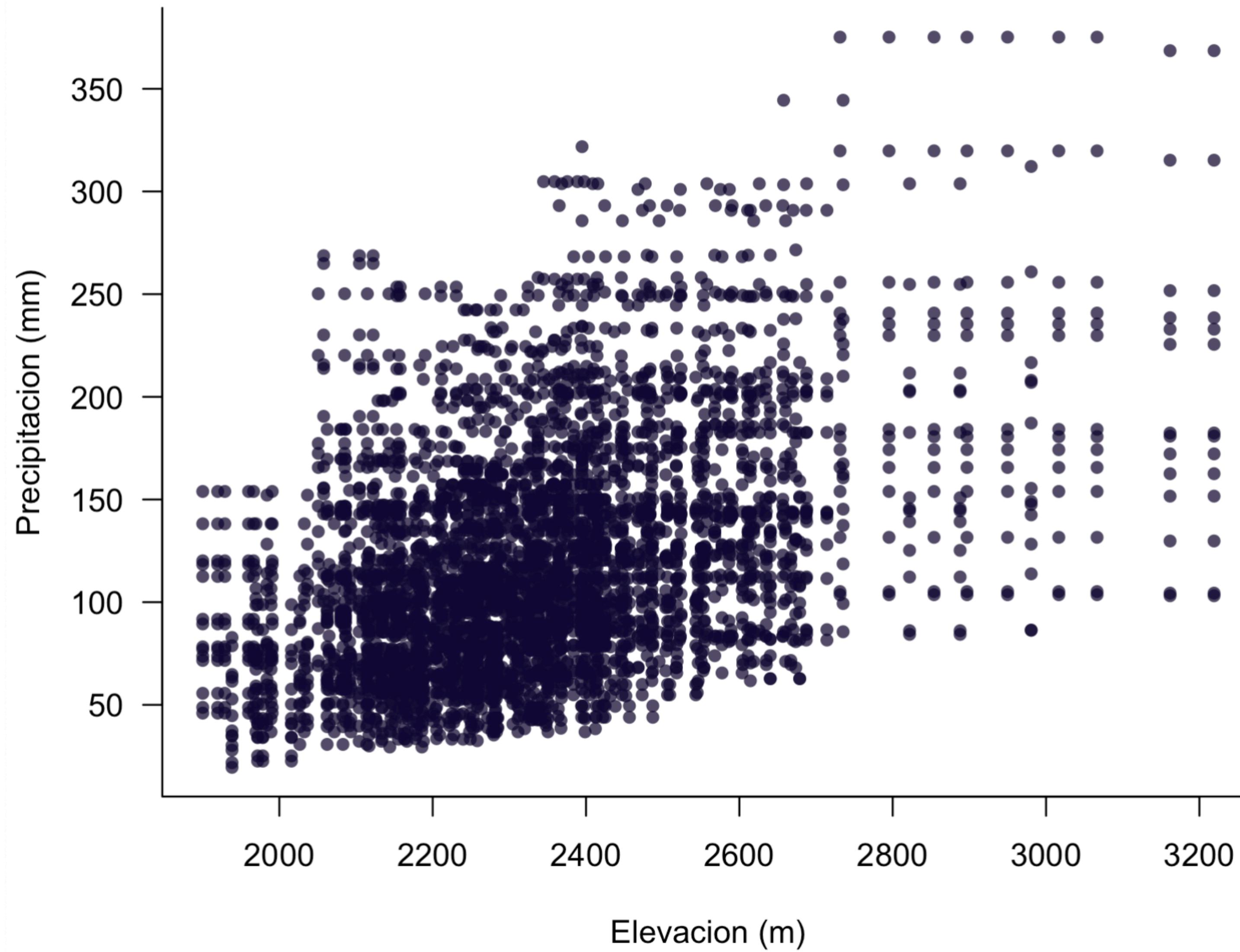
---

**Una pregunta de investigación bien formulada debería poder representarse como modelo estadístico.**

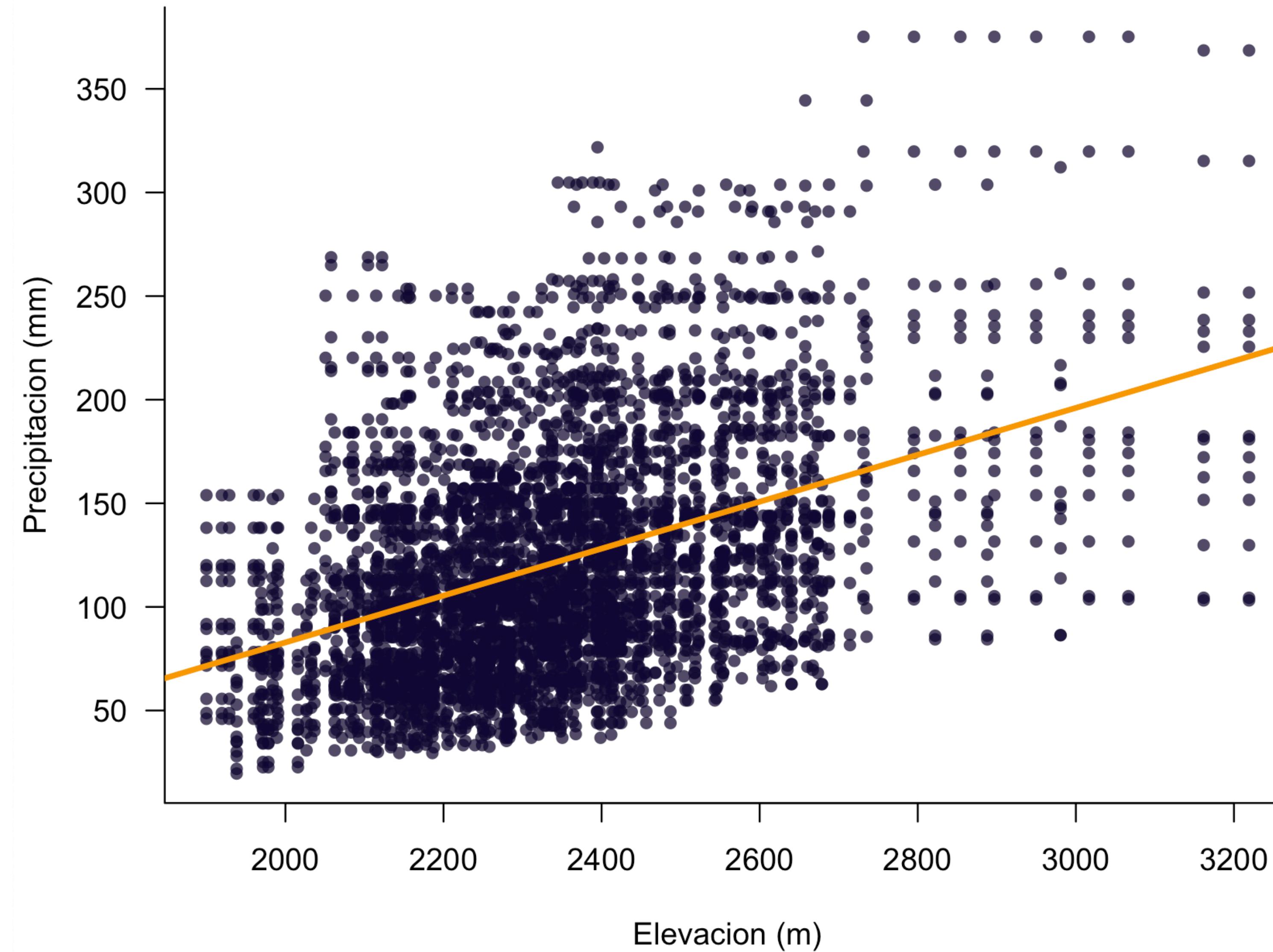
## **II. Modelos lineales, parte 1**

---

## Modelos lineales



## Modelos lineales

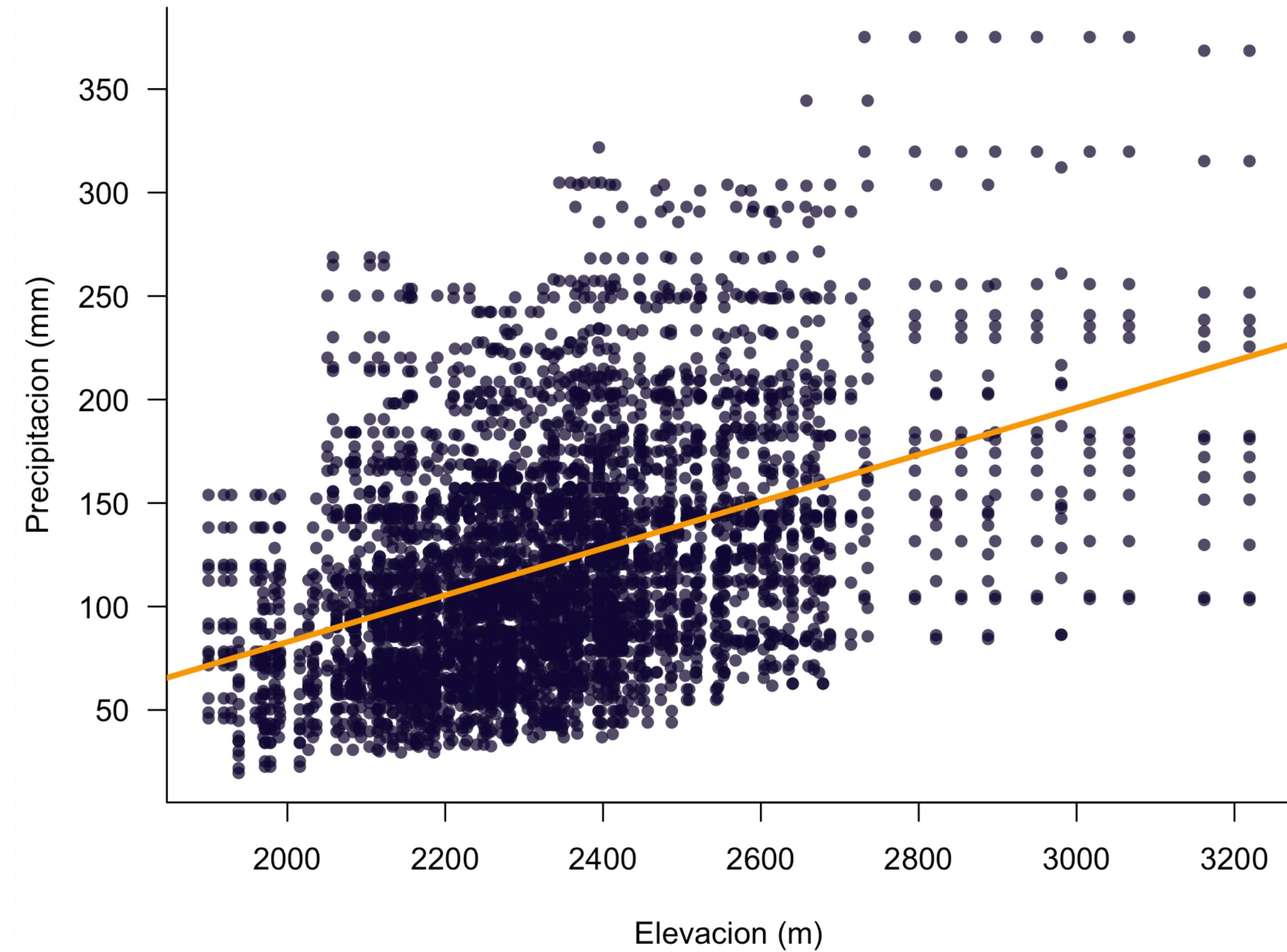


## Modelos lineales

Qué caracteriza este ejemplo?

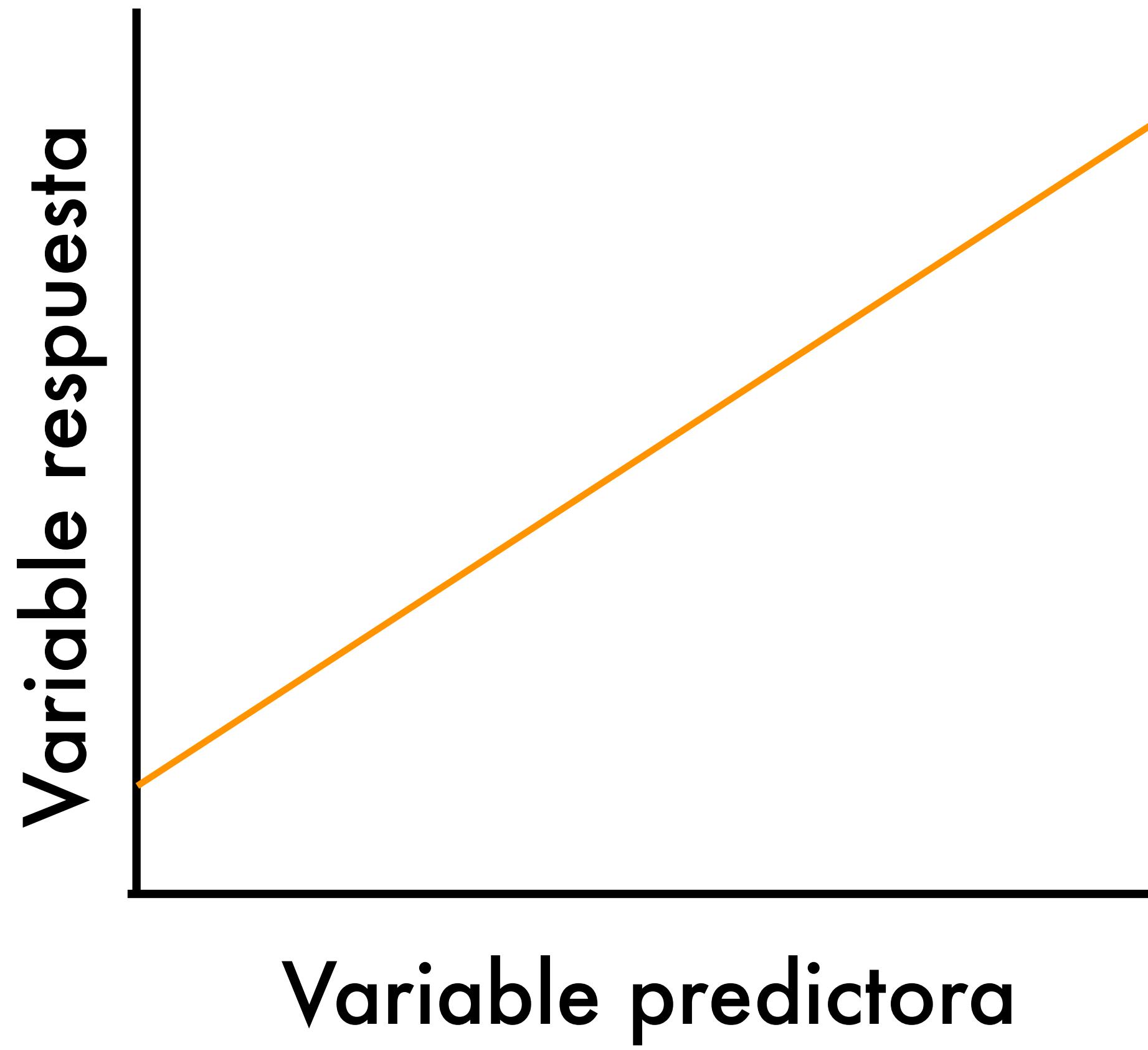
variable respuesta continua

variable predictoria continua



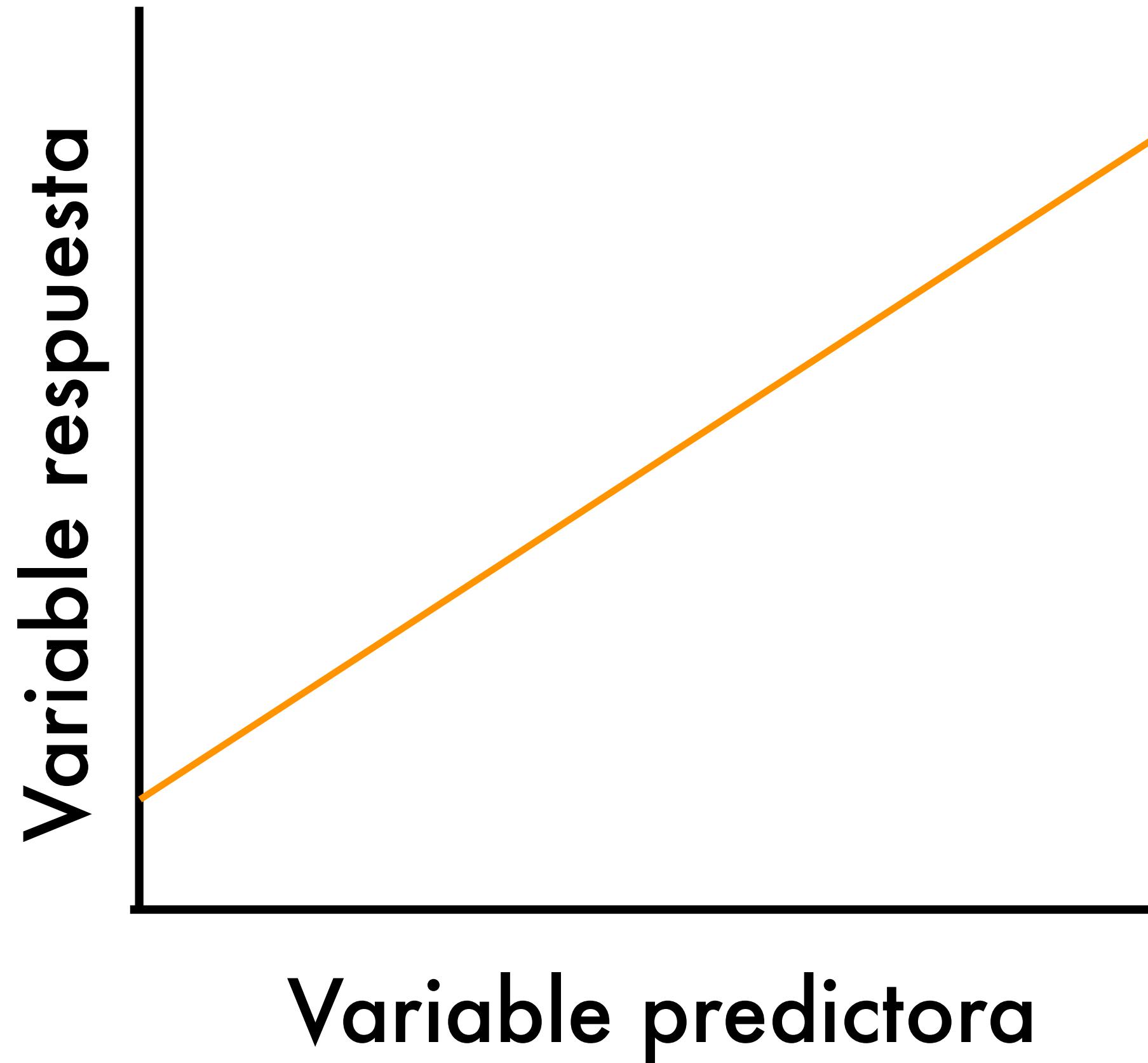
## Modelos lineales

---



## Modelos lineales

$$y = \alpha + \beta * x$$

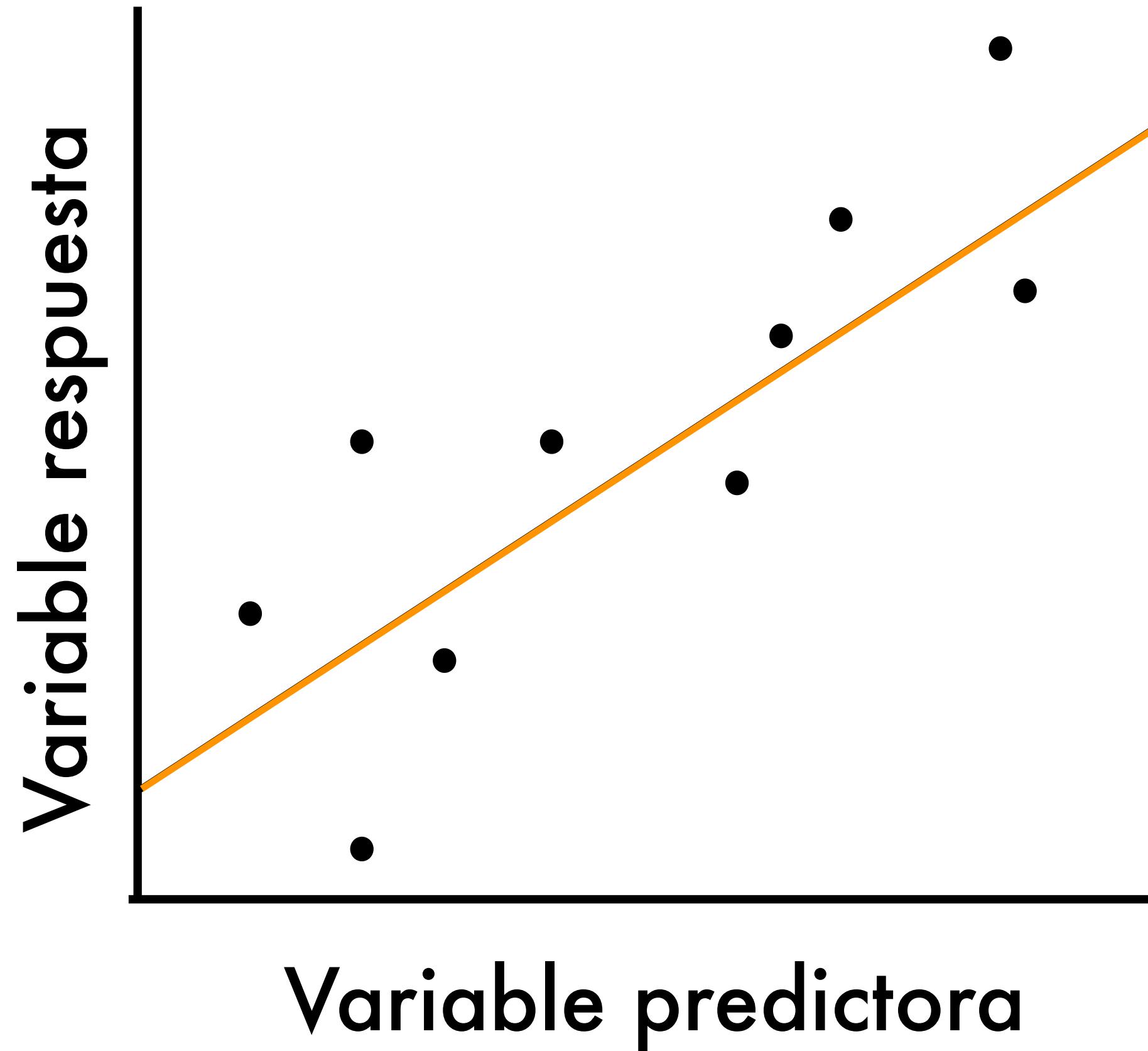


Queremos estimar los  
valores mas probables de  
 $\alpha$  y  $\beta$

## Modelos lineales

```
for ( i in 1:10 )
```

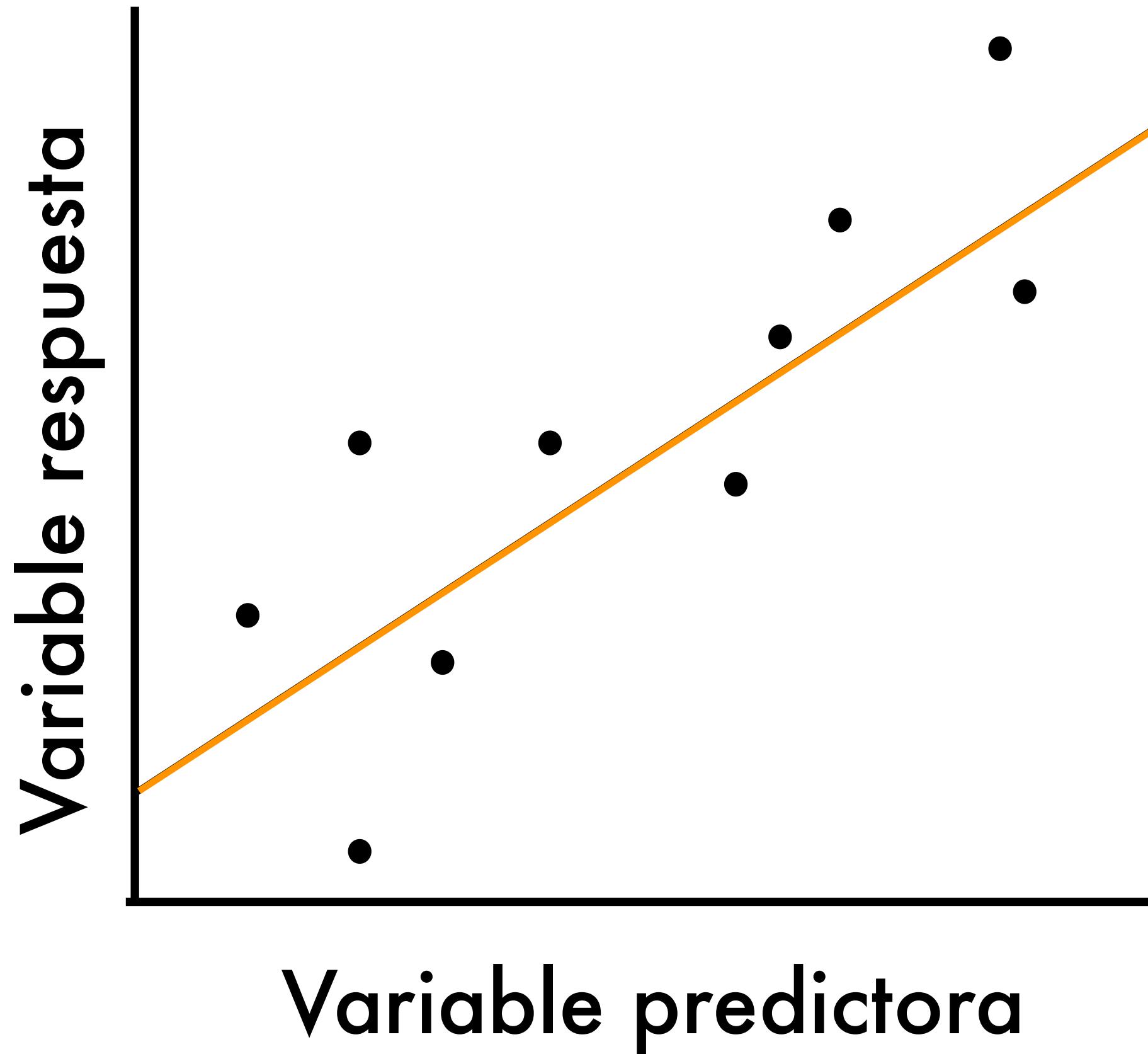
$$y_i = \alpha + \beta * x_i + \epsilon_i$$



## Modelos lineales

```
for ( i in 1:10 )
```

$$y_i = \alpha + \beta * x_i + \epsilon_i$$

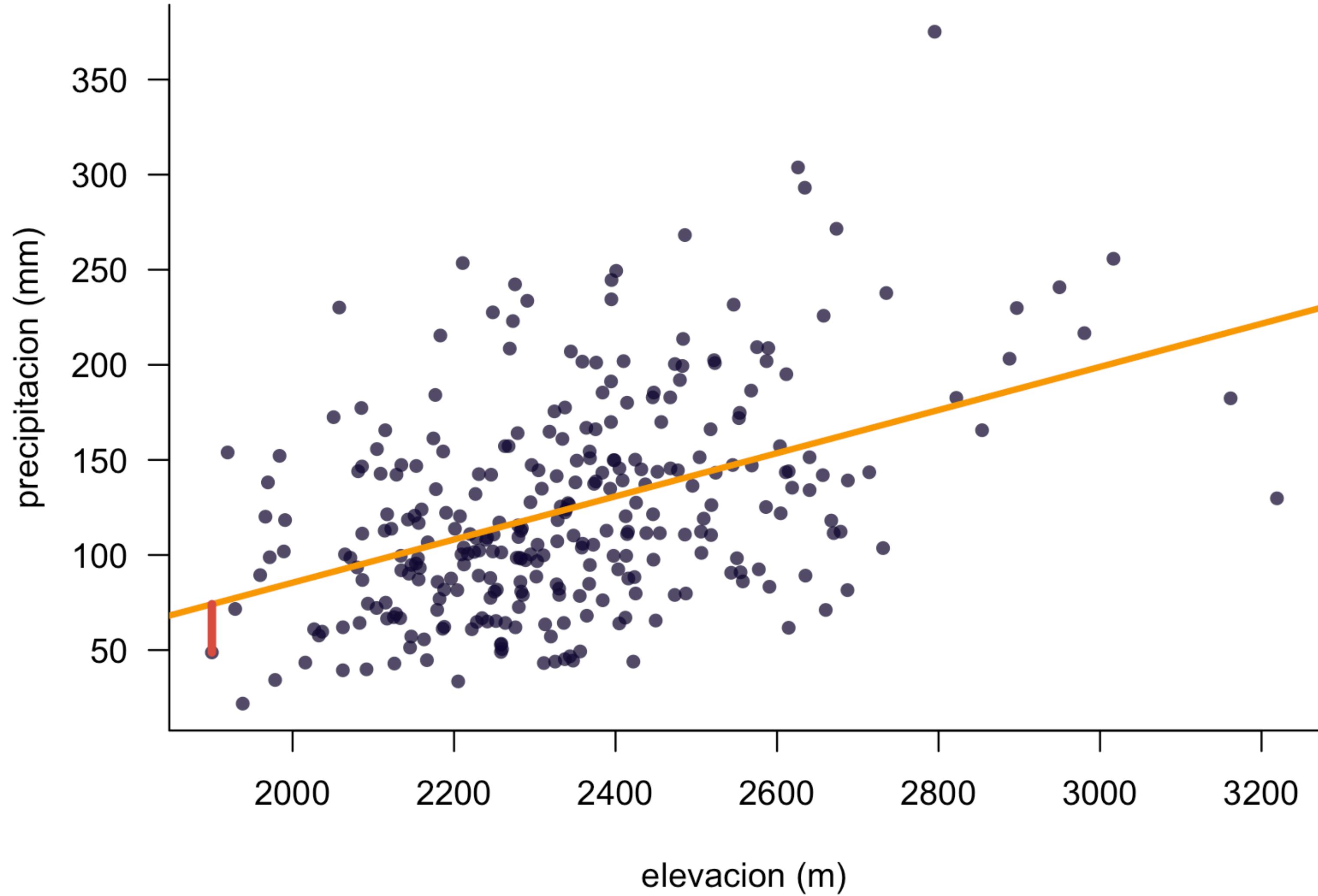


Las diferencias entre las observaciones y la linea es el error residual  $\epsilon_i$

Como podemos  
entender el  $\epsilon$ ?

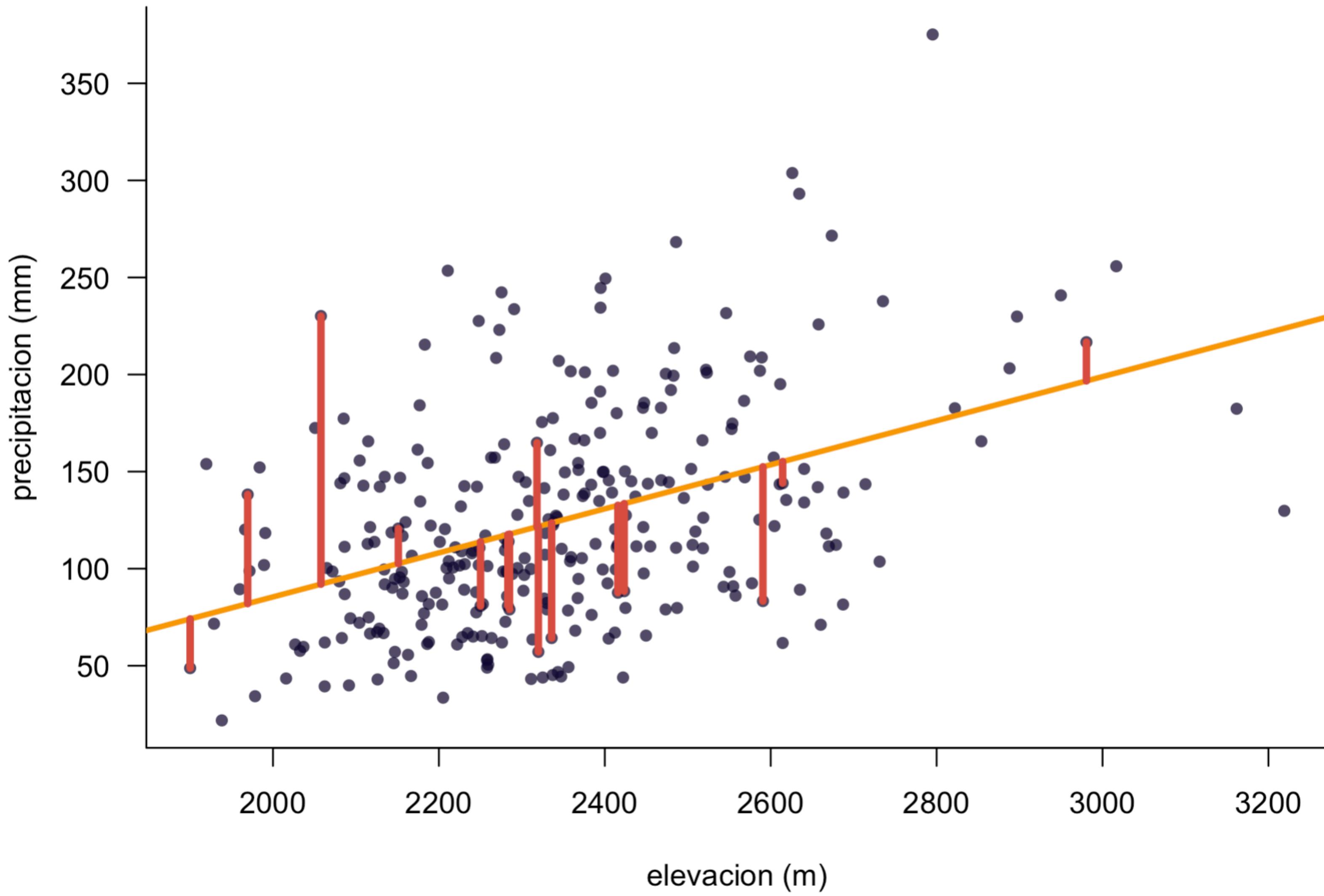
## Modelos lineales

Como podemos entender el  $\epsilon$ ?



## Modelos lineales

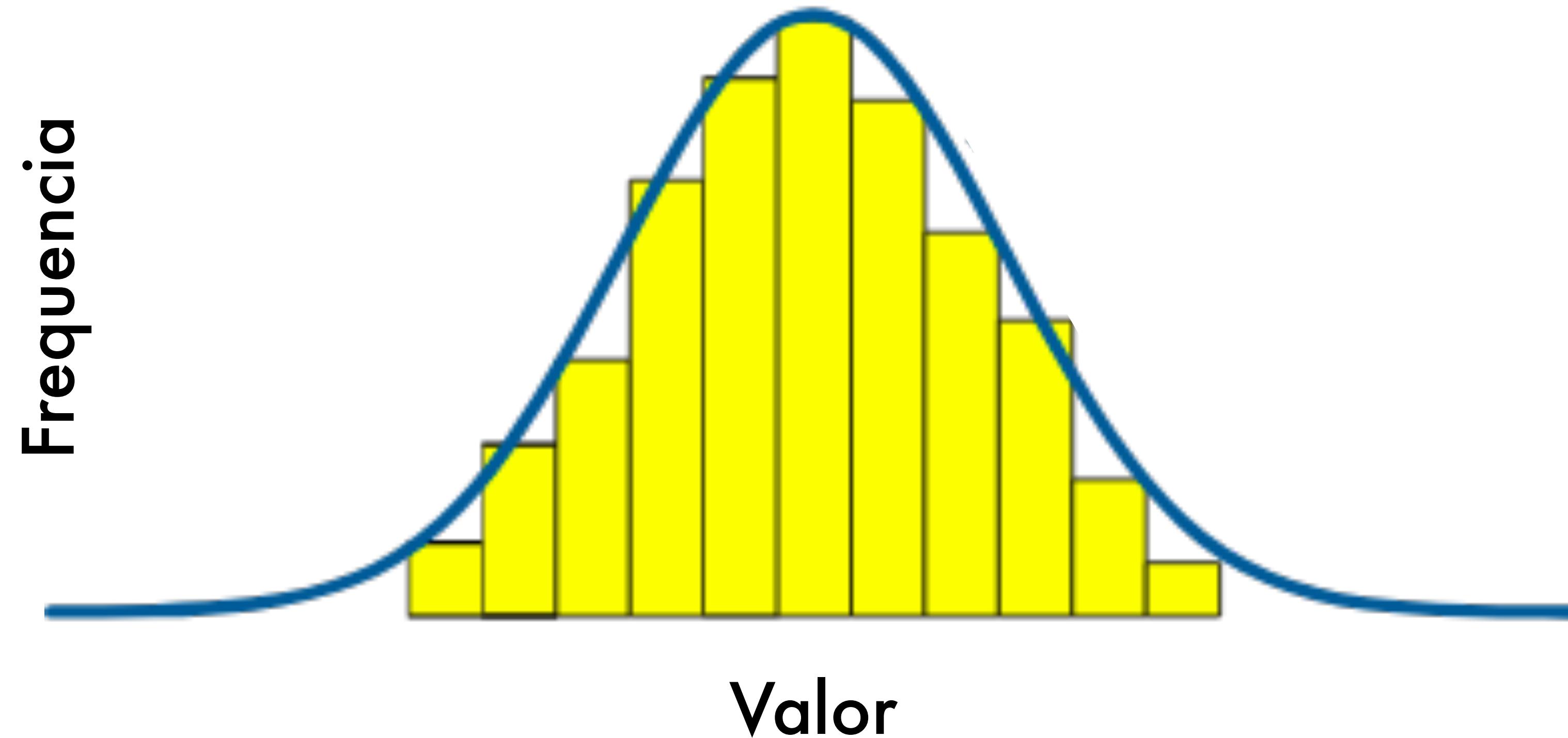
Como podemos entender el  $\epsilon$ ?



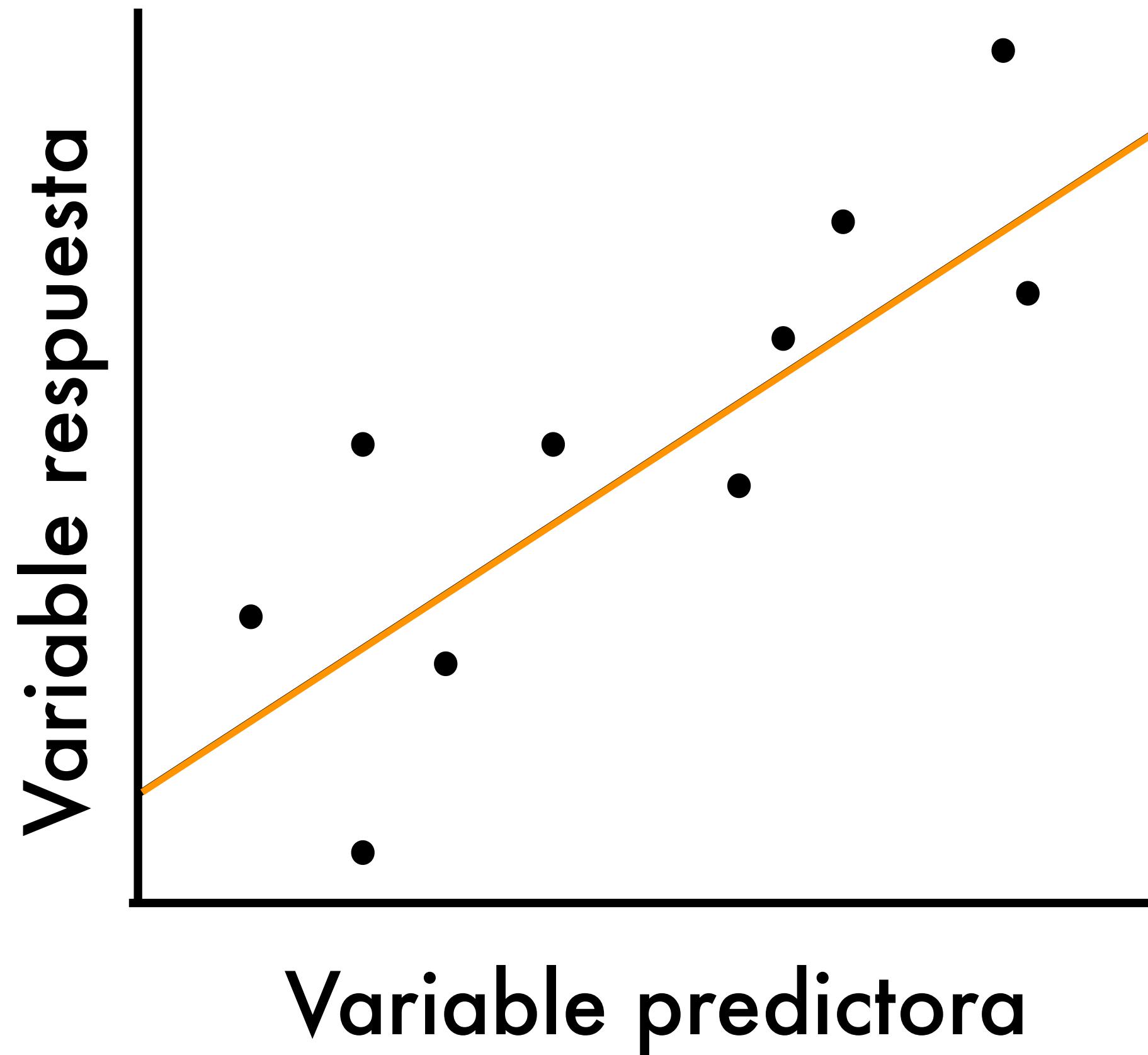
### Supuesto de regresión lineal 1: Observaciones son independientes

- cada observación provee nueva información
- observaciones no-independientes dan menos información
- buen diseño de estudio puede eliminar problemas de independencia

Supuesto 2: Residuales se ajustan a una distribución normal

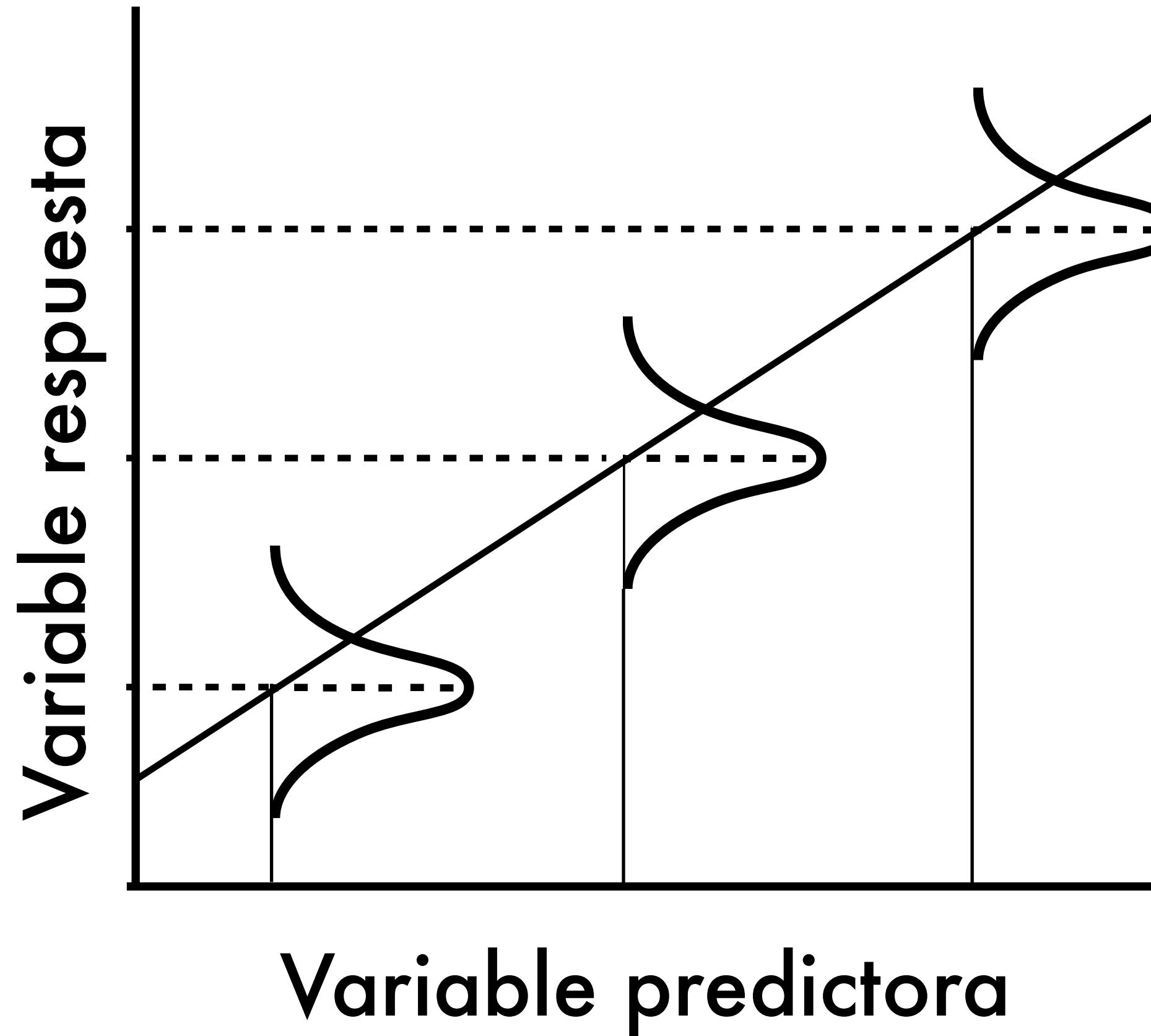


Supuesto 2: Residuales se ajustan a una distribución normal



$$y_i = \alpha + \beta * x_i + \epsilon_i$$

### Supuesto 2: Residuales se ajustan a una distribución normal

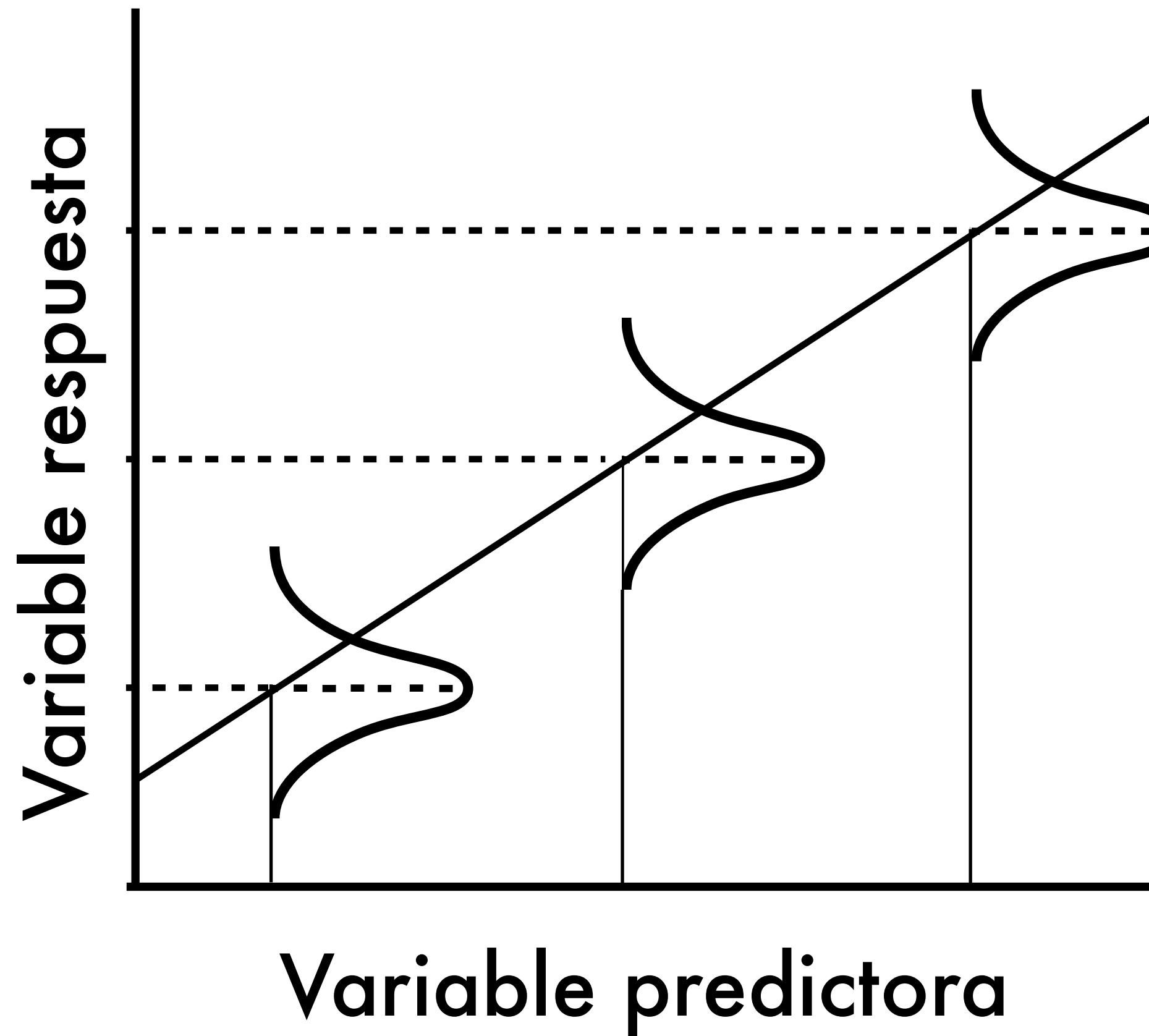


$$y_i = \alpha + \beta * x_i + \epsilon_i$$

$$\epsilon_i \sim N(0, \sigma^2)$$

La ubicación mas probable para cada valor de  $x$ , se encuentra en la recta.

### Supuesto 2: Residuales se ajustan a una distribución normal

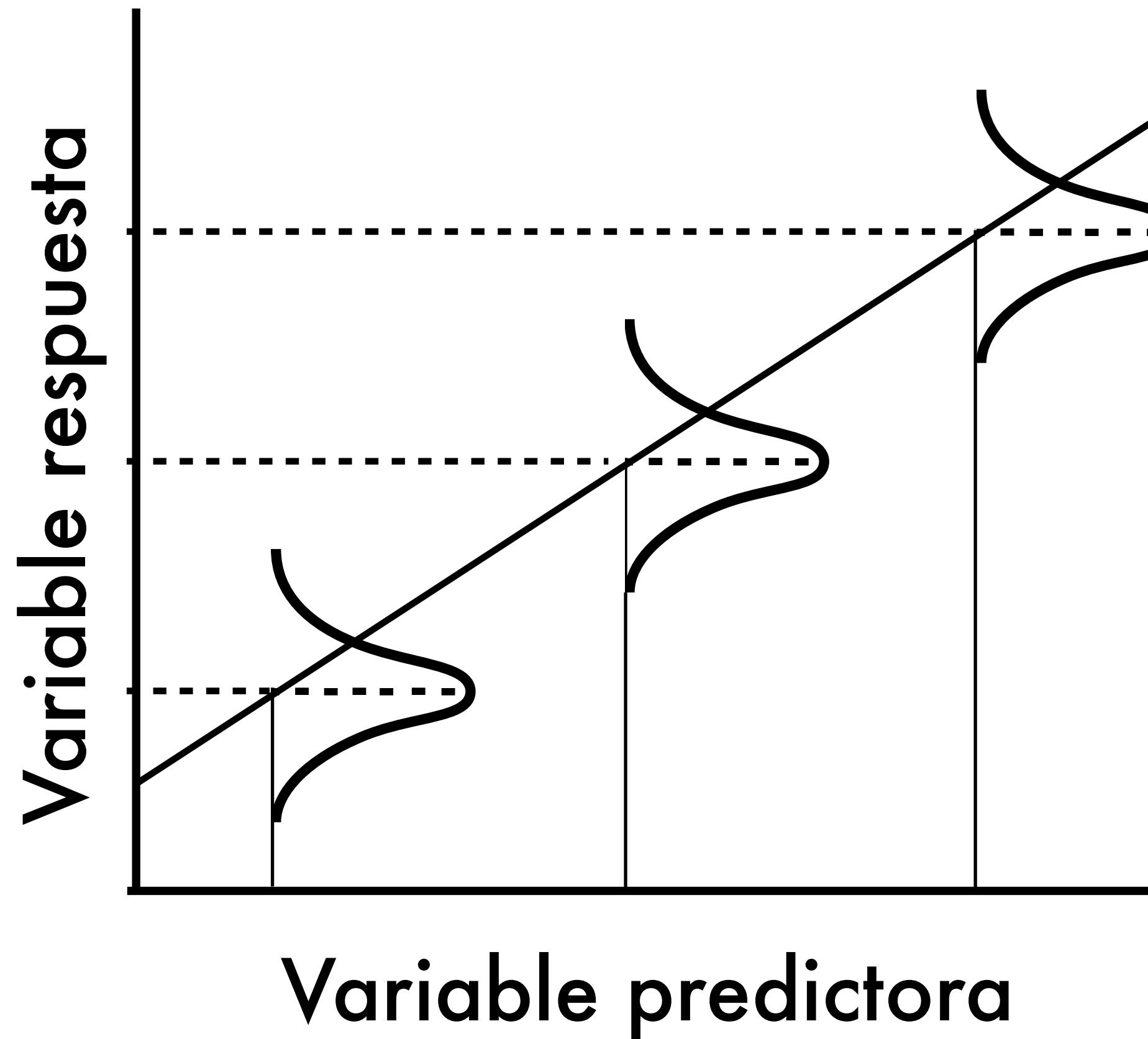


$$y_i = \alpha + \beta * x_i + \epsilon_i$$

$$\epsilon_i \sim N(0, \sigma^2)$$

known

### Supuesto 2: Residuales se ajustan a una distribución normal



$$y_i = \alpha + \beta * x_i + \epsilon_i$$

$$\epsilon_i \sim N(0, \sigma^2)$$

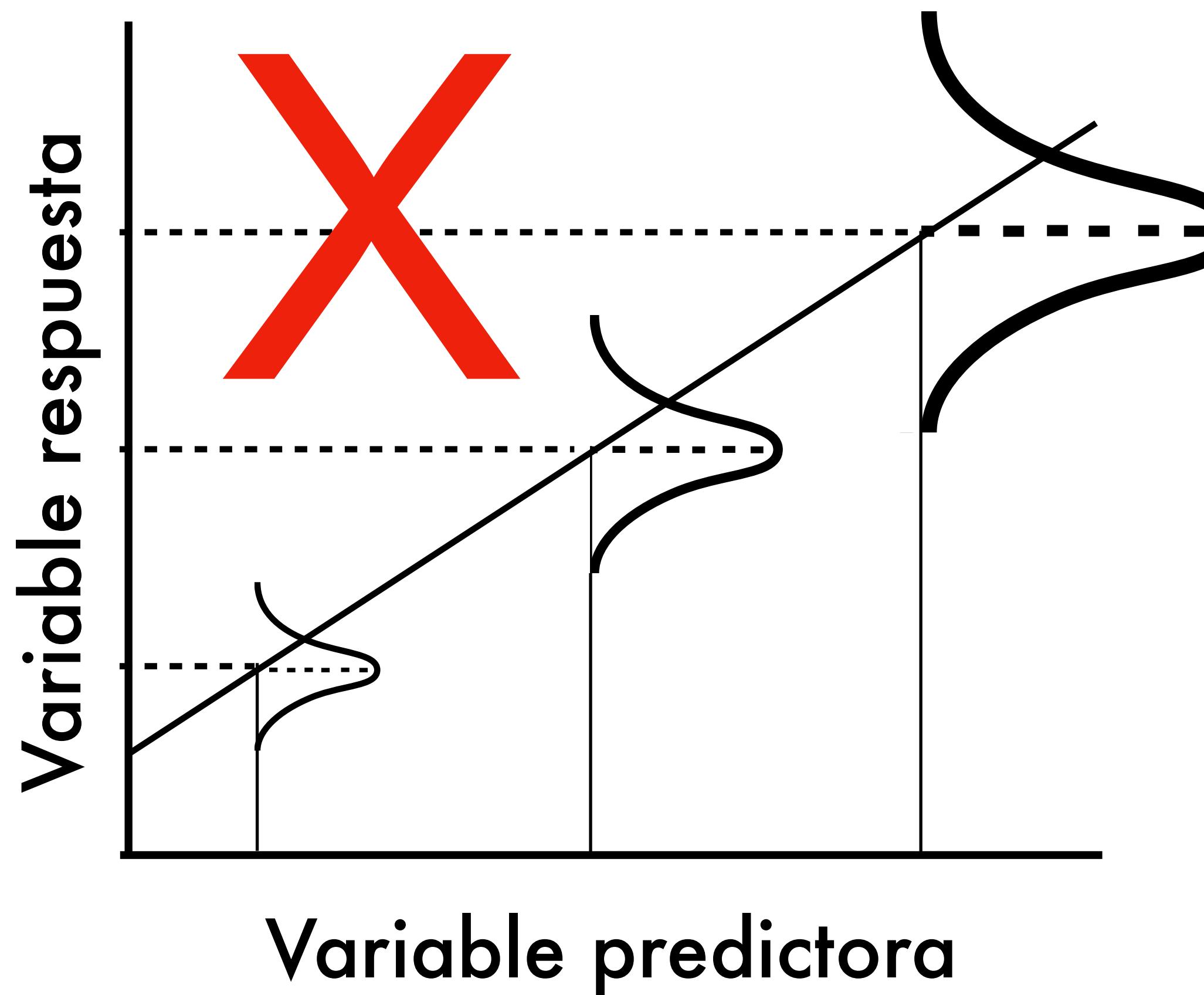
known

estimated

### Supuesto 2: Residuales se ajustan a una distribución normal

- Siempre hay que poner a prueba el supuesto, ya sea visualmente o con una prueba de ajuste (por ej. Kolmogorov-Smirnov)
- Si el supuesto no se cumple, se puede usar transformaciones, o modelos lineales generalizados

### Supuesto 3: Homocedasticidad



- La varianza de los residuales debe ser constante a través de la variable predictora

## **Normalidad de variables predictoras**

---

**Normalidad de las variables predictoras >> usa transformaciones**

## **Modelos lineales**

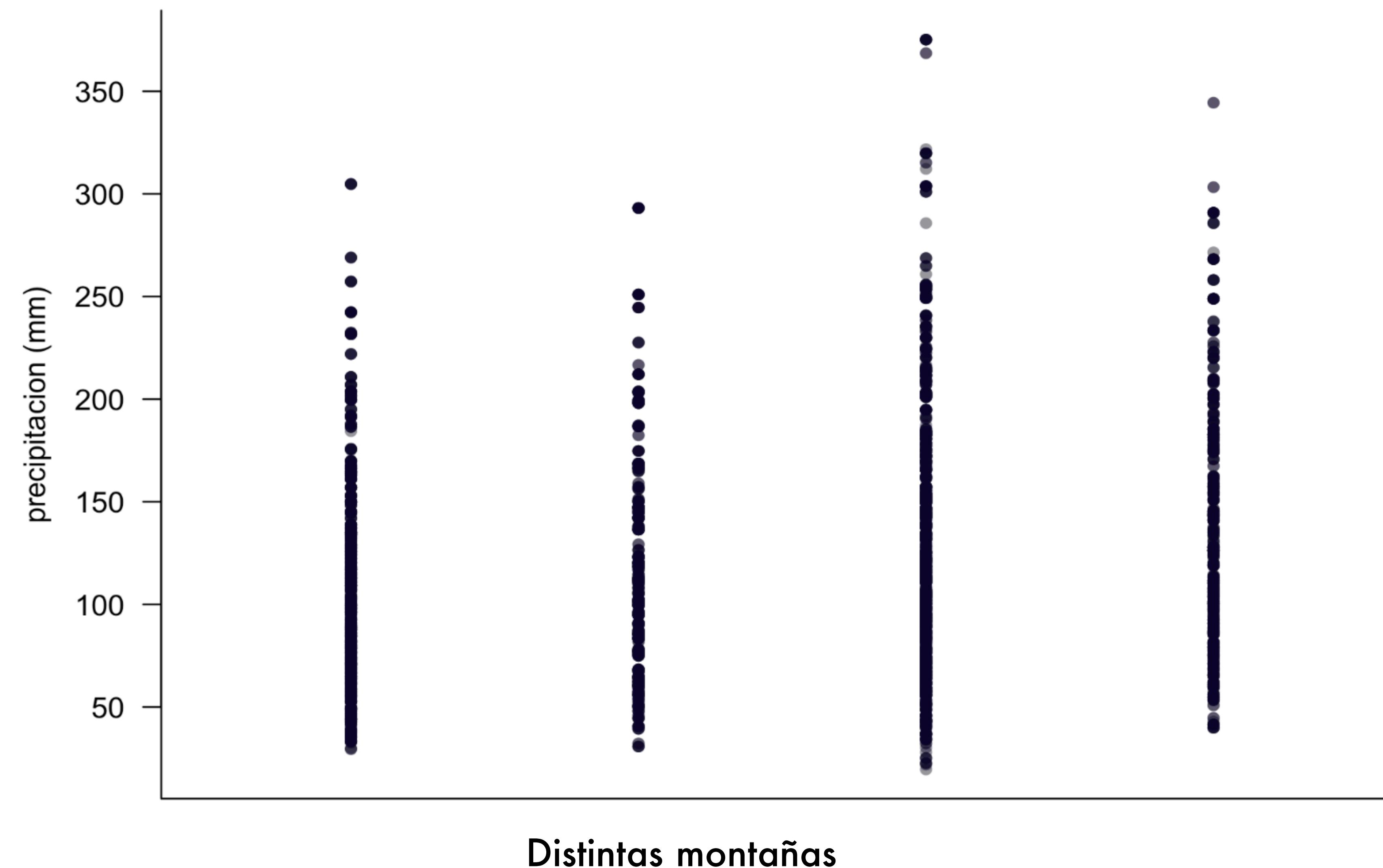
---

**practica en R: *script-part1.R***

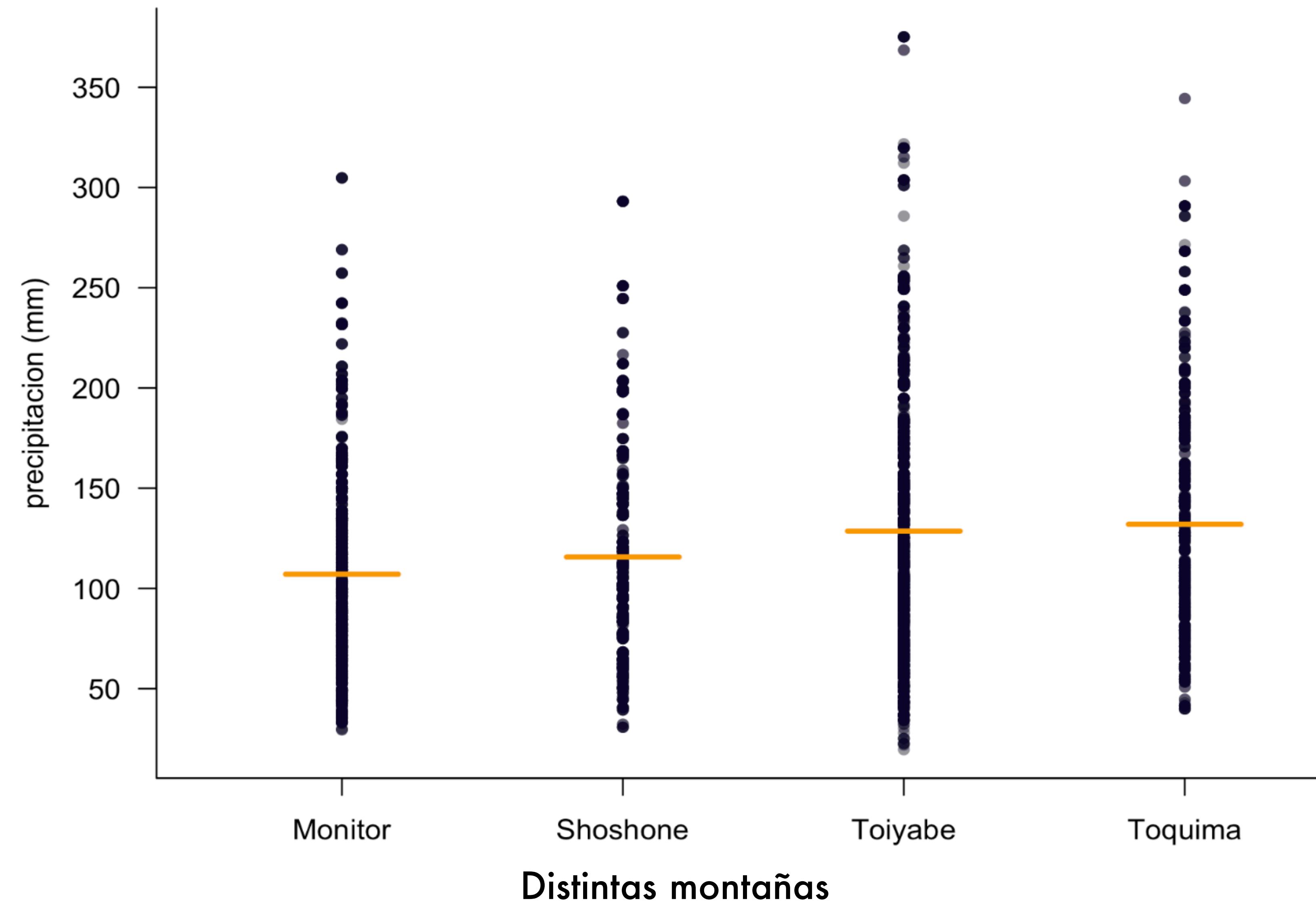
# **III. Modelos lineales, parte 2**

---

# ANOVA



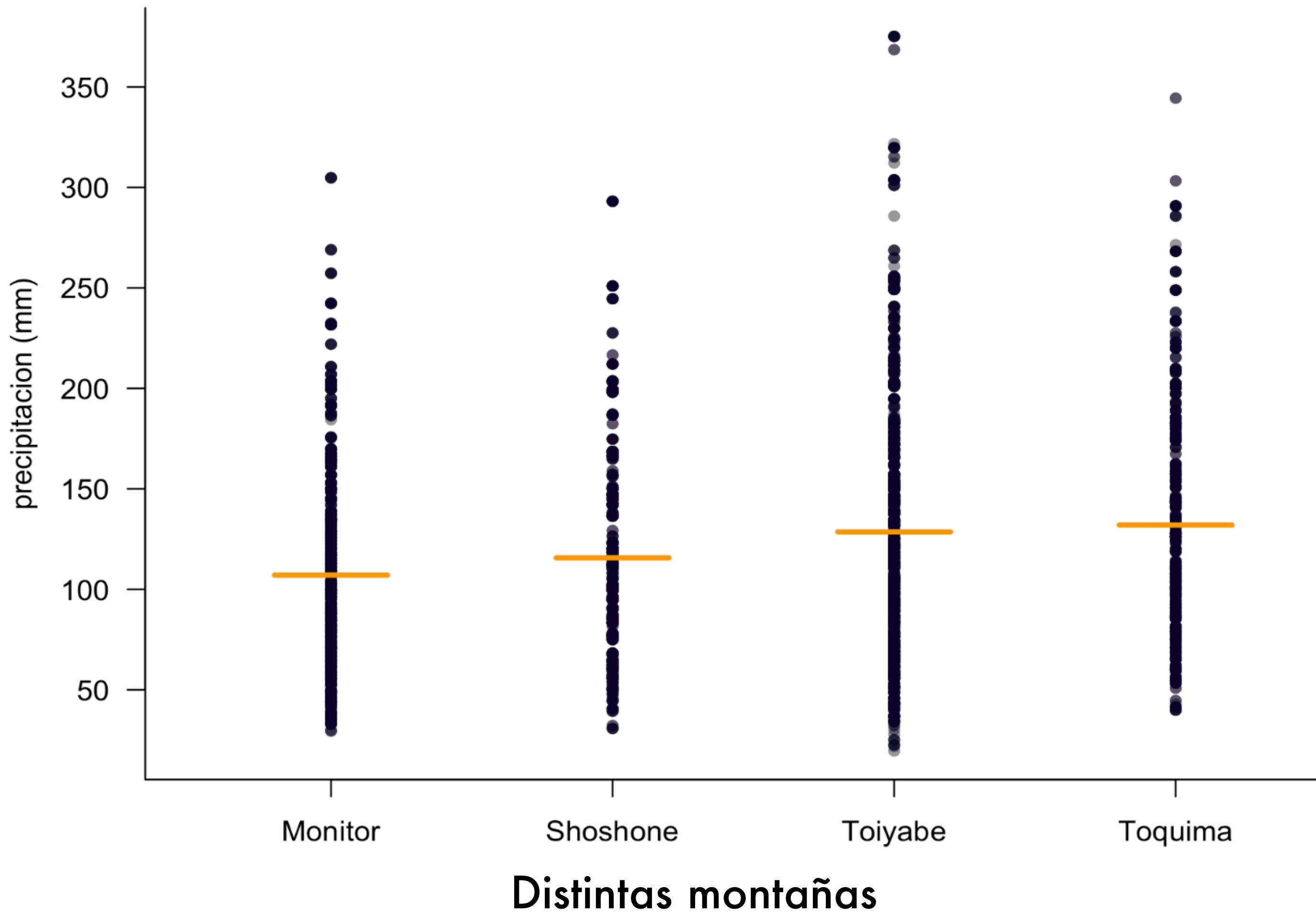
# ANOVA



Qué caracteriza esta  
ejemplo?

variable respuesta continua

variable predictoria  
categorica



## Supuestos: idénticos a los de regresiones lineales

En este caso  $x$  es una categoría, entonces se puede entender como el intercepto que cambia dado al variable.

En otras palabras, diferencias en la media entre distintas categorías

$$y_i = \alpha + \beta * x_i + \epsilon_i$$

$$\epsilon_i \sim N(0, \sigma^2)$$

**t-test: un ANOVA caso especial, con solo dos niveles**

# Cómo modelamos una variable categórica con multiples niveles?

montaña

a

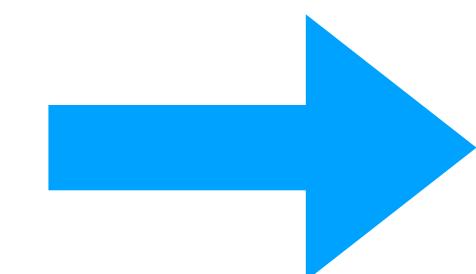
b

c

# Cómo modelamos una variable categórica con multiples niveles?

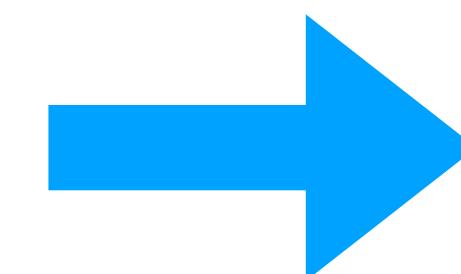
Tenemos que crear variables ficticias.

montaña	montaña-a	montaña-b	montaña-c
a	1	0	0
b	0	1	0
c	0	0	1



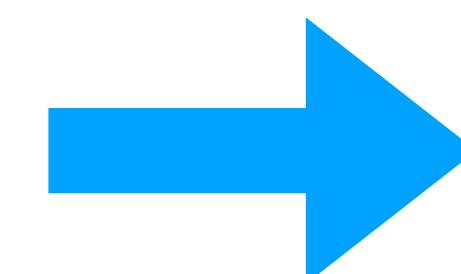
El primer nivel se convierte en la categoría por defecto.

montaña	intercepto	montaña-b	montaña-c
a	1	0	0
b	1	1	0
c	1	0	1



Las otras categorías están modeladas en comparación con la por defecto.

montaña	intercepto	montaña-b	montaña-c
a	1	0	0
b	1	1	0
c	1	0	1



Las otras categorías están modelado en comparación con el defecto.

montaña	intercepto	montaña-b	montaña-c
a	$\beta_a * 1$	$\beta_b \cancel{*} 0$	$\beta_c \cancel{*} 0$
b	1	1	0
c	1	0	1

Las otras categorías están modelado en comparación con el defecto.

montaña	intercepto	montaña-b	montaña-c
a	1	0	0
b	$\beta_a * 1$	$\beta_b * 1$	$\beta_c \cancel{*} 0$
c	1	0	1

Las otras categorías están modelado en comparación con el defecto.

montaña	intercepto	montaña-b	montaña-c
a	1	0	0
b	1	1	0
c	$\beta_a * 1$	$\beta_b * \cancel{0}$	$\beta_c * 1$

## Multiples variables predictoras

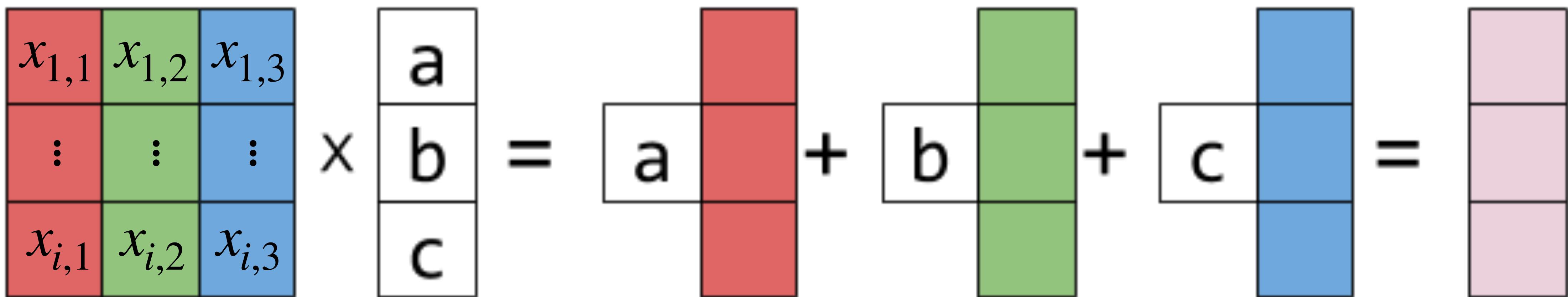
Con mas variables:

$$y_i = \beta_0 intercepto + \beta_1 cont1_i + \beta_2 cont2_i + \beta_3 cat_i + \epsilon_i$$

```
model.matrix(~ predictor1 + predictor2 + predictor3)
```

```
##          (Intercept) predictor1 predictor2 predictor3
## [1,]           1     2285.1      78.99     20.8
## [2,]           1     2304.6      67.31     17.1
## [3,]           1     2330.1      64.27     16.7
## [4,]           1     2589.2     144.02     15.8
## [5,]           1     3016.6     235.46      6.2
```

## Multiples variables predictoras



$$\beta^T x_i = \beta_1 x_{i,1} + \beta_2 x_{i,2} + \beta_3 x_{i,3}$$

## Multiples variables predictoras

Se pueden integrar variables predictoras continuas y categóricas en el mismo modelo:

```
mod <- lm(response ~ continuous1 + continuous2 + discrete)
```

## Multiples variables predictoras

Se pueden integrar variables predictoras continuas y categóricas en el mismo modelo:

```
mod <- lm(response ~ continuous1 + continuous2 + discrete)
```

```
##          (Intercept) continuous1 continuous2 discrete1 discrete2 discrete3
## [1,]           2285.1       78.99         0         0         0
## [2,]           2304.6       67.31         1         0         0
## [3,]           2330.1       64.27         0         0         0
## [4,]           2589.2      144.02         0         0         1
## [5,]           3016.6      235.46         0         1         0
```

## Multiples variables predictoras

---

Otra vez los supuestos son los mismos tres, mas un cuarto supuesto: las distintas variables predictoras son relativamente independientes entre sí.

Si no lo son, lo llamamos "multi-colinearidad".

Es difícil para el modelo de distinguir entre los efectos de una variable predictora y de otra. También llamado problema de distinguibilidad de los efectos. En simple, el modelo se confunde...

## Multiples variables predictoras

---

Otra vez los supuestos son los mismos tres, mas un cuarto supuesto: las distintas variables predictoras son relativamente independientes entre sí.

Si no lo son, lo llamamos "multi-colinearidad".

Es difícil para el modelo de distinguir entre los efectos de una variable predictora y de otra. También llamado problema de distinguibilidad de los efectos. En simple, el modelo se confunde...

Debemos evitar romper este supuesto eligiendo cuidadosamente las variables, o incluso sacando variables fuertemente correlacionadas.

## **Multiples variables predictoras**

---

Si nuestra pregunta de investigación requiere la comparación de los tamaños de los efectos de distintas variables predictoras [modelos explicativos], entonces es necesario pensar en la escala de los variables.

## Multiples variables predictoras

Si nuestra pregunta de investigación requiere la comparación de los tamaños de los efectos de distintas variables predictoras [modelos explicativos], entonces es necesario pensar en la escala de los variables.

Para hacerlas comparables, podemos estandarizar los valores de las variables predictoras.

```
# standardise continuous predictors  
predictors_std <- scale(predictors)
```

```
##      predictor1 predictor2 predictor3  
## [1,] -0.7065051 -0.5328227  1.00678496  
## [2,] -0.6438887 -0.6923145  0.32702139  
## [3,] -0.5620058 -0.7338261  0.25353344  
## [4,]  0.2699889  0.3551696  0.08818554  
## [5,]  1.6424108  1.6037937 -1.67552534
```

## **Multiples variables predictoras**

---

**practica en R: *script-part2.R***

## **Interacciones**

---

### **Interacciones variables predictoras**

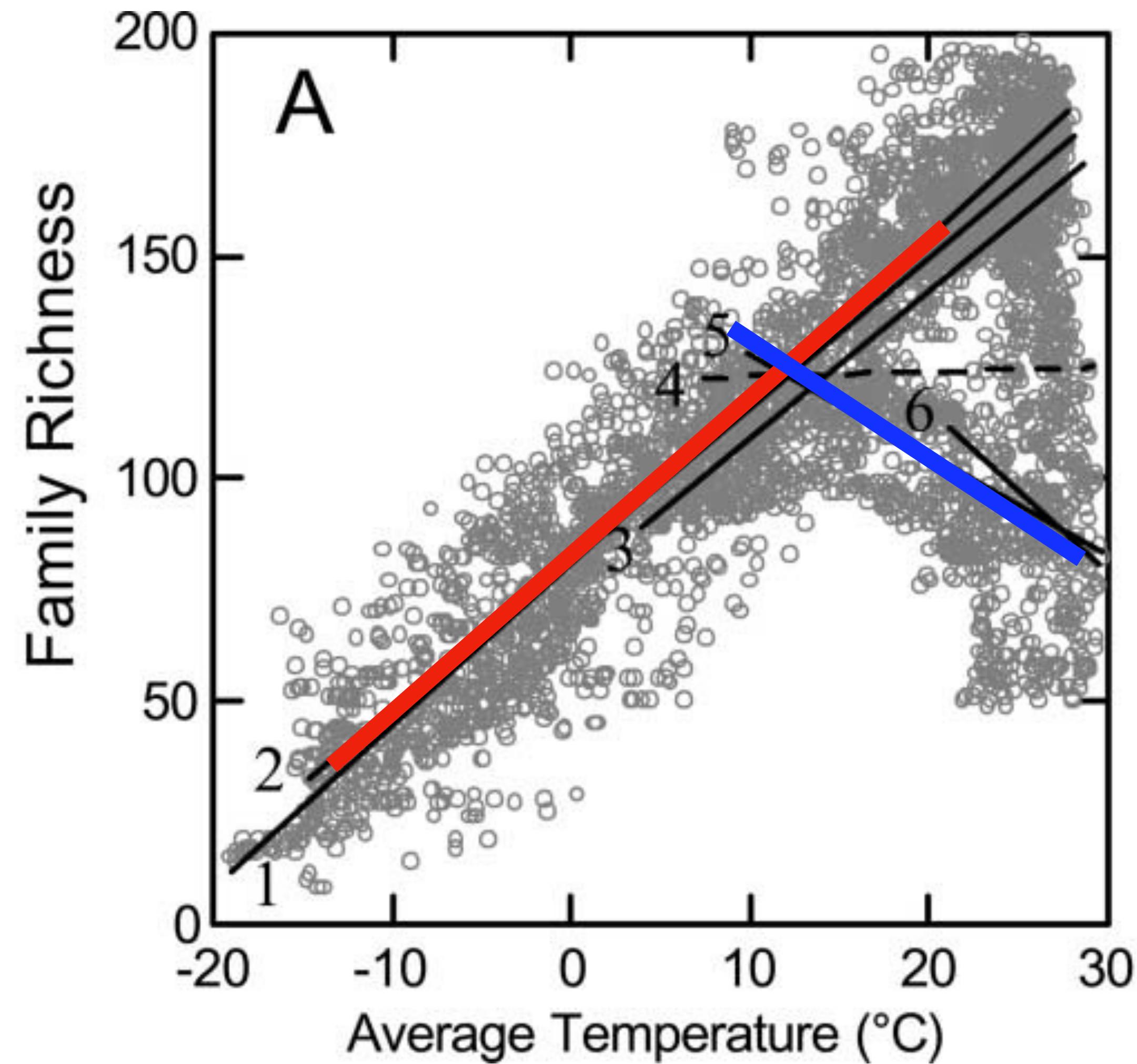
## Interacciones

### Interacciones variables predictoras

VOL. 161, NO. 4 THE AMERICAN NATURALIST APRIL 2003

#### A Globally Consistent Richness-Climate Relationship for Angiosperms

Anthony P. Francis and David J. Currie\*



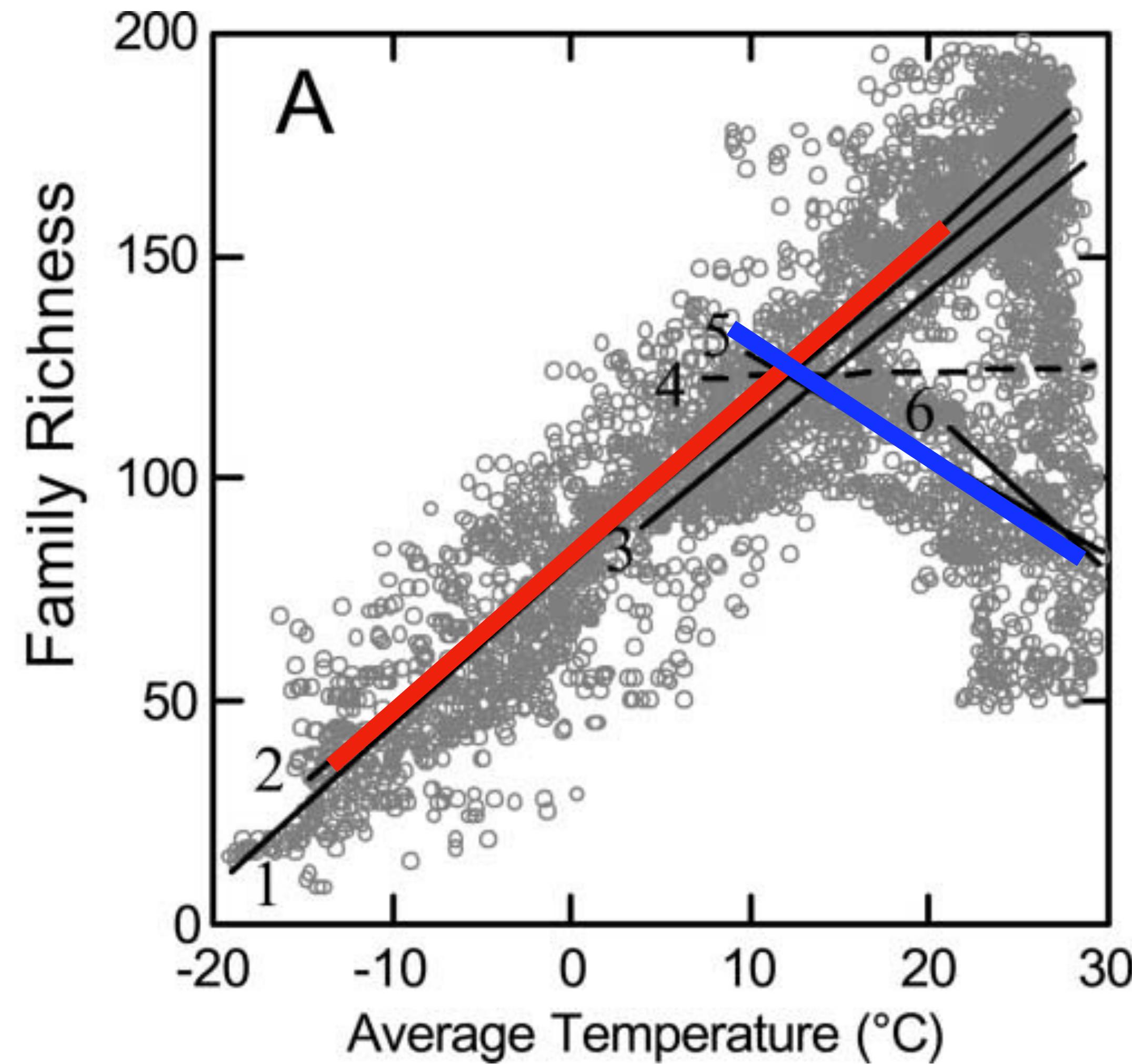
# Interacciones variables predictoras

VOL. 161, NO. 4 THE AMERICAN NATURALIST APRIL 2003

### A Globally Consistent Richness-Climate Relationship for Angiosperms

Anthony P. Francis and David J. Currie\*

- Riqueza de especies aumenta con la temperatura, en climas húmedos



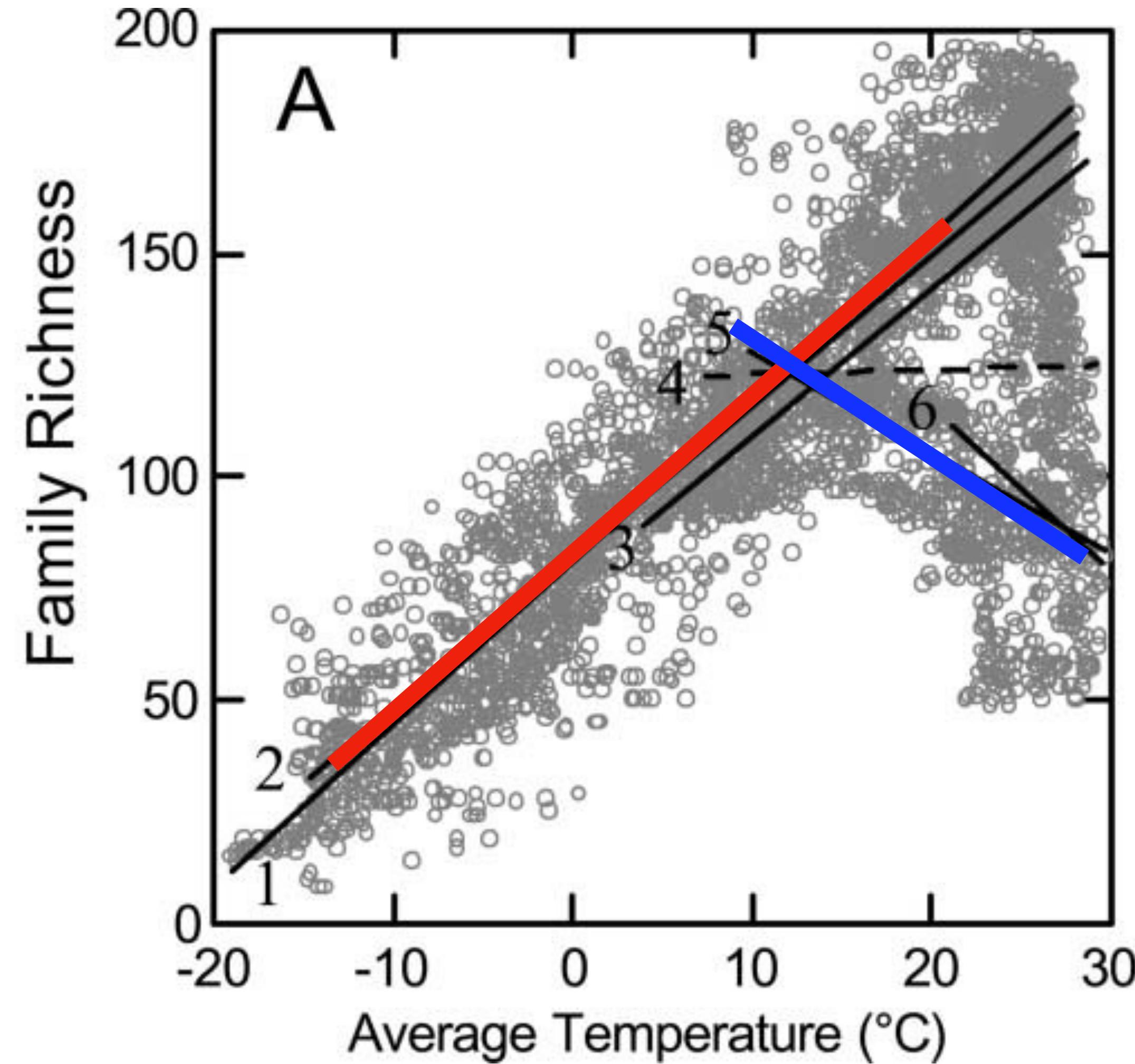
# Interacciones variables predictoras

VOL. 161, NO. 4 THE AMERICAN NATURALIST APRIL 2003

### A Globally Consistent Richness-Climate Relationship for Angiosperms

Anthony P. Francis and David J. Currie\*

- Riqueza de especies aumenta con la temperatura, en climas húmedos
- Pero decrece con la temperatura, en climas secos



## Interacciones

$$prec_i = \beta_0 intercepto + \beta_1 elev_i + \beta_2 montana_i + \beta_3 elev_i * montana_i + \epsilon_i$$

$$\epsilon_i \sim N(0, \sigma^2)$$

```
##  
## Call:  
## lm(formula = precipitation ~ elevation * mountain_range)  
##  
## Residuals:  
##      Min       1Q   Median      3Q      Max  
## -120.75  -35.97   -9.48   30.03  202.73  
##  
## Coefficients:  
##                               Estimate Std. Error t value Pr(>|t|)  
## (Intercept)                 -2.104e+02  3.185e+01 -6.607  4.4e-11  
## elevation                  1.364e-01  1.366e-02  9.981 < 2e-16  
## mountain_rangeShoshone     1.842e+01  4.067e+01  0.453  0.65053  
## mountain_rangeToiyabe      9.473e+01  3.346e+01  2.832  0.00465  
## mountain_rangeToquima     1.798e+01  4.192e+01  0.429  0.66795  
## elevation:mountain_rangeShoshone -3.521e-03  1.747e-02 -0.202  0.84028  
## elevation:mountain_rangeToiyabe -3.089e-02  1.435e-02 -2.152  0.03145  
## elevation:mountain_rangeToquima -2.771e-03  1.767e-02 -0.157  0.87538  
##
```

## Interacciones

- Es difícil interpretar coeficientes

```
mod <- lm(precipitation ~ elevation * mountain_range)
summary(mod)
```

```
##
## Call:
## lm(formula = precipitation ~ elevation * mountain_range)
##
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -120.75  -35.97   -9.48   30.03  202.73 
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)    
## (Intercept)                 -2.104e+02  3.185e+01 -6.607  4.4e-11  
## elevation                      1.364e-01  1.366e-02  9.981  < 2e-16  
## mountain_rangeShoshone        1.842e+01  4.067e+01  0.453  0.65053  
## mountain_rangeToiyabe         9.473e+01  3.346e+01  2.832  0.00465  
## mountain_rangeToquima        1.798e+01  4.192e+01  0.429  0.66795  
## elevation:mountain_rangeShoshone -3.521e-03  1.747e-02 -0.202  0.84028  
## elevation:mountain_rangeToiyabe -3.089e-02  1.435e-02 -2.152  0.03145  
## elevation:mountain_rangeToquima -2.771e-03  1.767e-02 -0.157  0.87538 
##
```

## Interacciones

- Es difícil interpretar coeficientes
- El efecto de cada parámetro depende del valor estimado de otros parámetros (al igual que en variables categóricas)

```
mod <- lm(precipitation ~ elevation * mountain_range)
summary(mod)

##
## Call:
## lm(formula = precipitation ~ elevation * mountain_range)
##
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -120.75  -35.97   -9.48   30.03  202.73 
##
## Coefficients:
## (Intercept)          elevation mountain_rangeShoshone
##             -2.104e+02            1.364e-01            1.842e+01
##             1.842e+01            9.473e+01            1.798e+01
##             9.473e+01           -3.521e-03           -3.089e-02
##             1.798e+01           -2.771e-03           1.767e-02
##             1.767e-02
##             4.067e+01            3.346e+01            4.192e+01
##             3.346e+01            2.832e+01            4.192e+01
##             0.453e+01            0.429e+01            0.453e+01
##             0.65053e+00            0.66795e+00            0.429e+01
##             0.00465e+00            0.84028e+00            0.429e+01
##             0.03145e+00            0.87538e+00            0.429e+01
##             0.87538e+00
##             4.4e-11               < 2e-16            0.65053e+00
##             0.00465e+00            0.66795e+00            0.429e+01
##             0.03145e+00            0.87538e+00            0.429e+01
##             0.87538e+00
```

## Interacciones

- Es difícil interpretar coeficientes
- El efecto de cada parámetro depende del valor estimado de otros parámetros (al igual que en variables categóricas)
- Particularmente difícil si la interacción es entre dos variables continuas

```
mod <- lm(precipitation ~ elevation * mountain_range)
summary(mod)

##
## Call:
## lm(formula = precipitation ~ elevation * mountain_range)
##
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -120.75  -35.97   -9.48   30.03  202.73 
##
## Coefficients:
## (Intercept)          elevation
##             -2.104e+02           1.364e-01
##             mountain_rangeShoshone
##                         1.842e+01
##             mountain_rangeToiyabe
##                         9.473e+01
##             mountain_rangeToquima
##                         1.798e+01
##             elevation:mountain_rangeShoshone
##                         -3.521e-03
##             elevation:mountain_rangeToiyabe
##                         -3.089e-02
##             elevation:mountain_rangeToquima
##                         -2.771e-03
## 
```

El manera de interpretar/ilustrar las estimaciones de un modelo depende de la pregunta.

El manera de interpretar/ilustrar las estimaciones de un modelo depende de la pregunta.

Qué tan bien se ajusta el modelo a los datos?

usa R<sup>2</sup>

## Modelos lineales

---

El manera de interpretar/ilustrar las estimaciones de un modelo depende de la pregunta.

Qué tan bien se ajusta el modelo a los datos?

usa R<sup>2</sup>

Hay apoyo estadística para una asociación?

usa p-values

## Modelos lineales

---

El manera de interpretar/ilustrar las estimaciones de un modelo depende de la pregunta.

Qué tan bien se ajusta el modelo a los datos?

usa R<sup>2</sup>

Hay apoyo estadística para una asociación?

usa p-values

Una asociación estadísticamente significativa tiene sentido?

mira los coeficientes  
(tamaño y signo)

Entrenamos modelo

Predecir a datos de ajuste == para calcular bondad de ajuste

Predecir a datos externos == para calcular capacidad predictiva

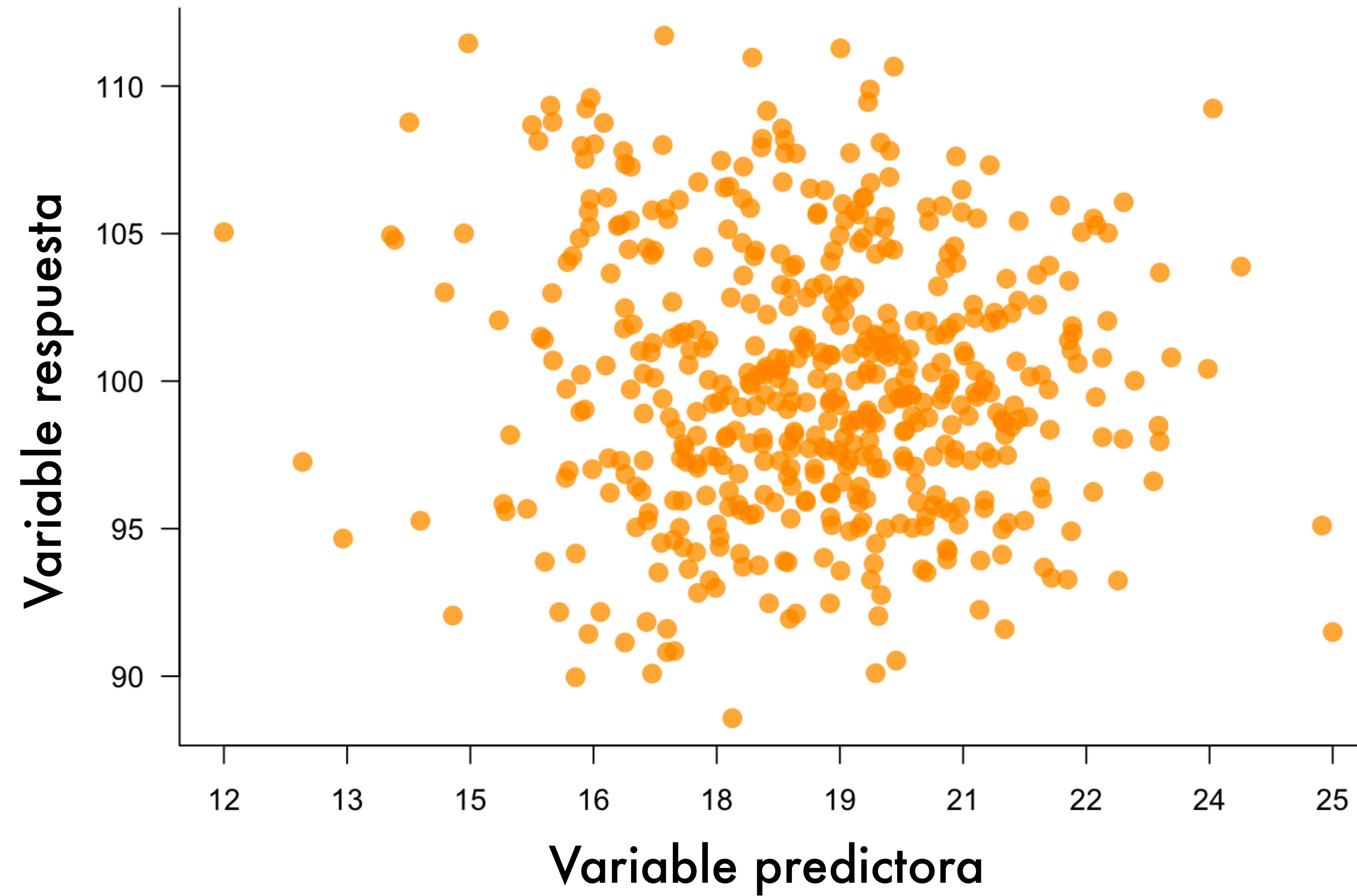
>> puede ser datos externos dentro del limite de datos de ajuste (== interpolacion)  
o fuera de rango de datos de ajuste (== extrapolacion)

practica en R: *script-part2.R*

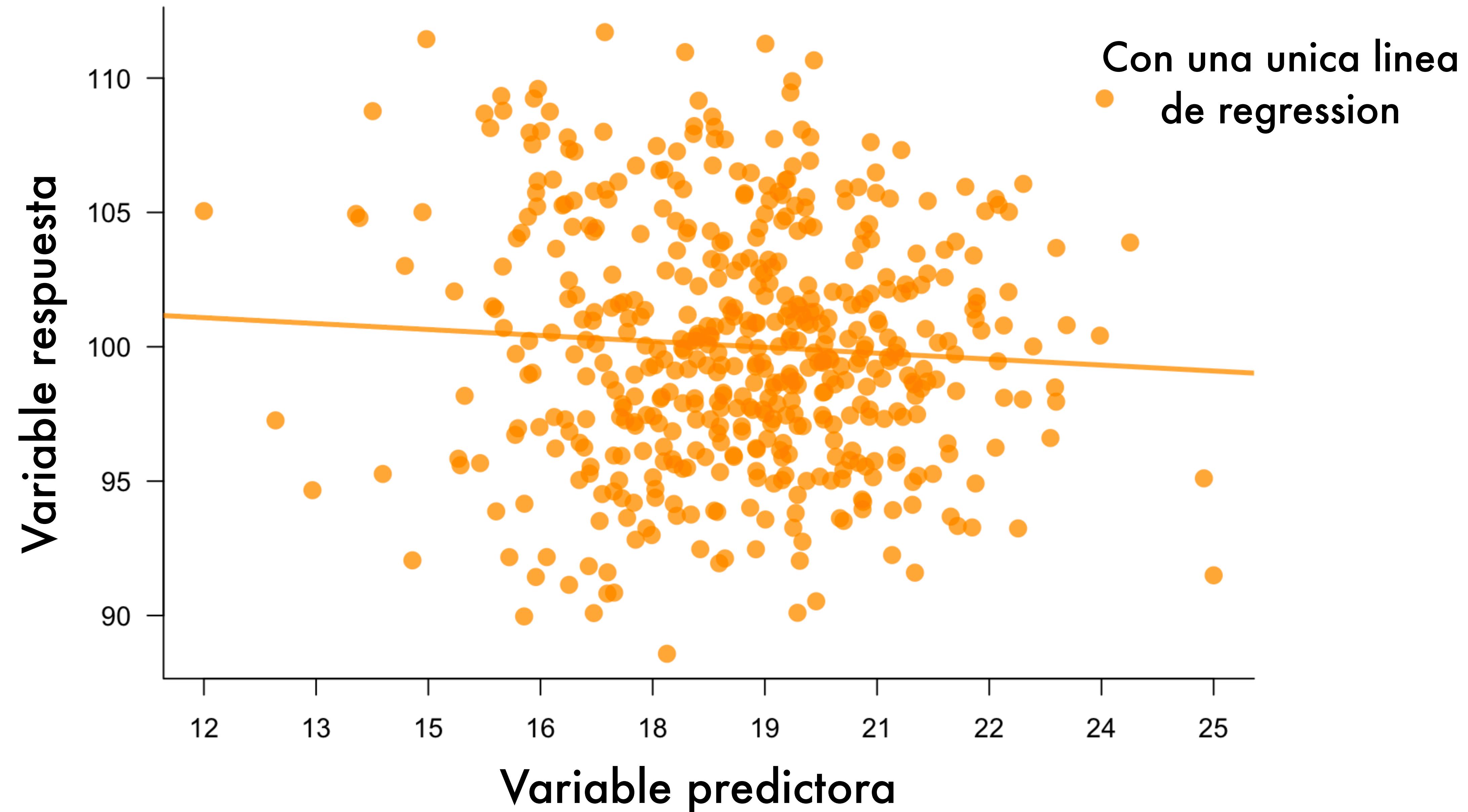
## **IV. Modelos mixtos**

---

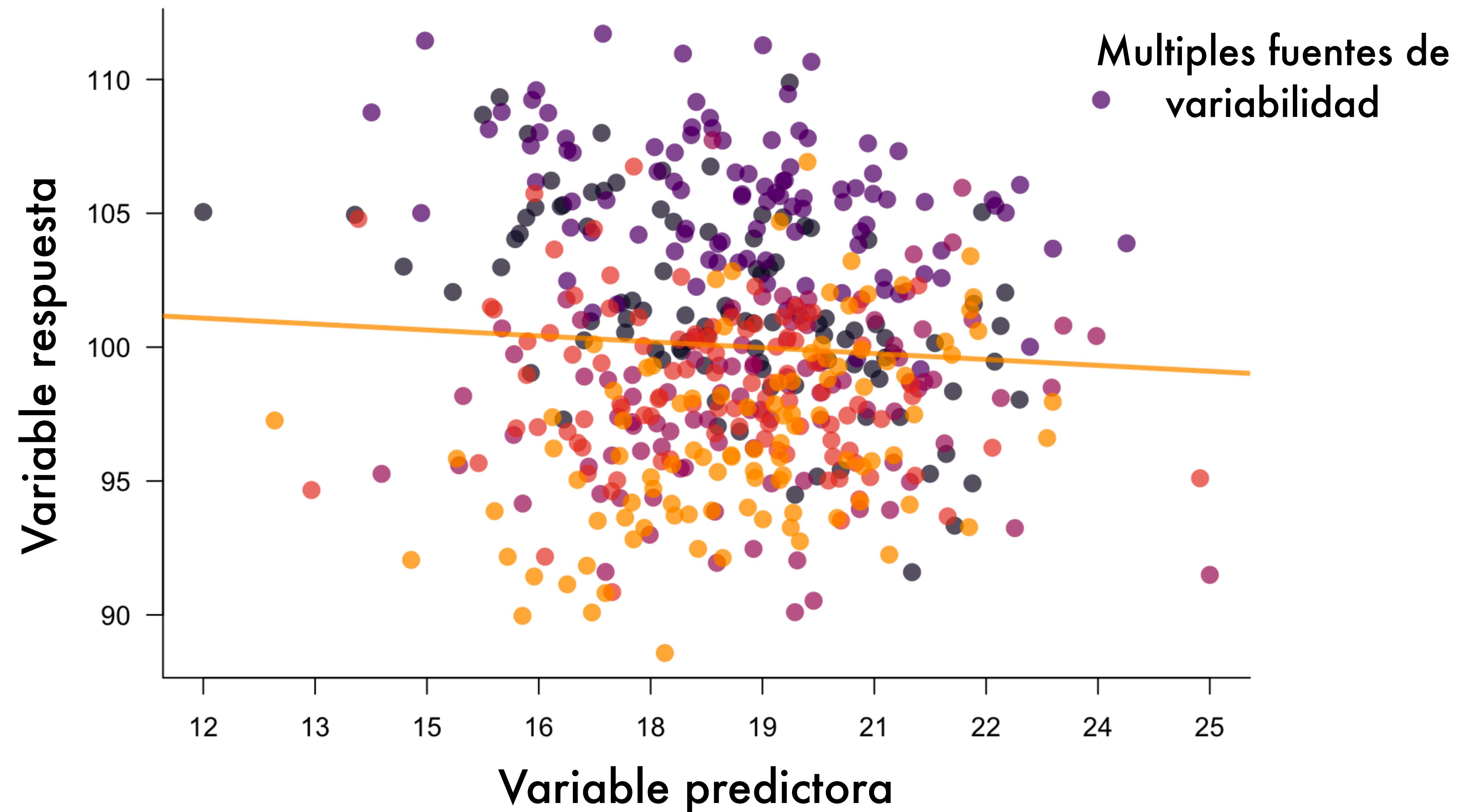
## Modelos mixtos



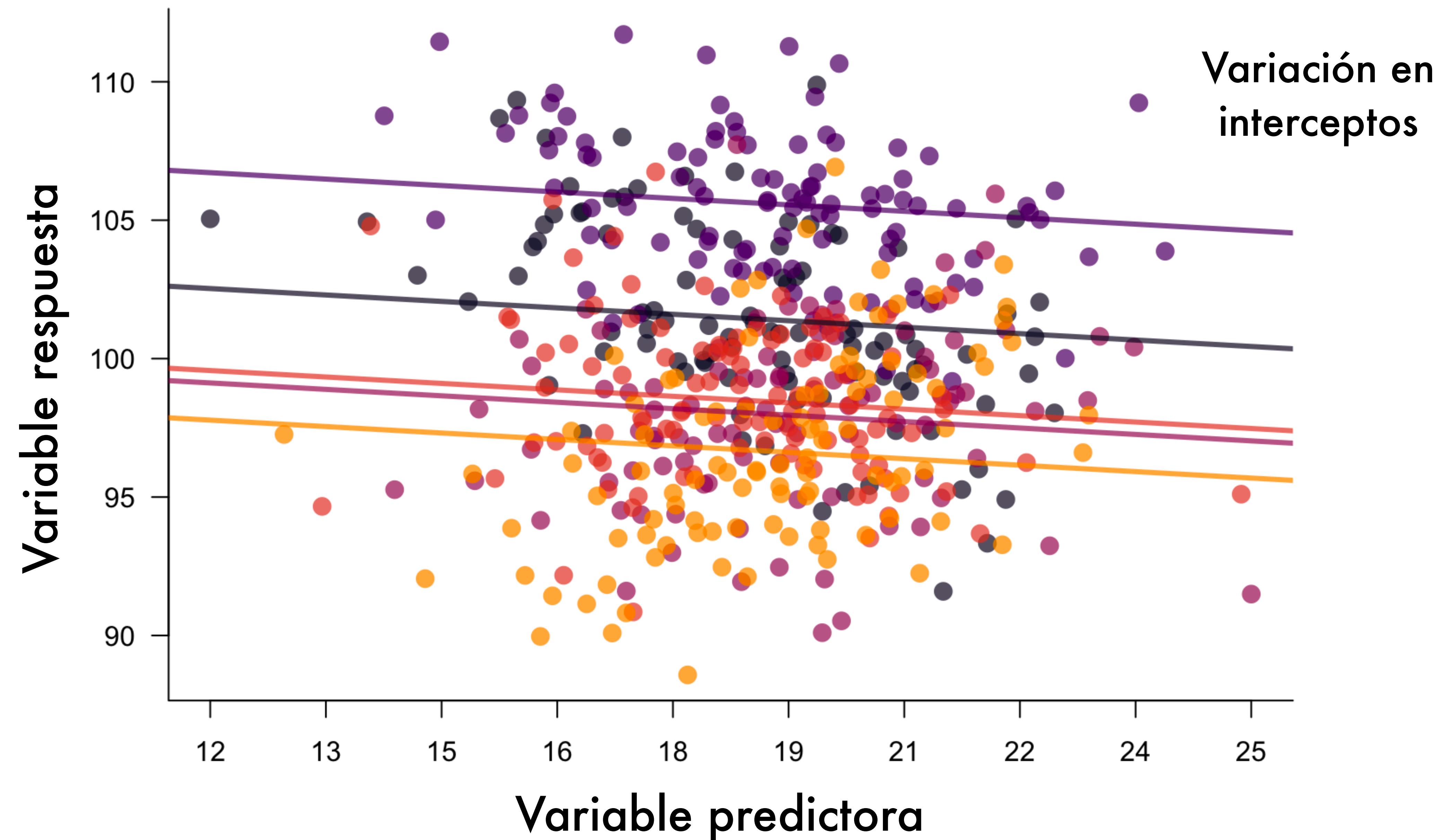
## Modelos mixtos



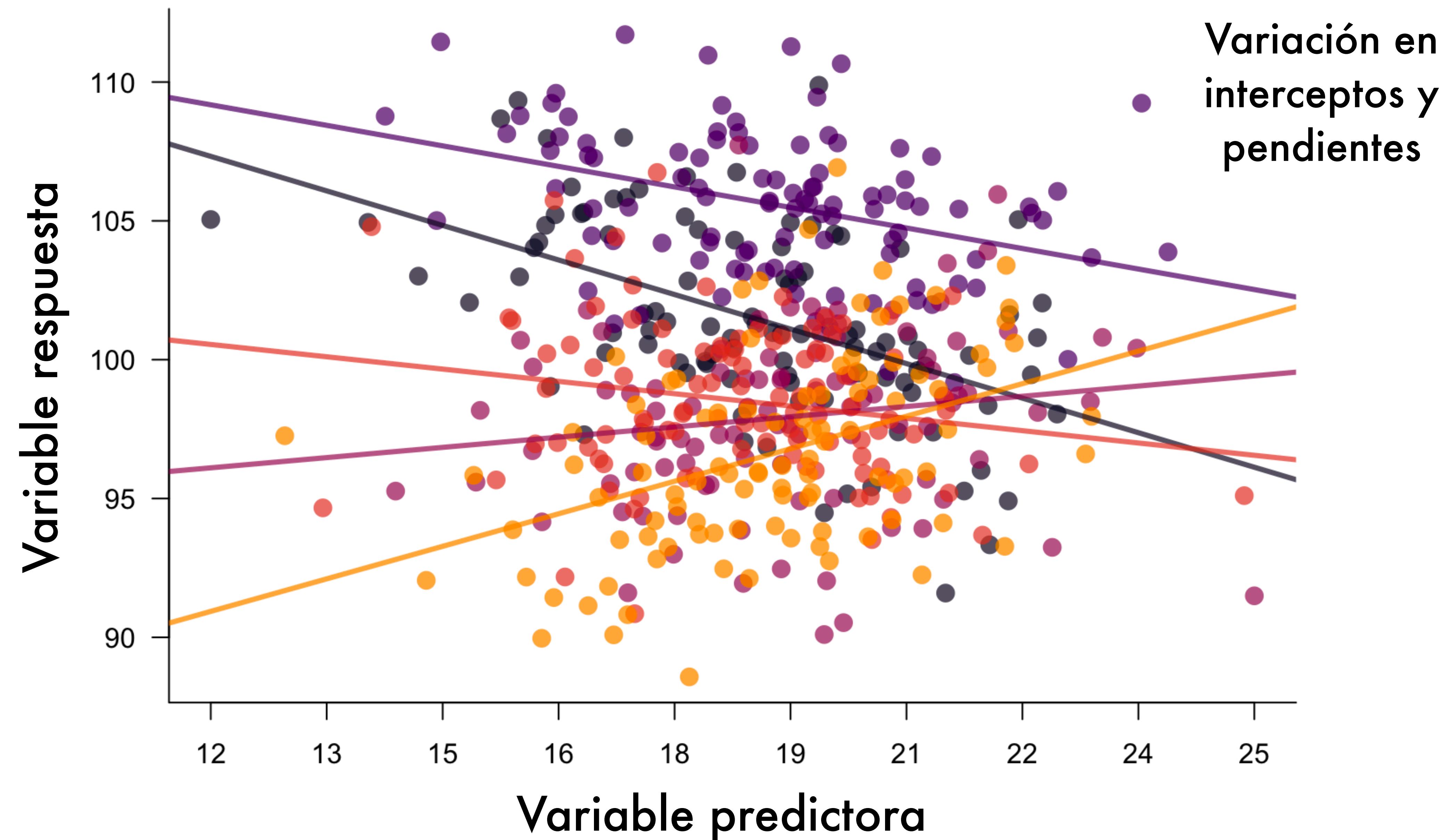
## Modelos mixtos



## Modelos mixtos

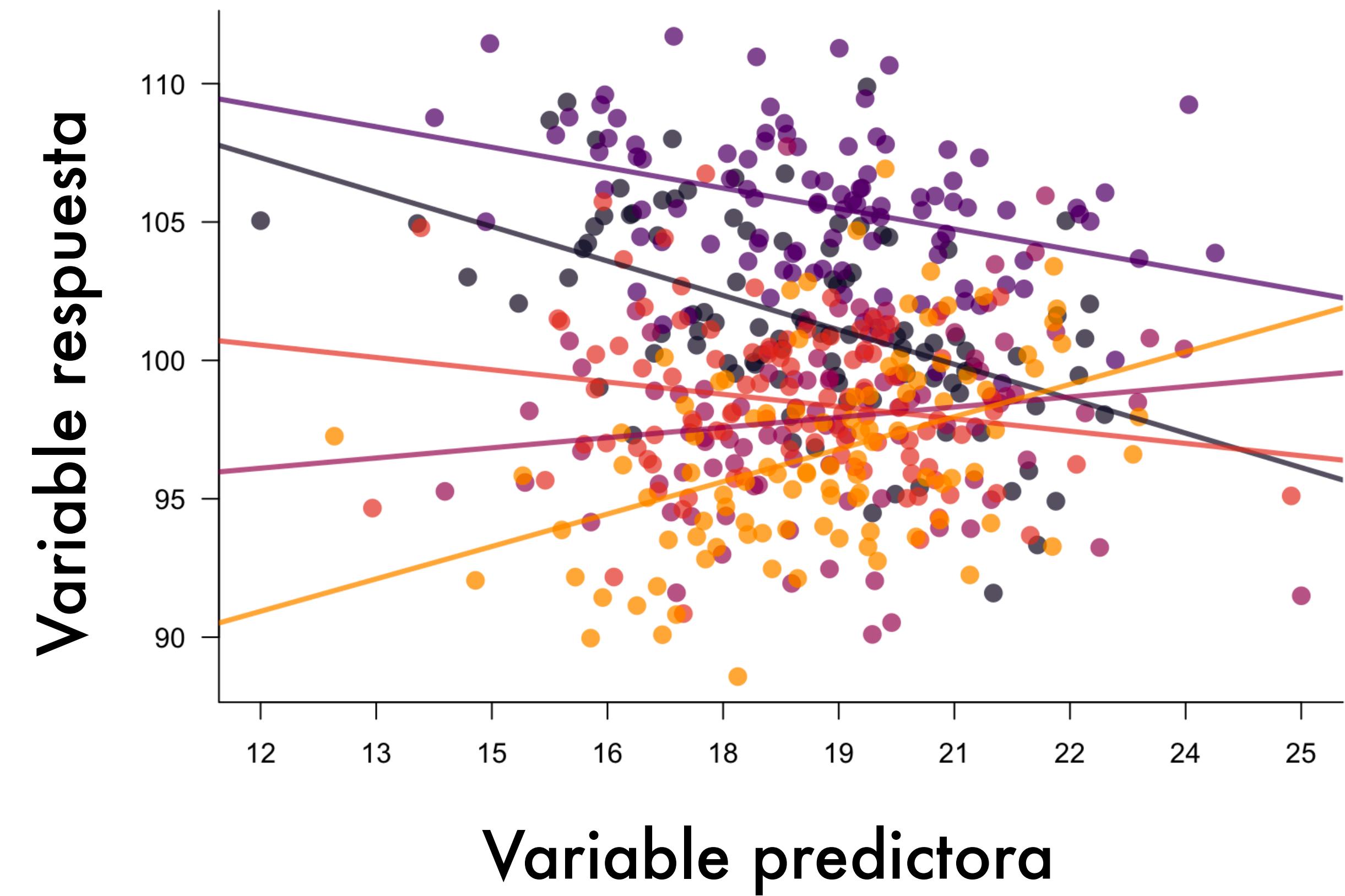


## Modelos mixtos



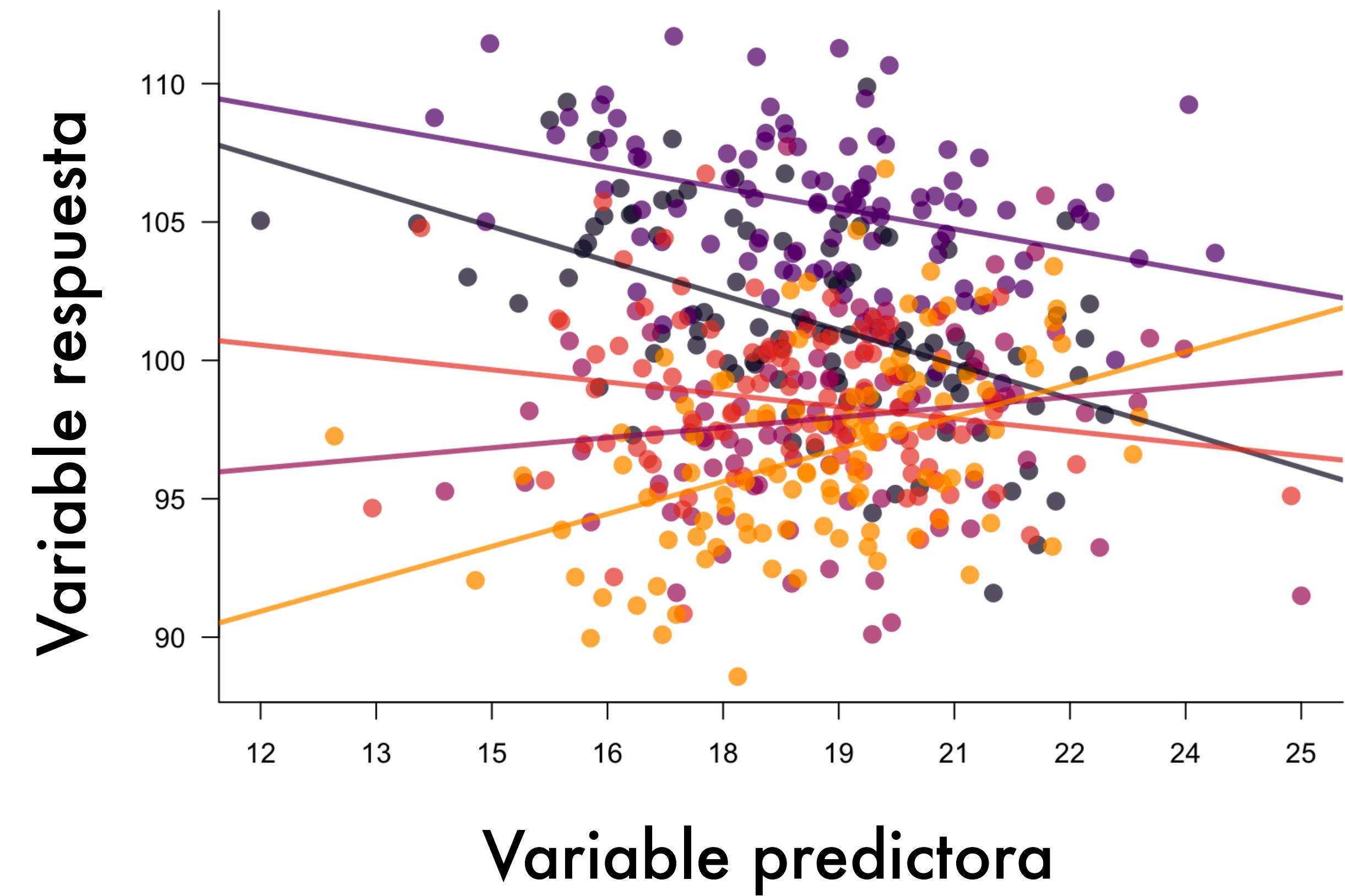
## Modelos mixtos

- Como vimos con variables categóricas, grupos pueden variar en su relación con la variable respuesta.



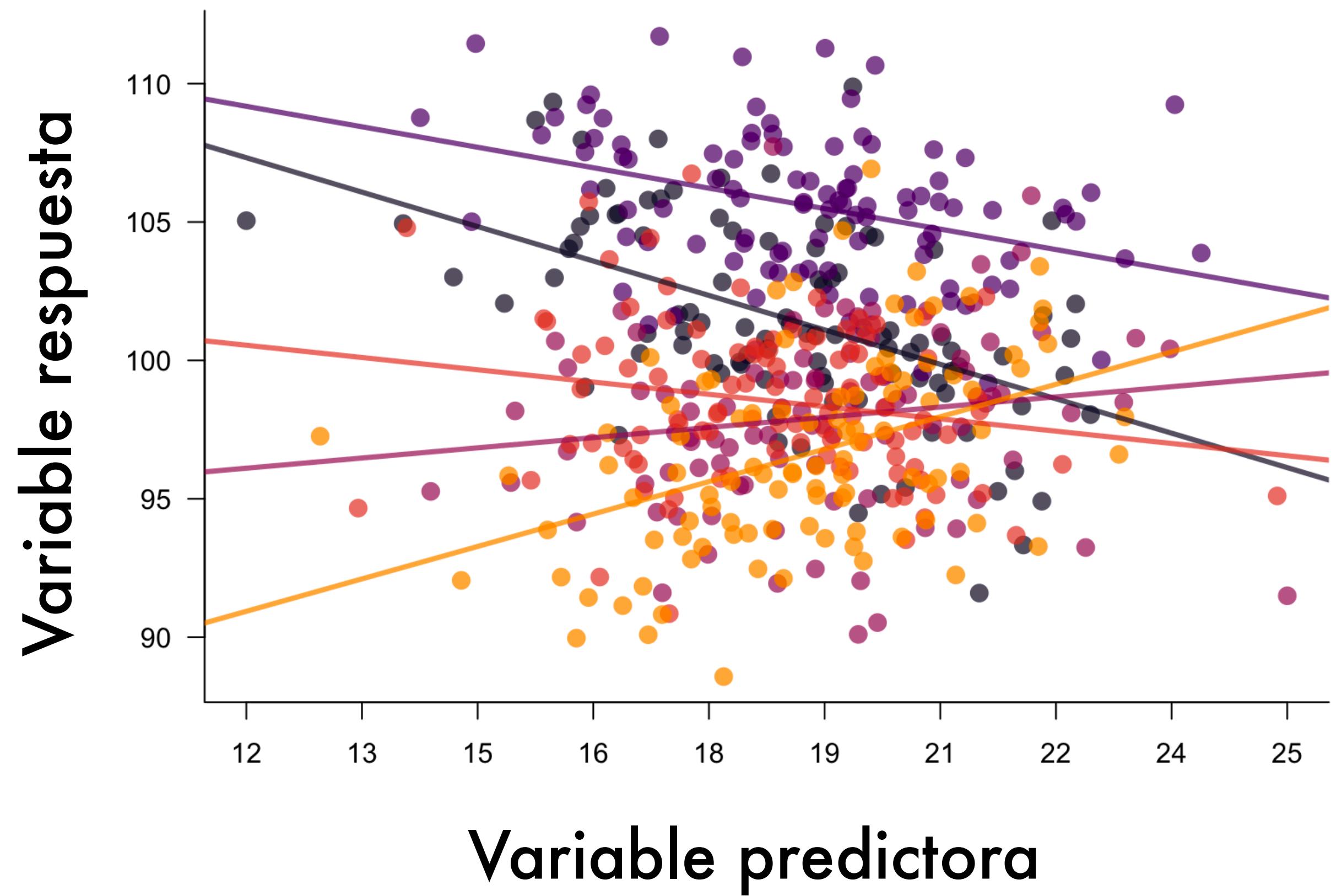
## Modelos mixtos

- Como vimos con variables categóricas, grupos pueden variar en su relación con la variable respuesta.
- Con variables predictoras categóricas, los grupos varian en su relación con la variable respuesta (distintas pendientes para cada grupo, es decir, efectos interactivos) en una manera que *nos interesa (relación directa con nuestra pregunta)* y que sean independientes.



## Modelos mixtos

- Cuando hay variación entre grupos o categorías que **NO** son de interés para nuestra hipótesis pero que sabemos existe, podemos reportarla en el modelo a través de efectos aleatorios.
- No reportarla puede ser problemático porque puede introducir ruido al modelo o romper el supuesto de independencia.



## Modelos mixtos



## Modelos mixtos



## Modelos mixtos

---

- Normalmente no interpretamos los efectos aleatorios. Si quieres interpretarlos, puede que sea mejor usarlo como efecto fijo.
- Recomendamos **decidir sus efectos aleatorios a priori porque dependen del diseño experimental y no de la hipótesis.**
- Aunque normalmente no interpretamos los valores de los efectos aleatorios, puede ser útil examinar la cantidad de variabilidad explicada por el modelo considerando o no los efectos aleatorios.

**Modelo conceptual**

**Formar la pregunta / escribir el modelo**

**Diseño experimental**

**Colección de datos**

**Armar el modelo**

**Creer resultados**

**Modelo conceptual**

**Formar la pregunta**

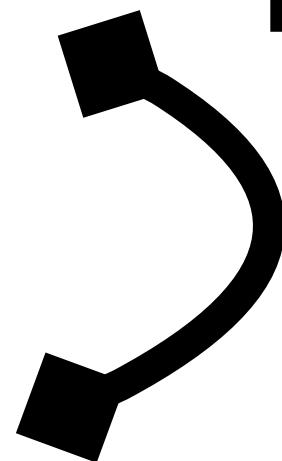
**Diseño experimental**

**Colección de datos**

**Armar el modelo**

**Creer resultados**

**Escribir el modelo**



## **Modelos mixtos**

---

En R usando paquete *lme4*

## Modelos mixtos

---

En R usando paquete *lme4*

$$y_i = \alpha + \beta * x_i$$

```
# Cargar paquete lme4
library(lme4)

# Ajustar modelo con intercepto unico y un predictor
mod_lm <- lm(response ~ predictor)
```

## Modelos mixtos

En R usando paquete *lme4*

$$y_i = \alpha_{j[i]} + \beta * x_i$$

```
# Cargar paquete lme4
library(lme4)
```

```
# Ajustar modelo con intercepto unico y un predictor
mod_lm <- lm(response ~ predictor)
```

```
# Ajustar modelo con interceptos aleatorios
mod_int <- lmer(response ~ predictor + (1 | block))
```

## Modelos mixtos

En R usando paquete *lme4*

$$y_i = \alpha_{j[i]} + \beta_{j[i]} * x_i$$

```
# Cargar paquete lme4
library(lme4)

# Ajustar modelo con intercepto unico y un predictor
mod_lm <- lm(response ~ predictor) == lm(response ~ 1 + predictor)

# Ajustar modelo con interceptos aleatorios
mod_int <- lmer(response ~ predictor + (1 | block))

# Ajustar modelo con interceptos y pendientes aleatorios
mod_slope <- lmer(response ~ predictor + (1 + predictor | block))
== lmer(response ~ predictor + (predictor | block))
```

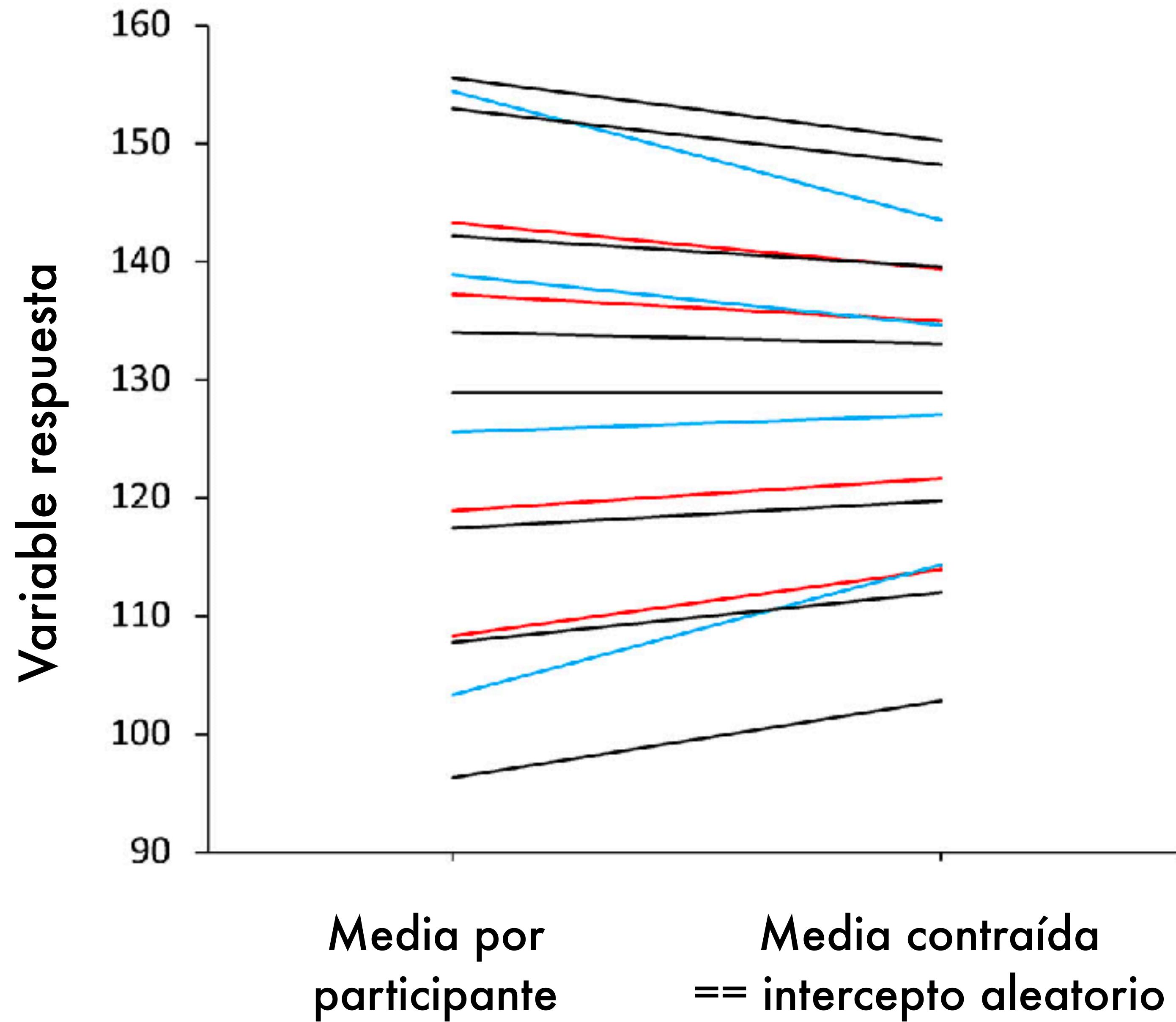
## **Modelos mixtos**

---

**practica en R: *script-part3.R***

## Modelos mixtos

Contracción del intercepto aleatorio para cada nivel, hacia la media del grupo



## Modelos mixtos

Recuerda: modelando un variable categorica

montaña	intercepto	montaña-b	montaña-c
a	1	0	0
b	1	1	0
c	$\beta_a * 1 + \beta_b * 0 + \beta_c * 1$	<del><math>\beta_b * 0</math></del>	

Recuerda: modelando un variable categorica

montaña	intercepto	montaña-b	montaña-c
a	1	0	0
b	1	1	0
c	$\beta_a * 1$	$\beta_b * 0$	$\beta_c * 1$

$\beta_a, \beta_b, \beta_c$       **son independientes**

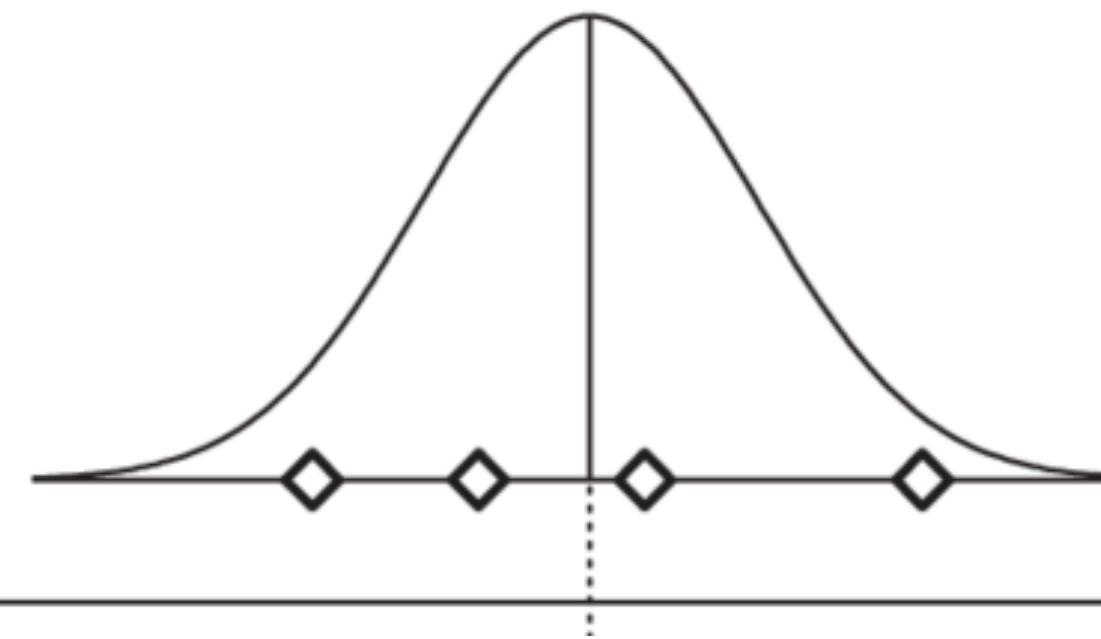
## Modelos mixtos

Un valor muestrado de una distribucion normal para el

parametro  $\theta_i \sim N(\mu, \sigma_i^2)$

### First Stage

A  $\theta_i$  value is randomly sampled from a normal distribution with expected value  $\mu$  and variance  $\tau^2_T$  for each study that is conducted.



## Modelos mixtos

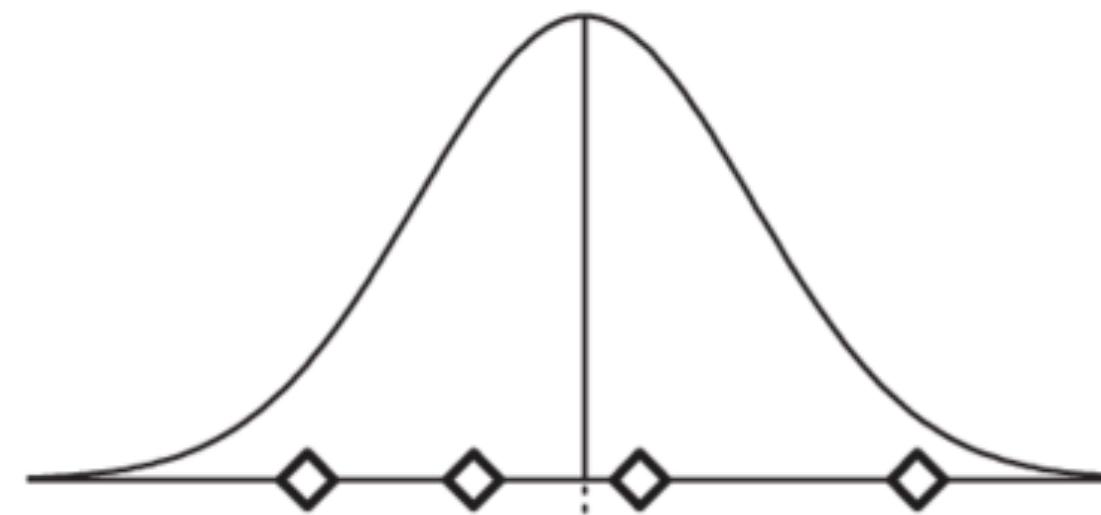
Un valor muestrado de una distribución normal para el parámetro

$$\theta_i \sim N(\mu, \sigma_i^2)$$

Condicional al  $\theta_i$ , una distribución de muestra existe para cada nivel aleatorio

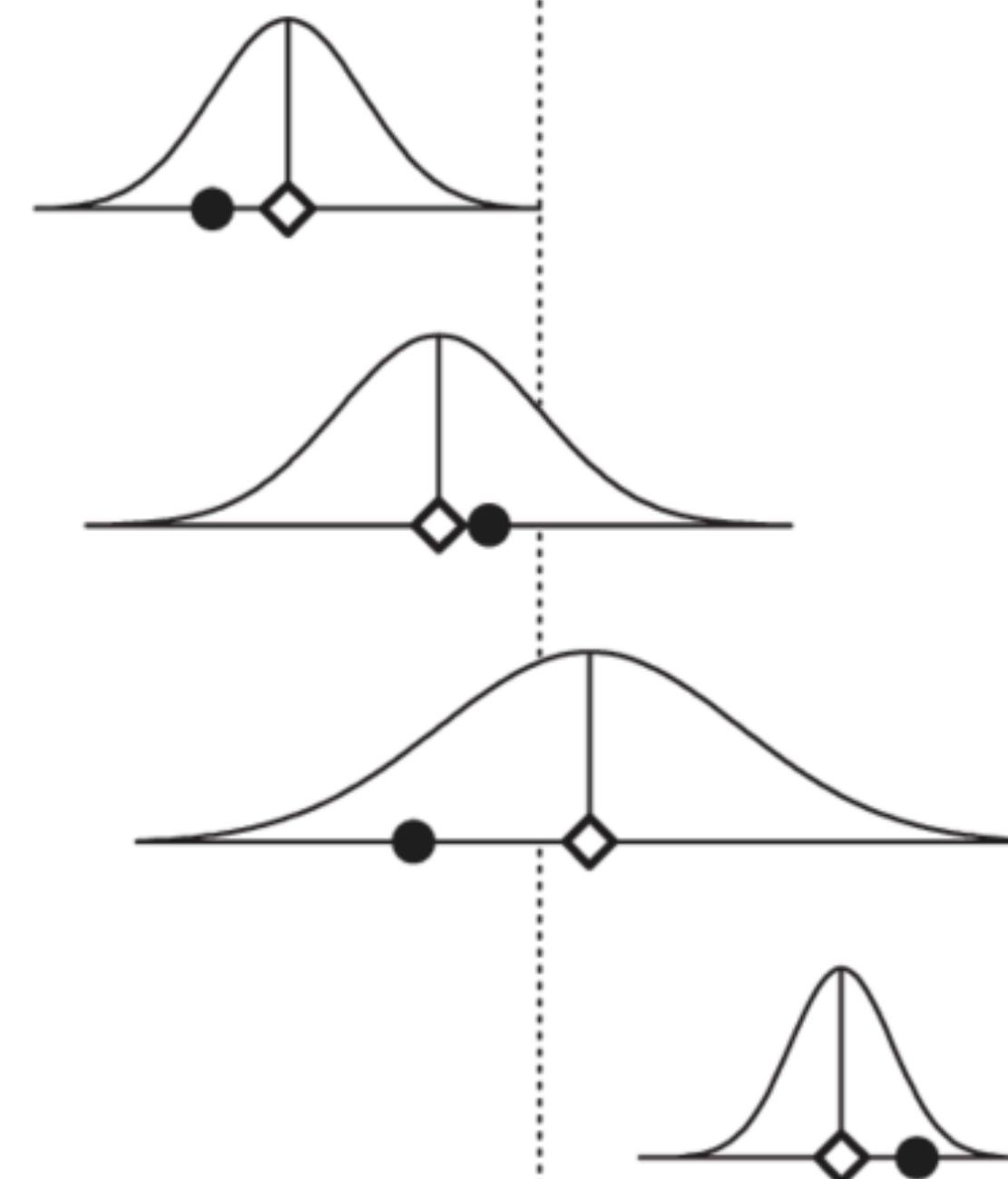
### First Stage

A  $\theta_i$  value is randomly sampled from a normal distribution with expected value  $\mu$  and variance  $\tau^2_\theta$  for each study that is conducted.



### Second Stage

Conditional on the  $\theta_i$  value for a particular study, a sampling distribution of  $y_i$  exists, but the selection of a specific sample leads to a single observed outcome. The amount of sampling variability is a function of the sample size in each study. Therefore, observed outcomes based on larger samples tend to be closer to their respective  $\theta_i$  value.



## Modelos mixtos

Un valor muestrado de una distribución normal para el parámetro

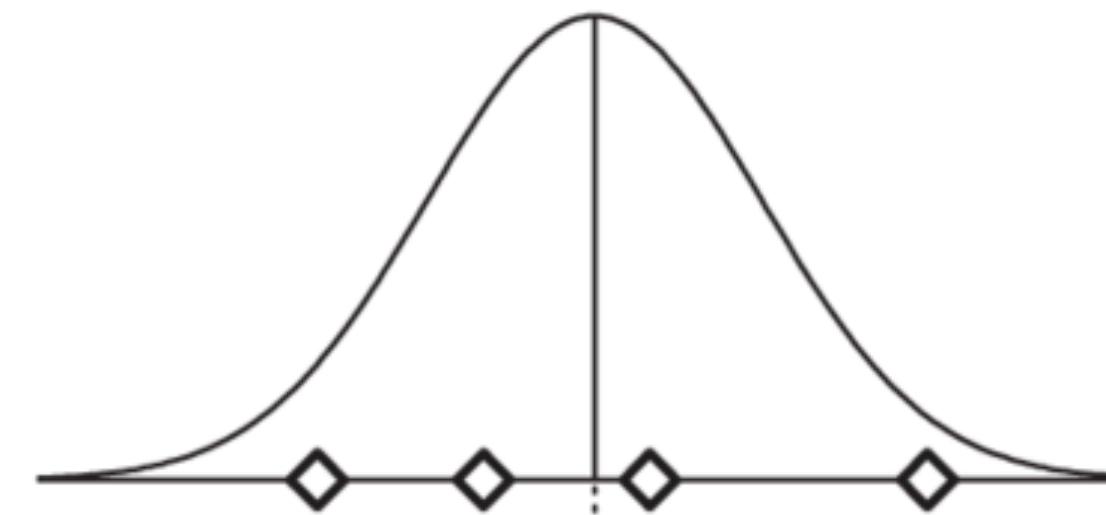
$$\theta_i \sim N(\mu, \sigma_i^2)$$

Condicional al  $\theta_i$ , una distribución de muestra existe para cada nivel aleatorio

data observado  $y_i$

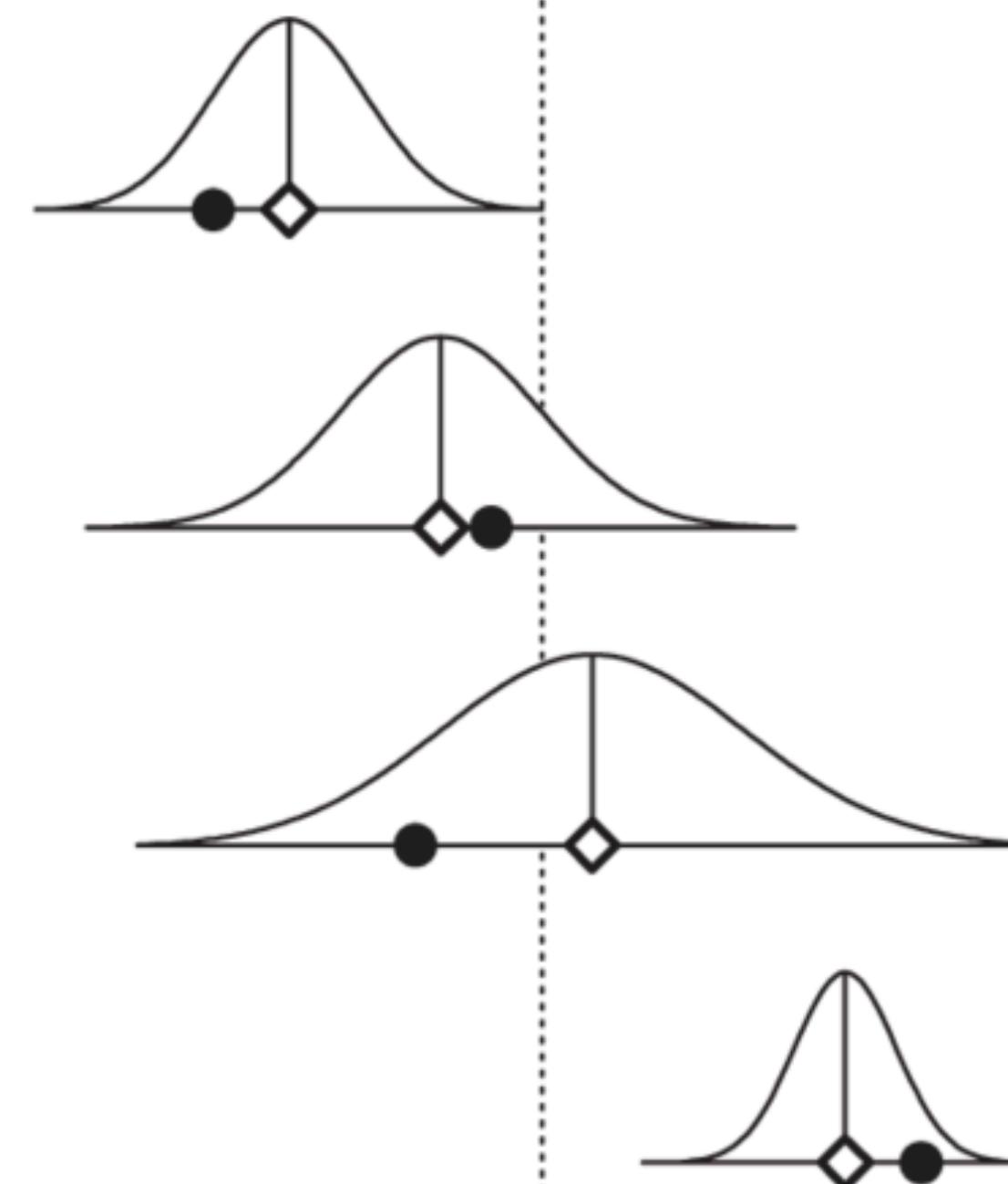
### First Stage

A  $\theta_i$  value is randomly sampled from a normal distribution with expected value  $\mu$  and variance  $\tau^2_\theta$  for each study that is conducted.



### Second Stage

Conditional on the  $\theta_i$  value for a particular study, a sampling distribution of  $y_i$  exists, but the selection of a specific sample leads to a single observed outcome. The amount of sampling variability is a function of the sample size in each study. Therefore, observed outcomes based on larger samples tend to be closer to their respective  $\theta_i$  value.



### Observed Data

The observed data consist of the  $y_i$  values and an estimate of the amount of sampling variability in each study.



# IV. Modelos lineales generalizados

---

Modelo lineal:

$$y_i = \alpha + \beta * x_i + \epsilon_i$$

$$\epsilon_i \sim N(0, \sigma^2)$$

Modelo lineal:

$$y_i = \alpha + \beta * x_i + \epsilon_i$$

$$\epsilon_i \sim N(0, \sigma^2)$$

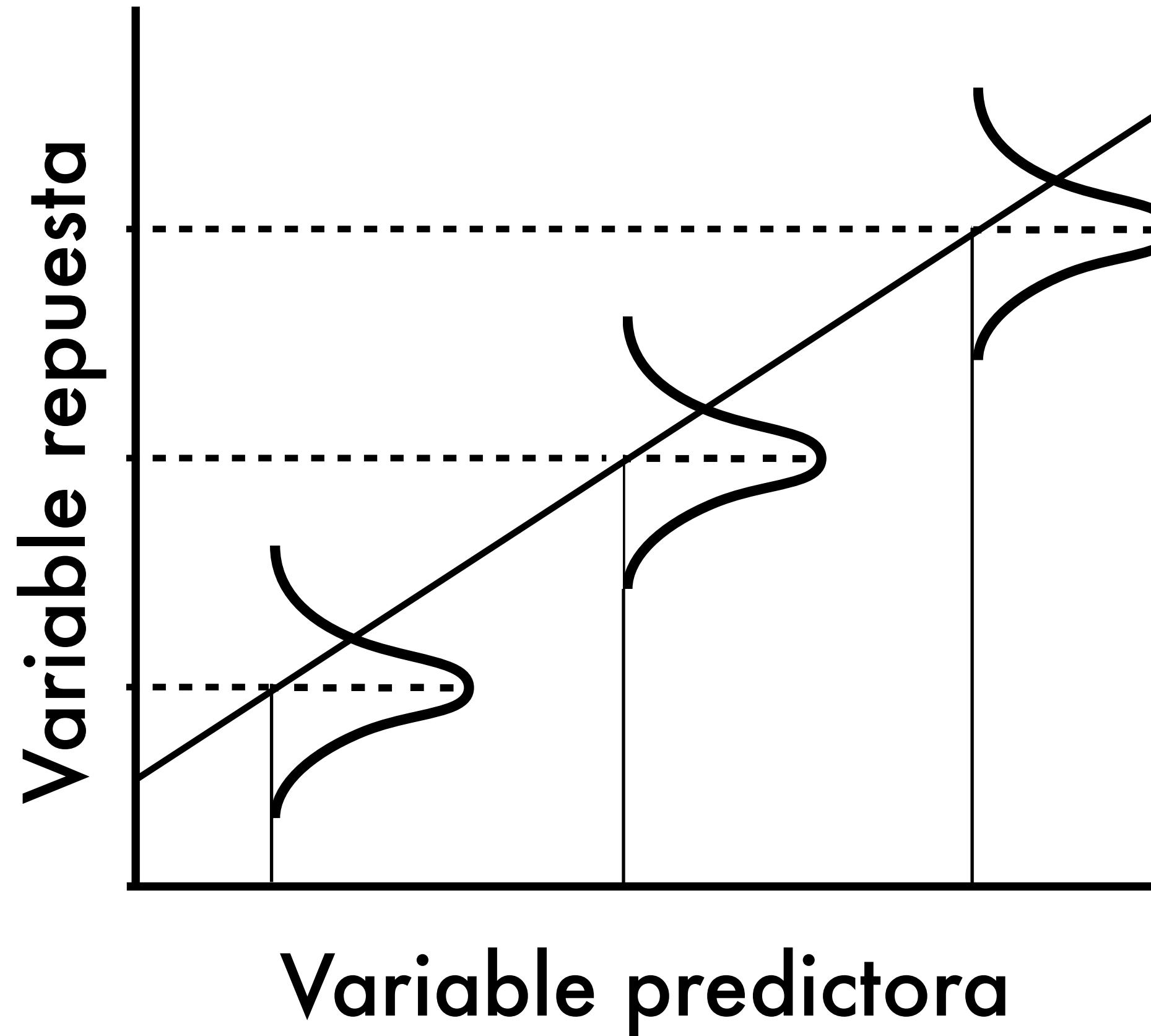
El matemática también  
puede ser escrito así:

$$y_i \sim N(\mu_i, \sigma^2)$$

$$\mu_i = \alpha + \beta * x_i$$

Recuerden supuesto 2 de modelos lineales:

*Residuales se ajustan a una distribución normal*



$$y_i = \alpha + \beta * x_i + \epsilon_i$$

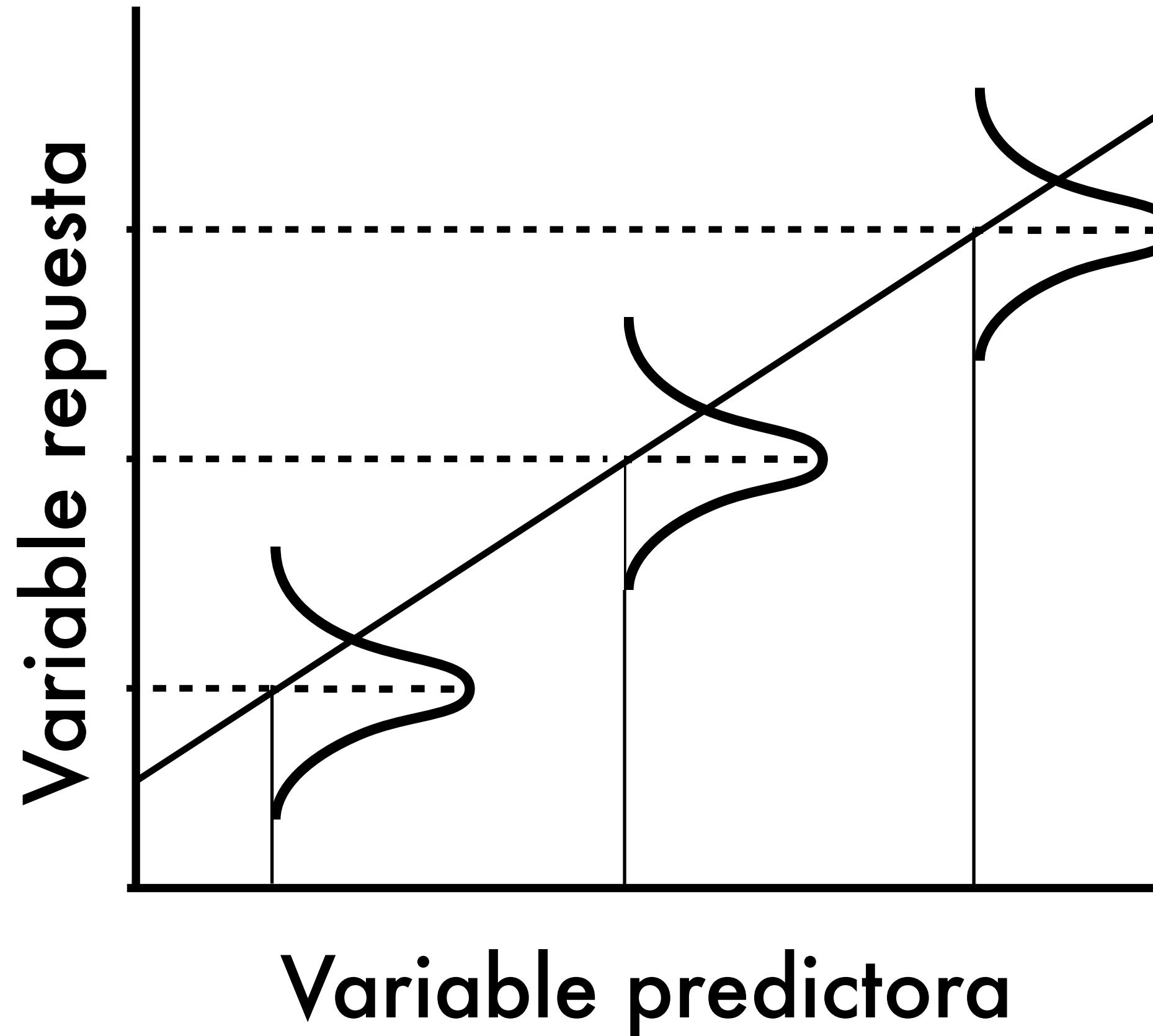
$$\epsilon_i \sim N(0, \sigma^2)$$

$$y_i \sim N(\mu_i, \sigma^2)$$

$$\mu_i = \alpha + \beta * x_i$$

Recuerden supuesto 2 de modelos lineales:

*Residuales se ajustan a una distribución normal*

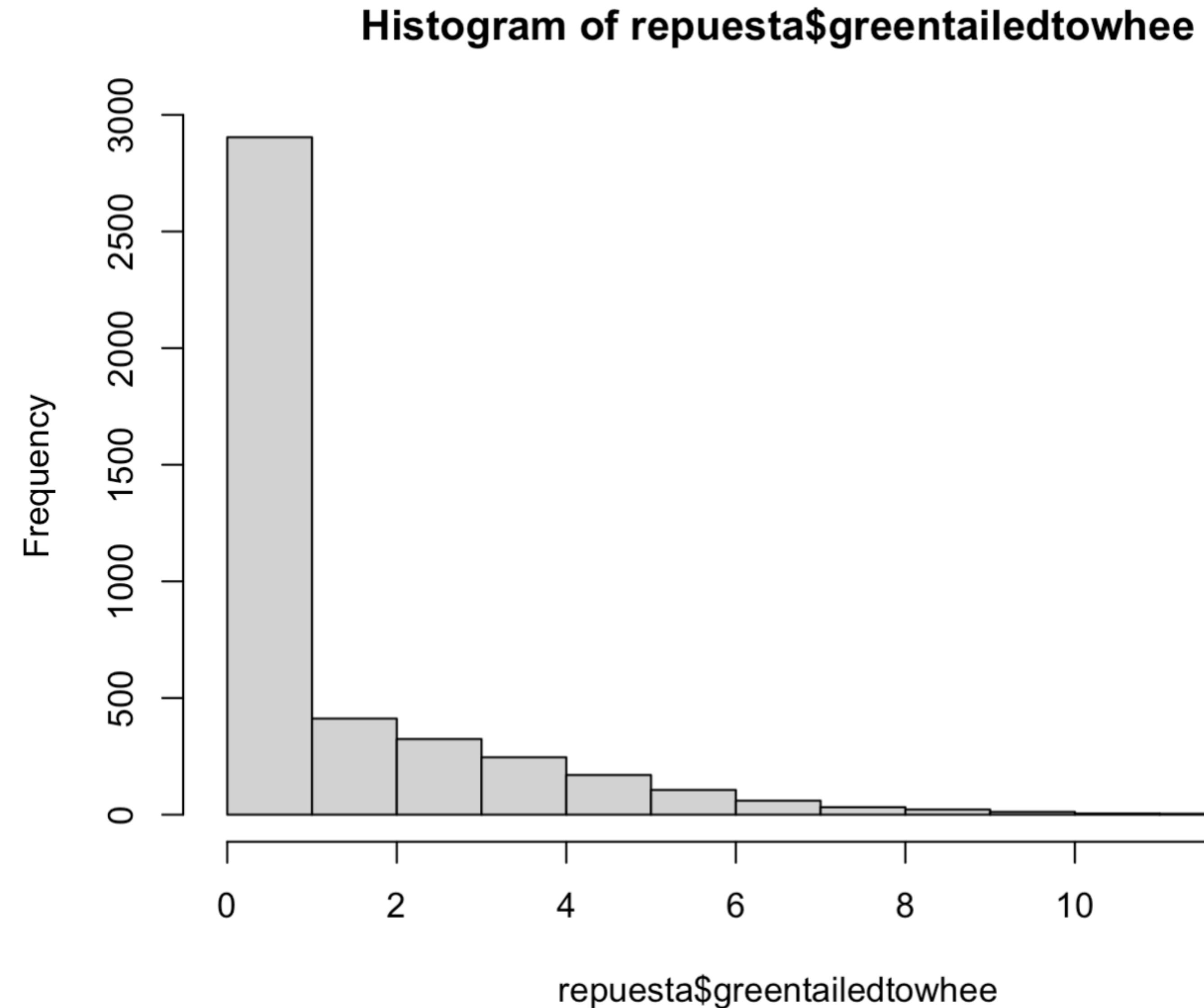


Pero no siempre tenemos  
variables respuestas  
continuos y distribuido en  
manera normal!

## Modelos lineales generalizados

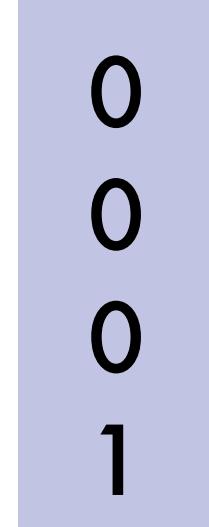
Por ejemplo:

1. **Conteos** no pueden ser negativo, o non-integer



Por ejemplo:

1. **Conteos no pueden ser negativo, o non-integer**
2. **Presencia/ausencia** solo tienen dos valores (0 o 1)



0  
0  
0  
1

## **Modelos lineales generalizados**

---

Podemos transformar la variable respuesta... e.g. `log()`

pero es menos poderoso comparado con las otras herramientas disponibles.

Podemos transformar la variable respuesta... e.g. log()

pero es menos poderoso comparado con las otras herramientas disponibles.

y también puede darte una solución incorrecta.

# Methods in Ecology and Evolution



British Ecological Society

*Methods in Ecology and Evolution* 2010, 1, 118–122

doi: 10.1111/j.2041-210X.2010.00021.x

## Do not log-transform count data

Robert B. O'Hara<sup>1\*</sup> and D. Johan Kotze<sup>2</sup>

<sup>1</sup>Biodiversity and Climate Research Centre, Senckenberganlage 25, D-60325 Frankfurt am Main, Germany and

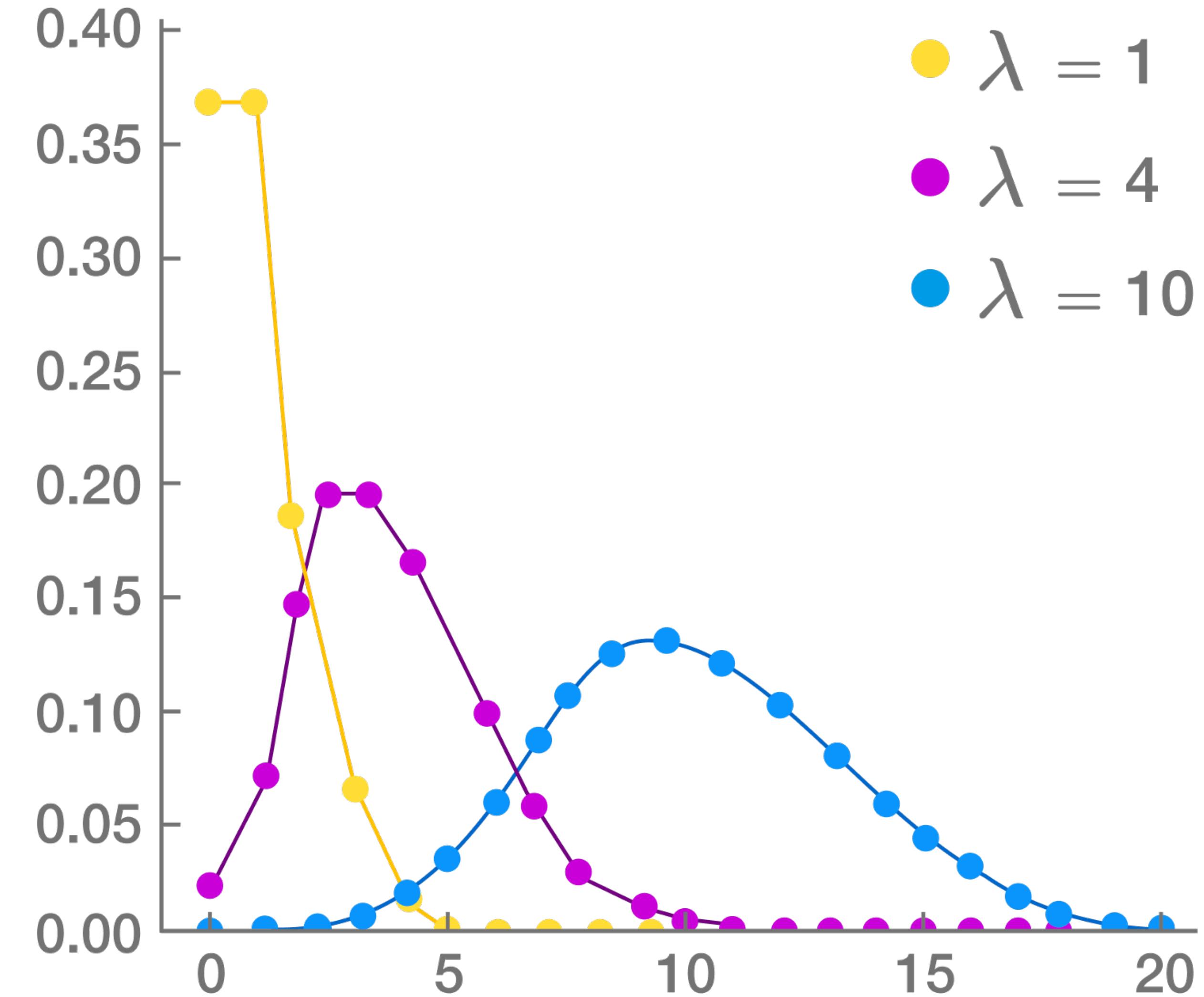
<sup>2</sup>Department of Environmental Sciences, PO Box 65, University of Helsinki, Helsinki FI-00014, Finland

Conteos

Distribución Poisson

con parametro  $\lambda_i$

que significa el valor  
esperado (i.e. el mas  
probable)



Presencia/ausencia

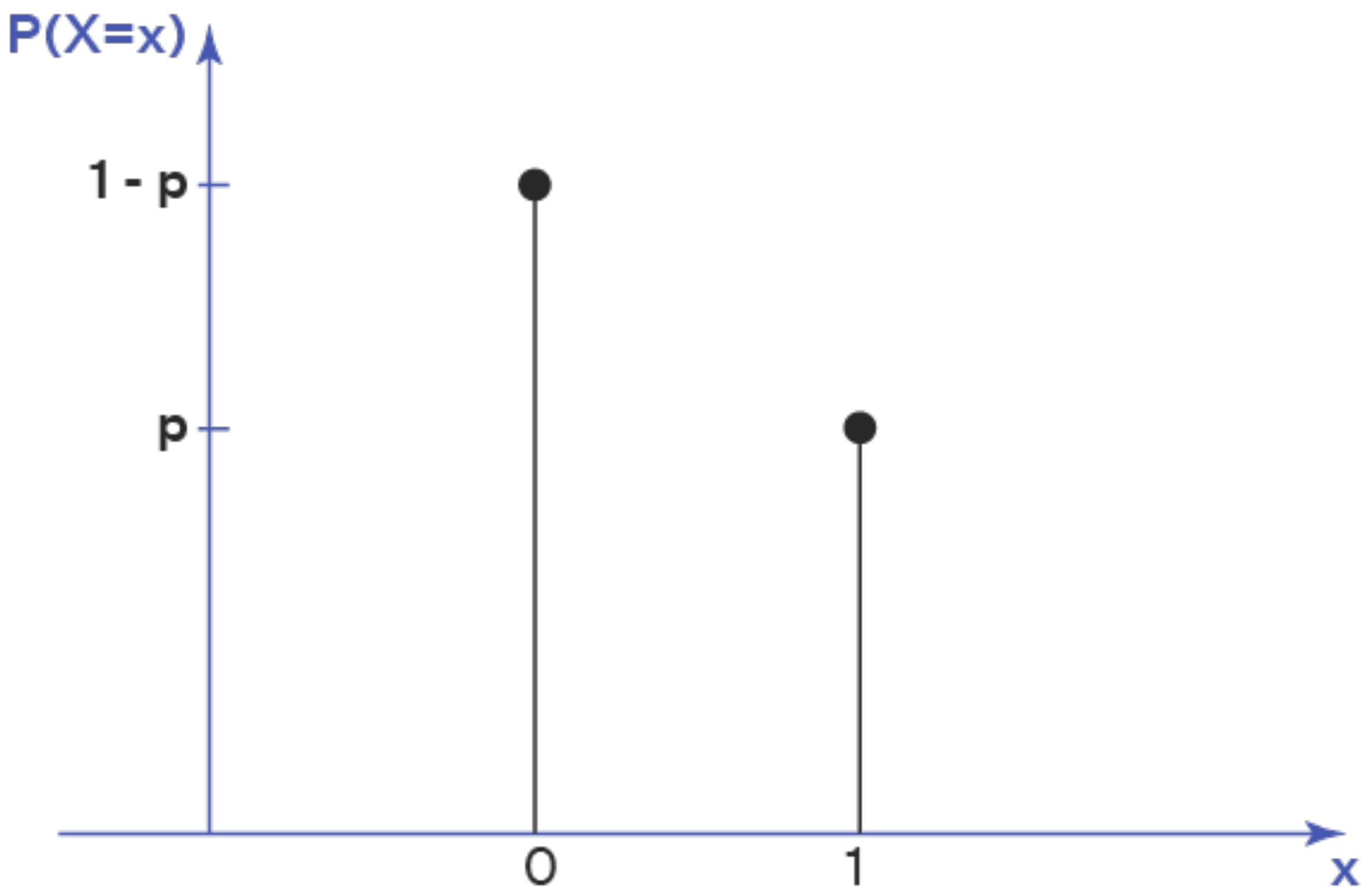
Distribución Bernoulli

con parametro  $p_i$

que significa el valor esperado (i.e. el mas probable)

Bernoulli Distribution Graph

$X \sim \text{Bernoulli}(p)$



Presencia/ausencia

Distribución Bernoulli

con parametro  $p_i$

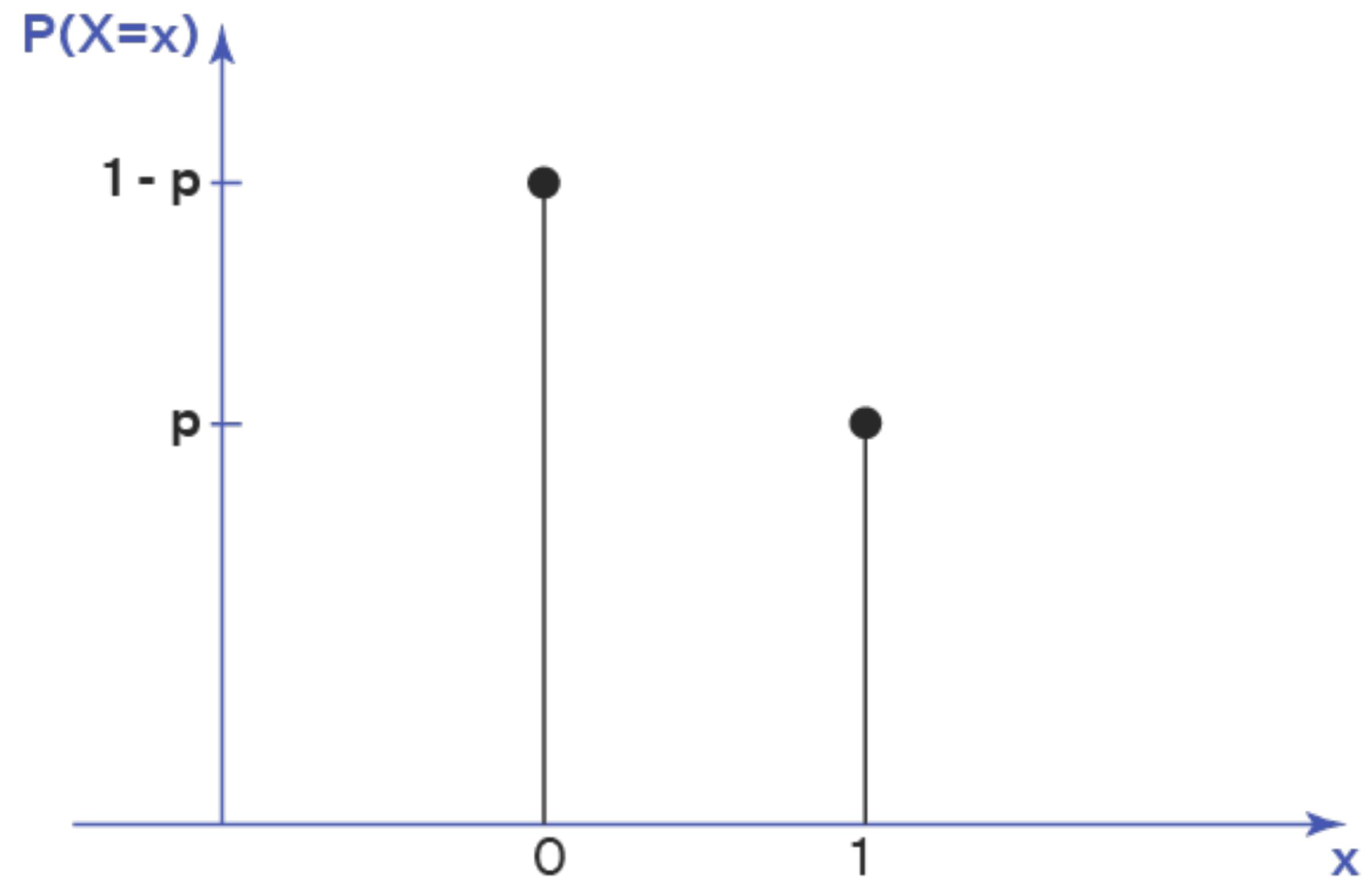
Distribución Binomial  
es un caso especial  
cuando hay un intento.



shutterstock.com · 1317073241

Bernoulli Distribution Graph

$X \sim \text{Bernoulli}(p)$



## Modelos lineales generalizados

---

Normal

Distribución  $y_i \sim N(\mu_i, \sigma^2)$

Componente  
lineal  $\mu_i = \alpha + \beta * x_i$

## Modelos lineales generalizados

	Normal	Poisson
Distribución	$y_i \sim N(\mu_i, \sigma^2)$	$y_i \sim Poisson(\lambda_i)$
Componente lineal	$\mu_i = \alpha + \beta * x_i$	$\mu_i = \alpha + \beta * x_i$

## Modelos lineales generalizados

	Normal	Poisson
Distribución	$y_i \sim N(\mu_i, \sigma^2)$	$y_i \sim Poisson(\lambda_i)$
Función de enlace		$\lambda_i = e^{\mu_i}$
Componente lineal	$\mu_i = \alpha + \beta * x_i$	$\mu_i = \alpha + \beta * x_i$

para asegurar que lambda sea positivo

## Modelos lineales generalizados

	Normal	Poisson	Bernoulli/Binomial
Distribución	$y_i \sim N(\mu_i, \sigma^2)$	$y_i \sim Poisson(\lambda_i)$	$y_i \sim Bernoulli(p_i)$
Función de enlace		$\lambda_i = e^{\mu_i}$	
Componente lineal	$\mu_i = \alpha + \beta * x_i$	$\mu_i = \alpha + \beta * x_i$	$\mu_i = \alpha + \beta * x_i$

para asegurar que lambda sea positivo

## Modelos lineales generalizados

	Normal	Poisson	Bernoulli/Binomial
Distribución	$y_i \sim N(\mu_i, \sigma^2)$	$y_i \sim Poisson(\lambda_i)$	$y_i \sim Bernoulli(p_i)$
Función de enlace	$\mu_i = \mu_i$	$\lambda_i = e^{\mu_i}$	$p_i = \frac{1}{1 + e^{-\mu_i}}$
Componente lineal	$\mu_i = \alpha + \beta * x_i$	$\mu_i = \alpha + \beta * x_i$	$\mu_i = \alpha + \beta * x_i$

**para asegurar que lambda sea positivo**

**para asegurar que p\_i sea entre 0 y 1**

## Modelos lineales generalizados

Funcione de  
enlace

Poisson

$$\lambda_i = e^{\mu_i}$$

Bernoulli/Binomial

$$p_i = \frac{1}{1 + e^{-\mu_i}}$$

## Modelos lineales generalizados

Funcione de  
enlace

Poisson

$$\lambda_i = e^{\mu_i}$$



$$\log(\lambda_i) = \mu_i$$

Bernoulli/Binomial

$$p_i = \frac{1}{1 + e^{-\mu_i}}$$



$$\text{logit}(p_i) = \mu_i$$

## Modelos lineales generalizados

Mas  
distribuciones

```
glm(formula, family=familytype(link=linkfunction), data=)
```

Funcione de enlace defecta

Family	Default Link Function
"normal"	binomial (link = "logit")
	gaussian (link = "identity")
	Gamma (link = "inverse")
	inverse.gaussian (link = "1/mu^2")
	poisson (link = "log")
	quasi (link = "identity", variance = "constant")
	quasibinomial (link = "logit")
	quasipoisson (link = "log")

**El manera de interpretar/ilustrar las estimaciones de tu GLM depende de la pregunta.**

**Prueba de ratio de probabilidad**

Cuanto mejor es un modelo comparado con otro?

**Criterio de información**

Qué tan bueno sea el modelo explicando los datos, considerando sobre-ajuste?

**Validación cruzada**

Qué tan bueno sea el modelo a predecir nuevos datos?

## **Modelos lineales generalizados**

---

**No podemos comparar modelos ajustado a datos diferentes, o modelos usando diferentes distribuciones.**

**Podemos solamente comparar diferentes combinaciones de parámetros, porque es lo mismo que imaginar estas variables predictoras puesto a 0.**

## Modelos lineales generalizados

Practica en R: *script-part4.R*



**Muchas gracias!!**

**Financia:**

  
**consortium**

**Invita:**

