Curso-Taller de R para investigadores

UMSA, La Paz, Bolivia

23 - 25 Feb 2023

Saras Windecker & David Uribe

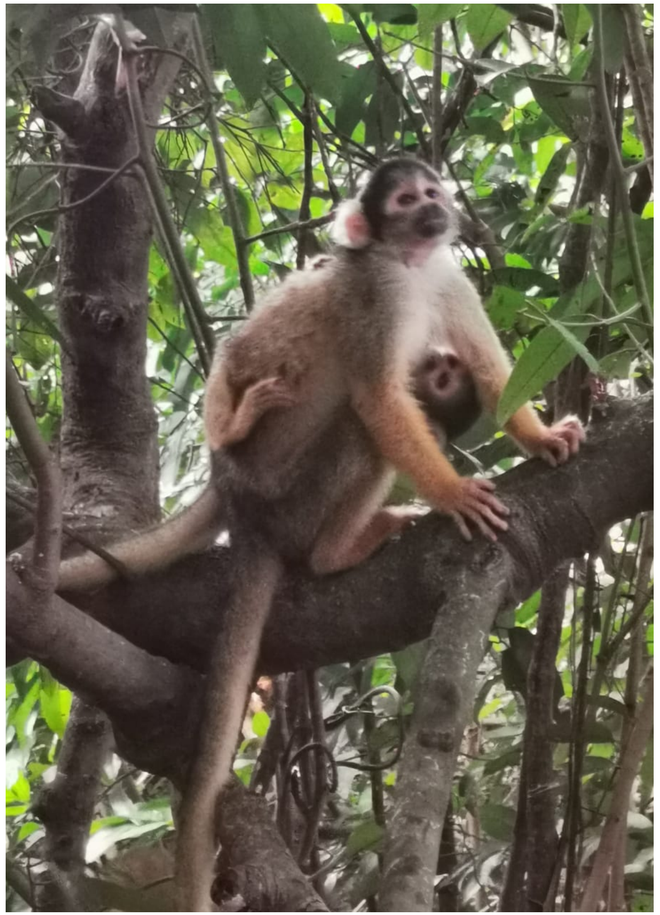
Código de Conducta

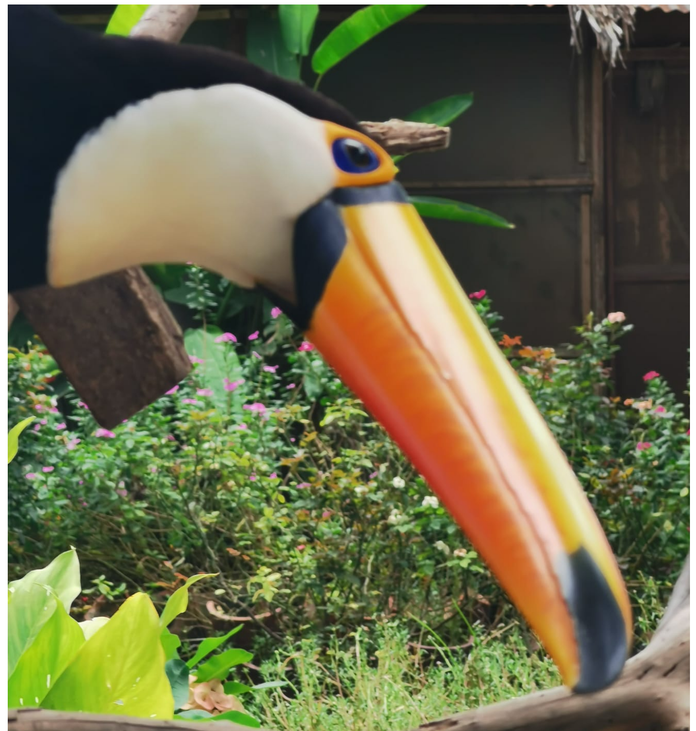
https://github.com/smwindecker/R-para-ecologia

Financia:

consortium

Invita:





Qué esperan aprender en este curso?

Con qué tipos de datos trabajan?

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

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

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

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

- 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

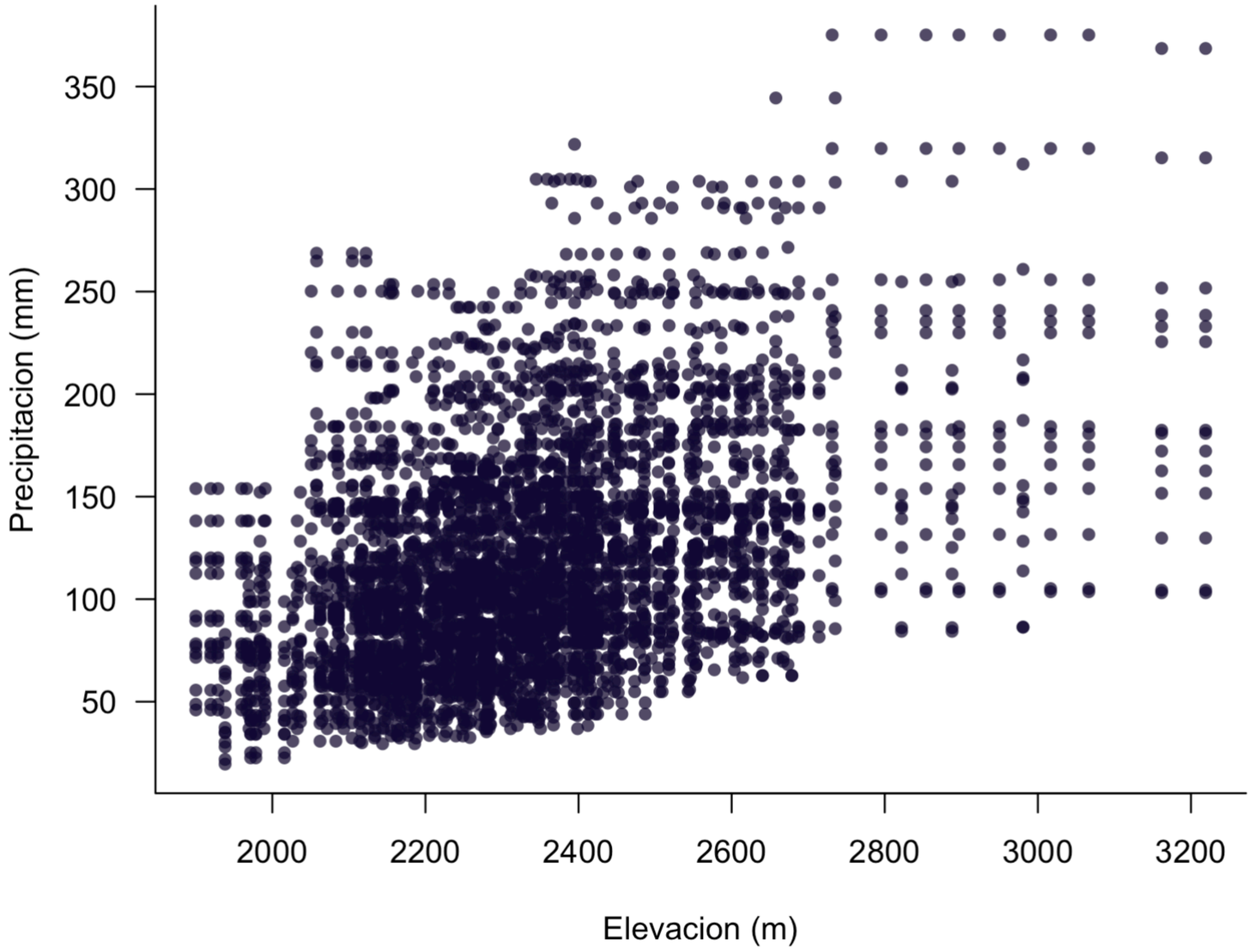
Presentar resultados

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

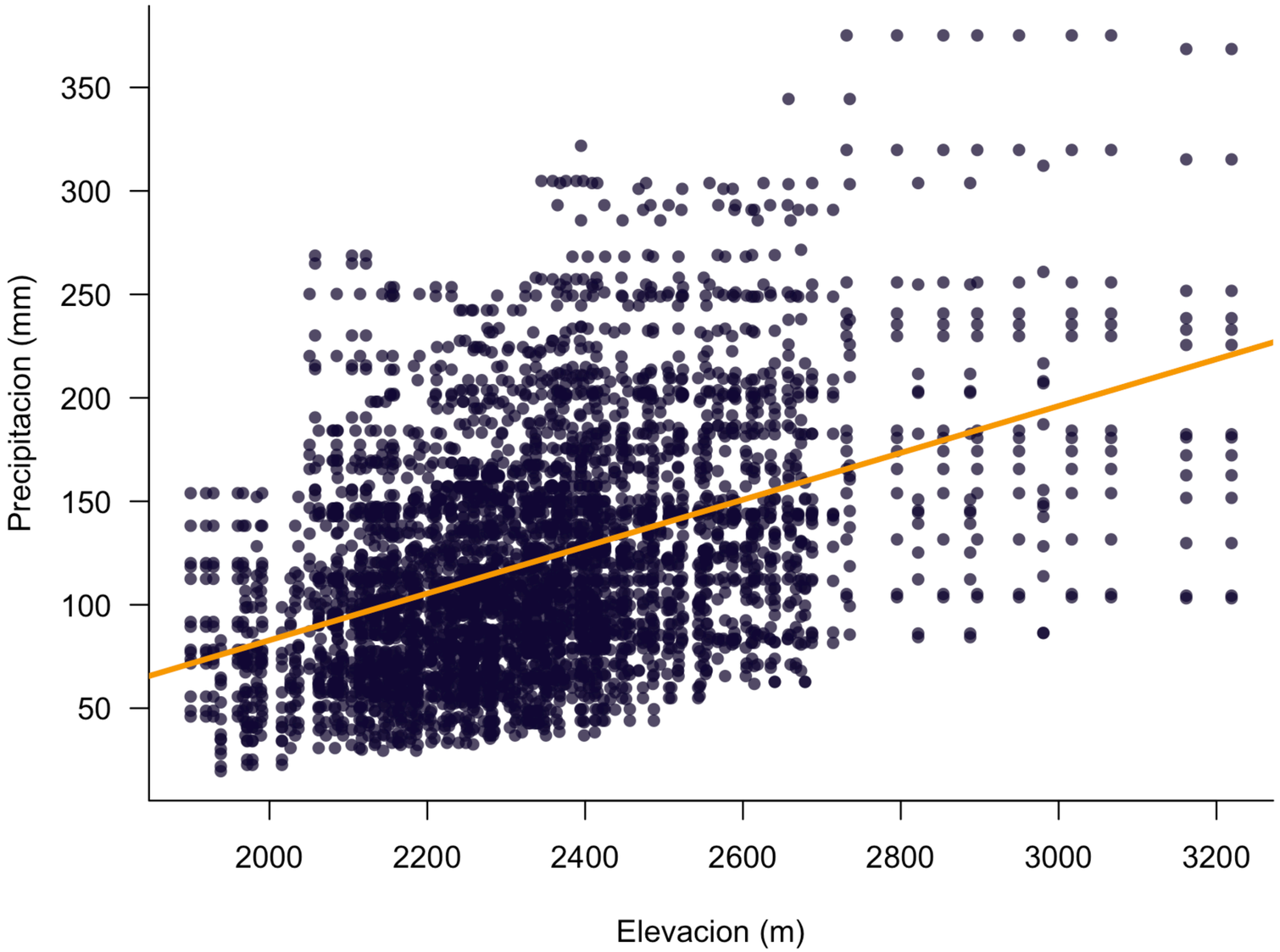
example with anova?

II. Modelos lineales

Modelos lineales



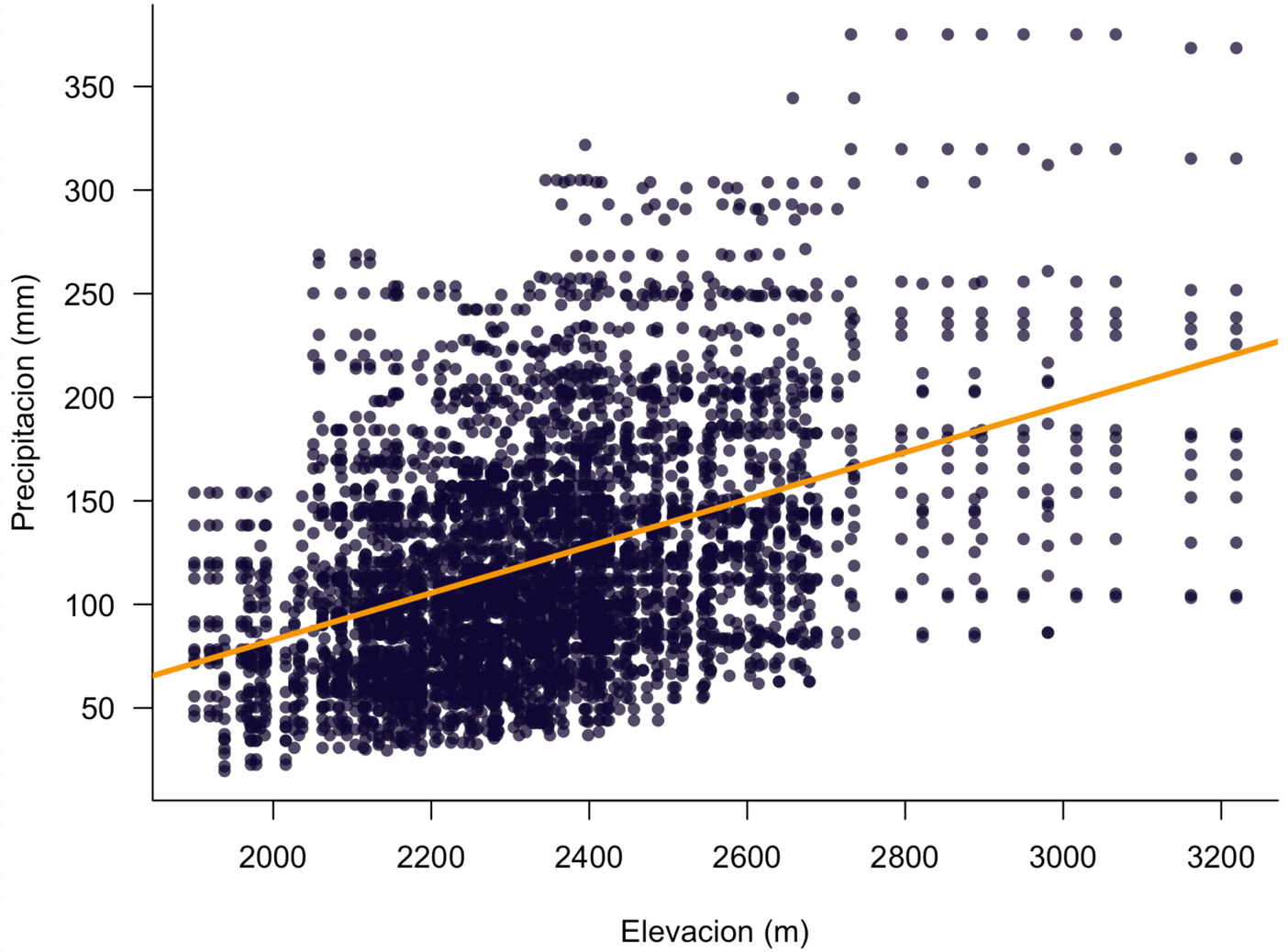
Modelos lineales

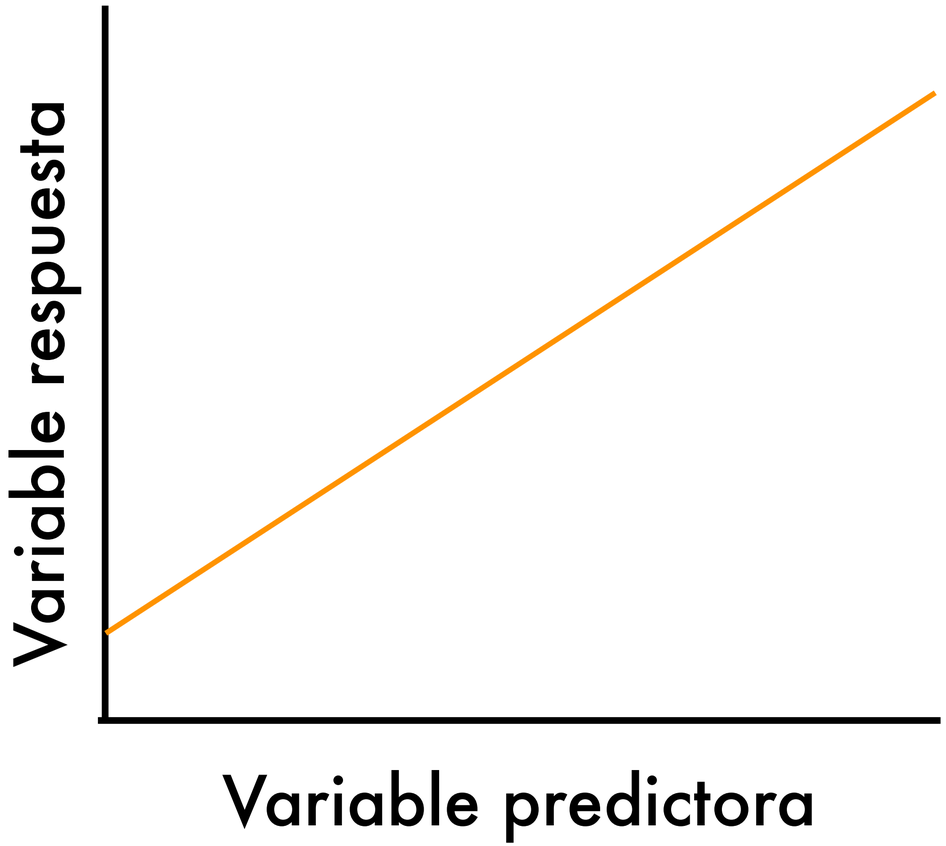


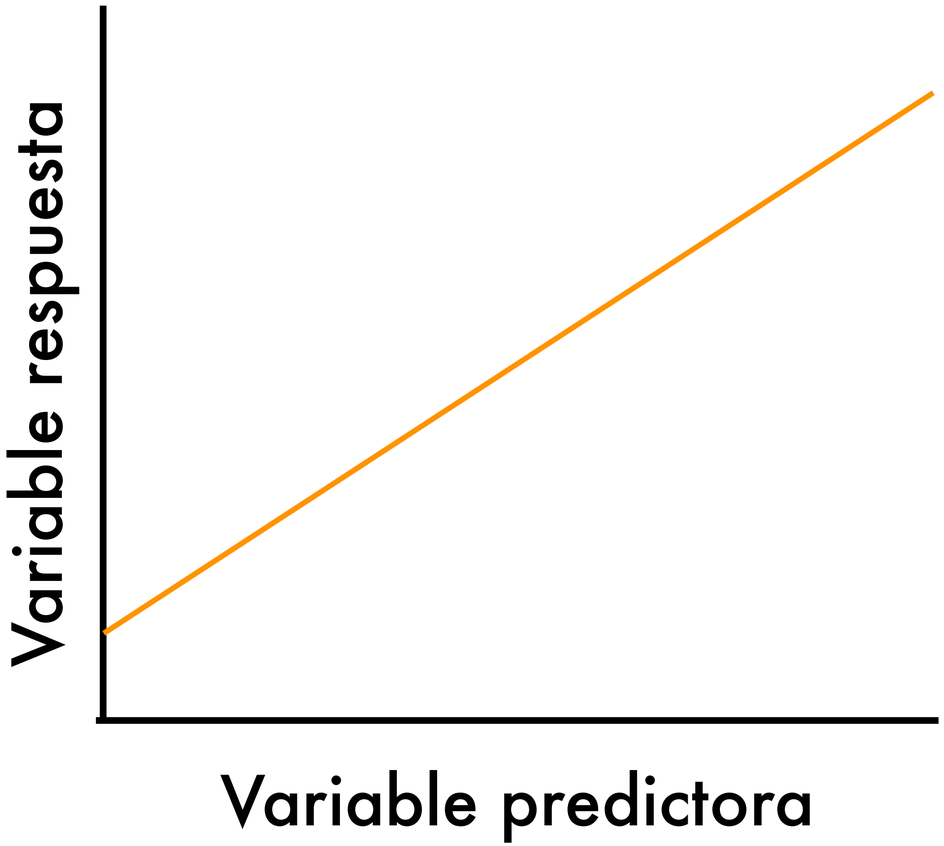
Qué caracteriza este ejemplo?

variable repuesta continua

variable predictora continua

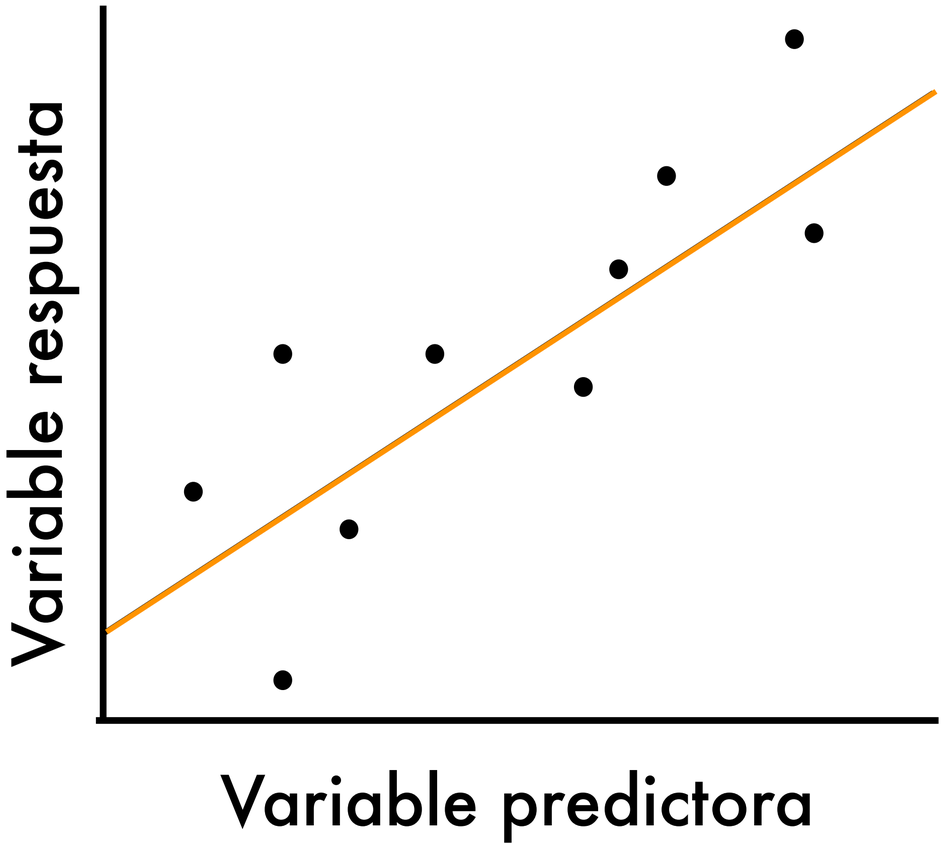



Variable respuesta
Variable predictora



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

Queremos estimar los valores mas probables de α y β

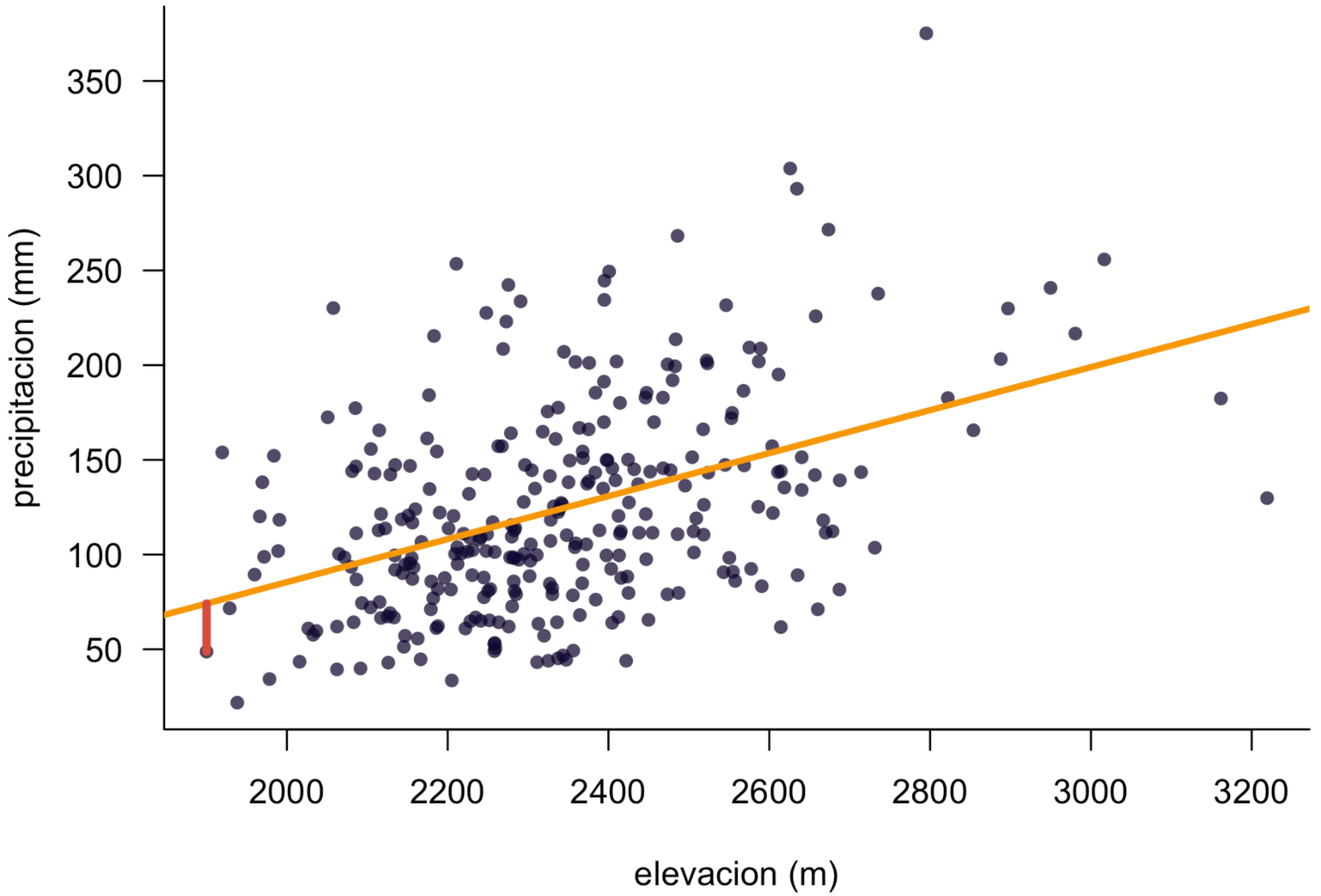


$$y_i = \alpha + \beta * x_i + \epsilon_i$$

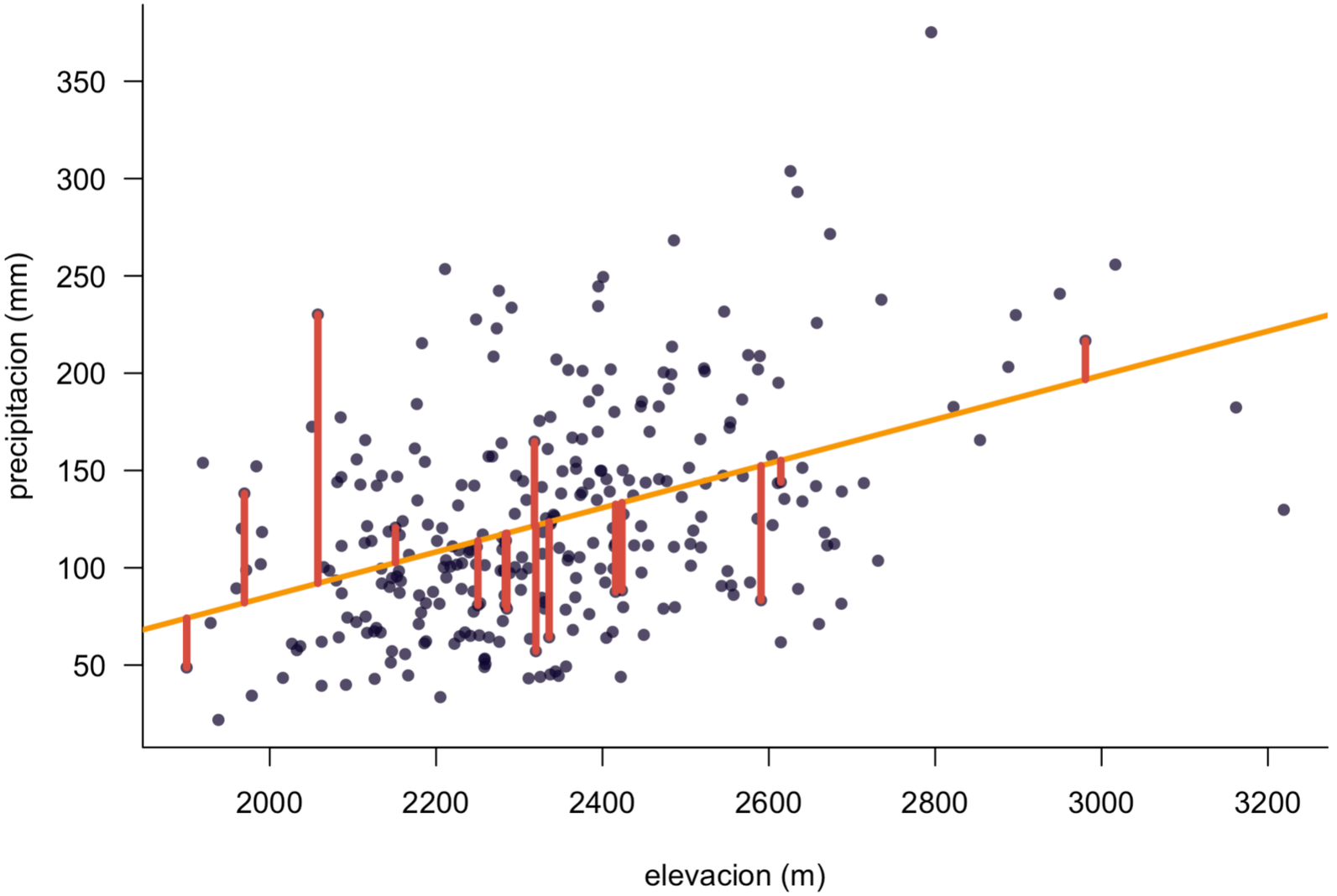
Las diferencias entre las observaciones y la linea es el error residual ϵ_i

Como podemos entender el ϵ?

Como podemos entender el ϵ?



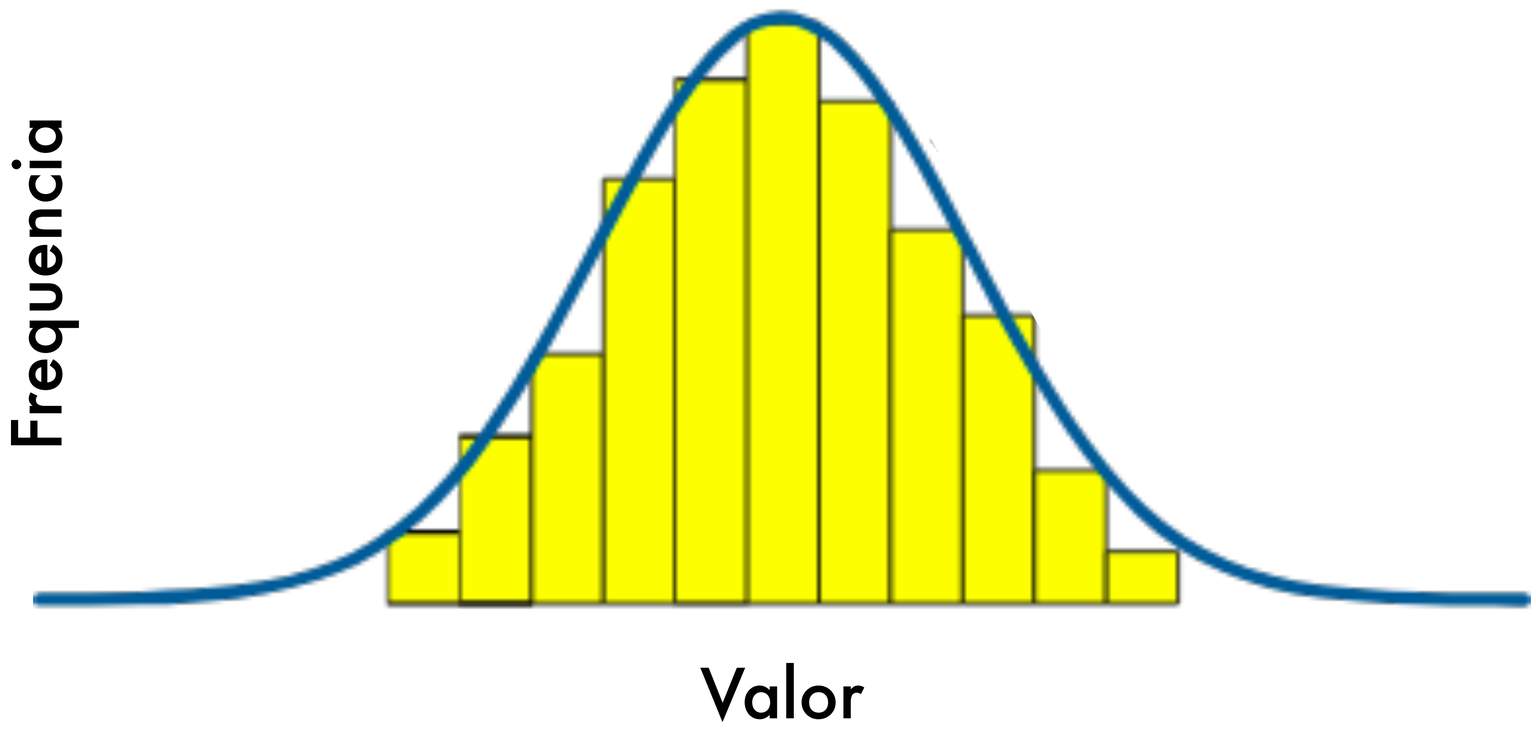
Como podemos entender el ϵ?



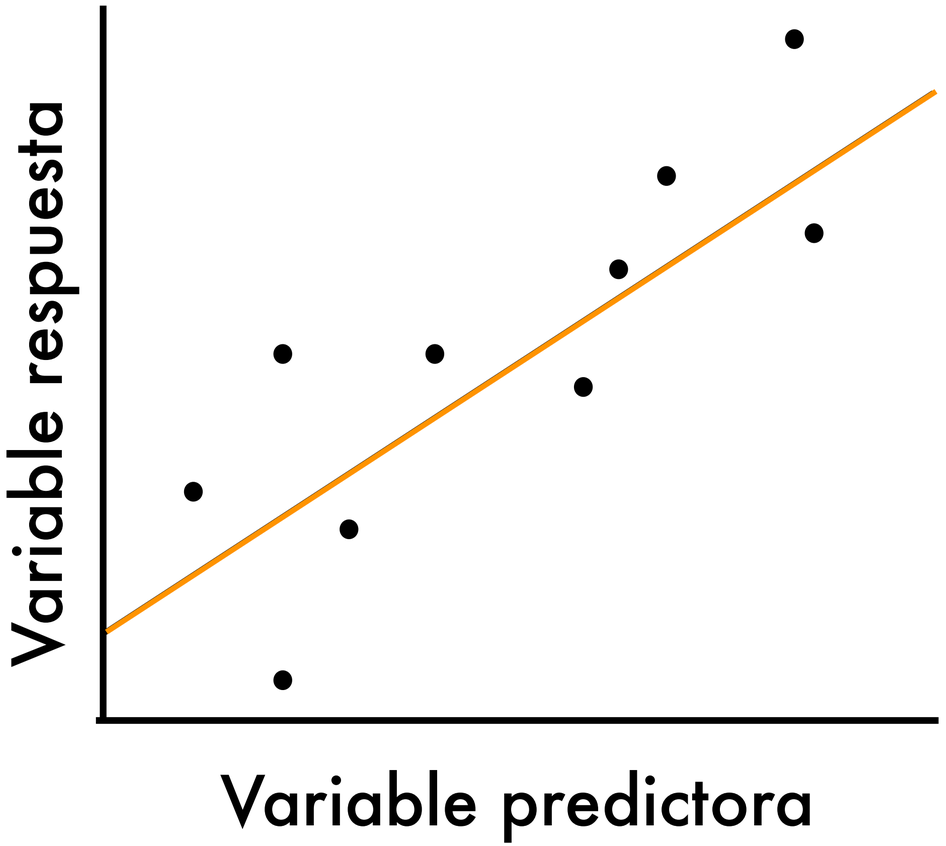
Supuesto de regresión lineal 1: Observaciones son independientes

- 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

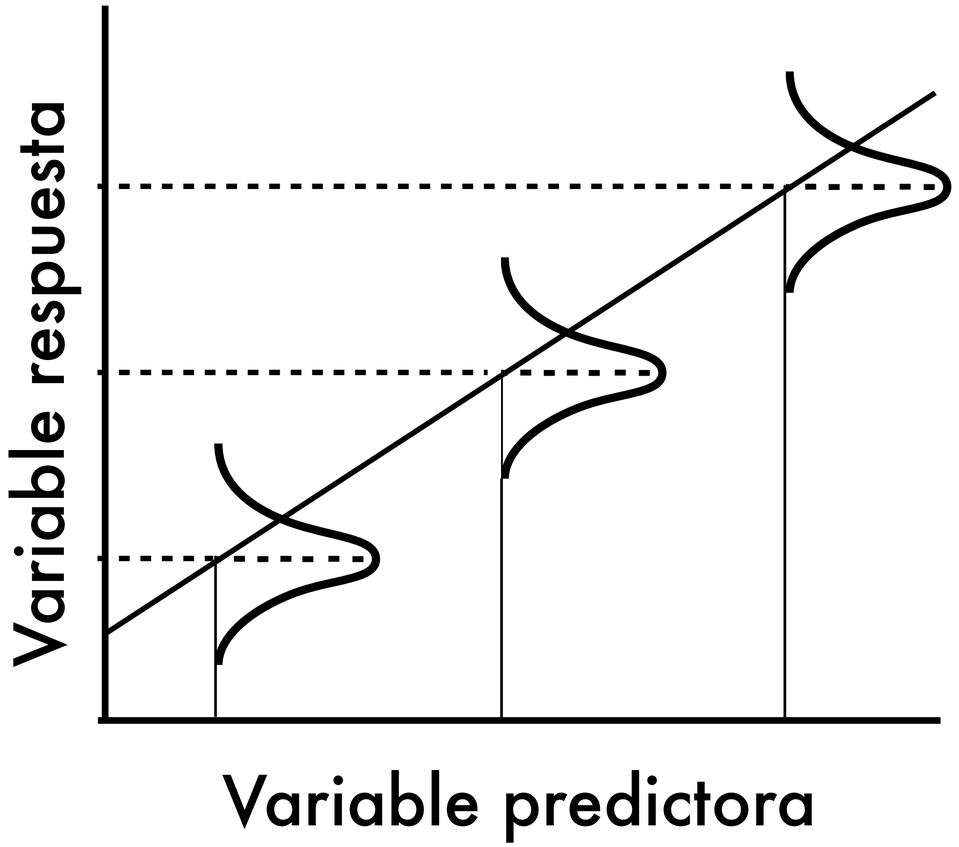


Supuesto 2: Residuales se ajustan a una distribución normal



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

Supuesto 2: Residuales se ajustan a una distribución normal



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

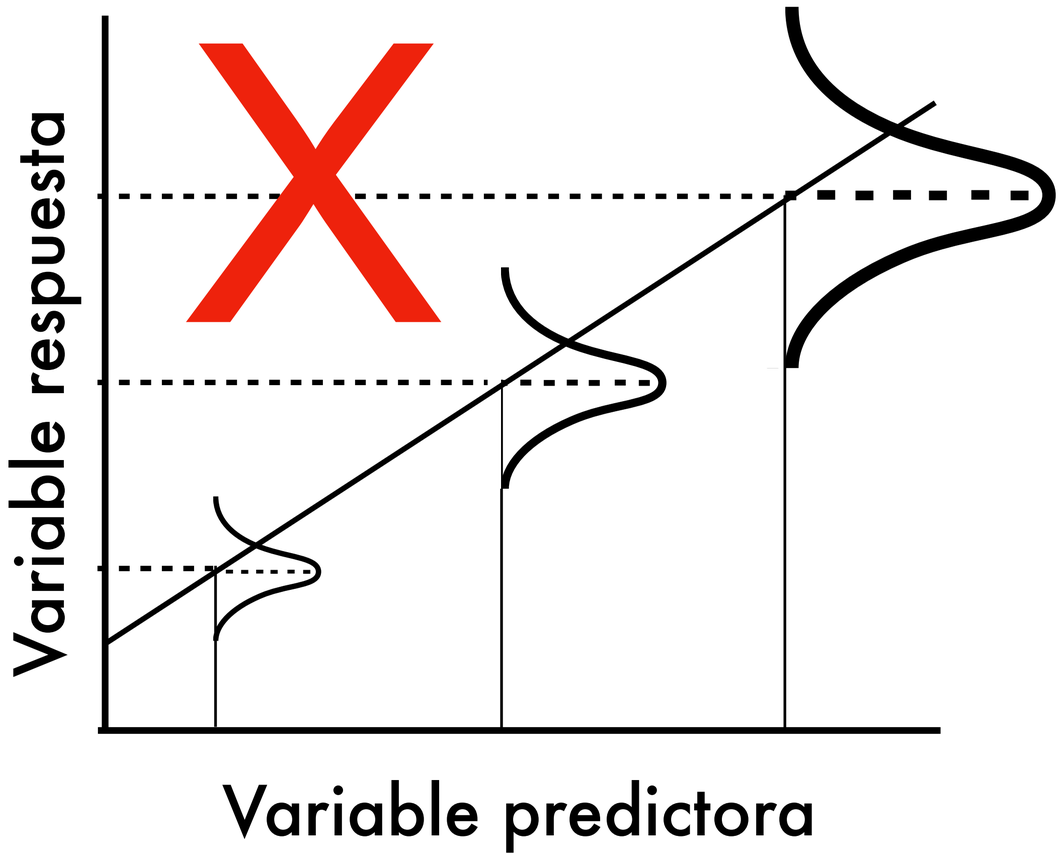
$$\epsilon_i \sim N(0,\sigma^2)$$

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

practica en R

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 tu pregunta.

Qué tan bien se ajusta el modelo a los datos? usa R2

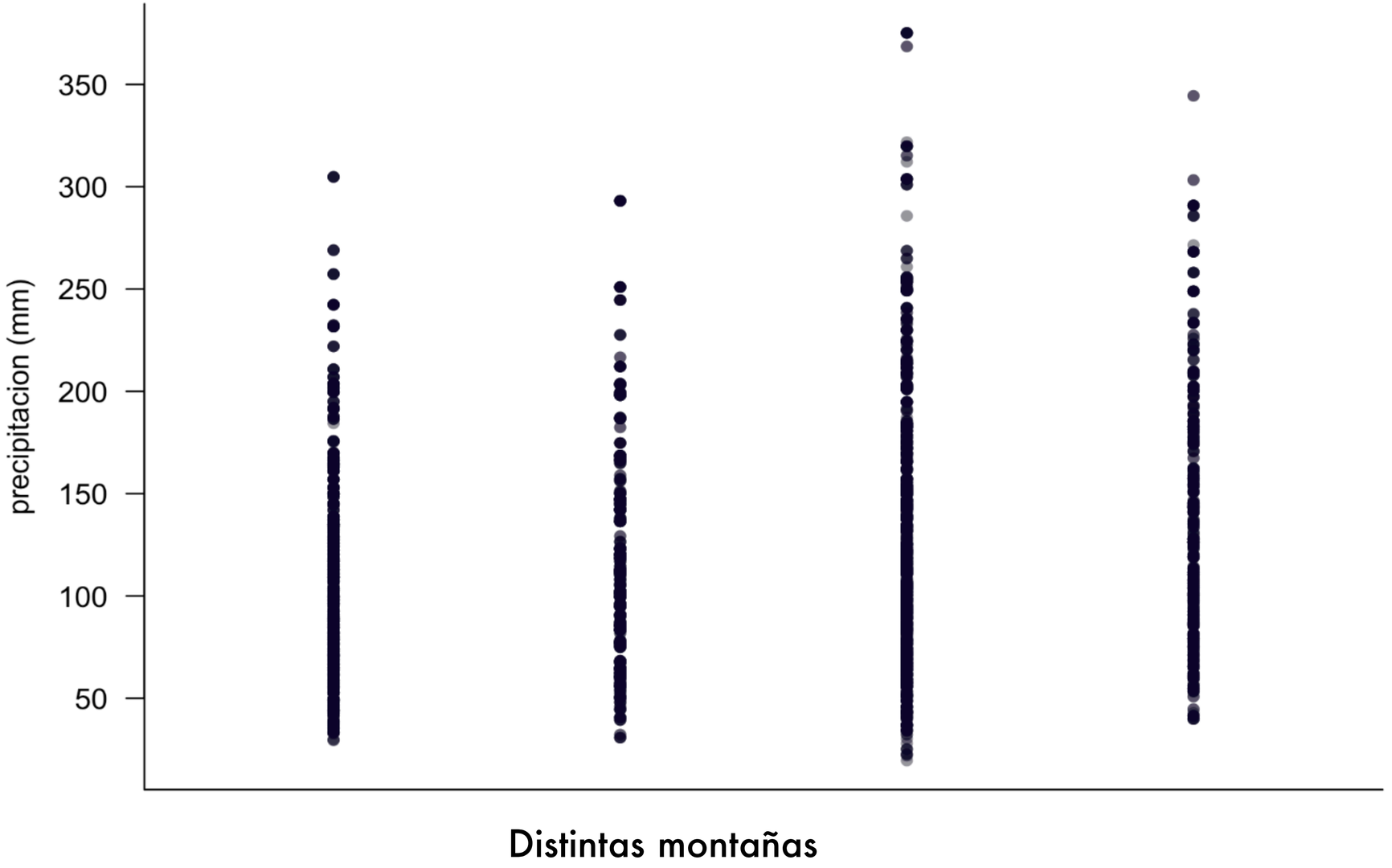
El manera de interpretar/ilustrar las estimaciones de un modelo depende de tu pregunta.

Qué tan bien se ajusta el modelo a los datos?	usa R2
Hay apoyo estadística para una asociación?	usa p-values

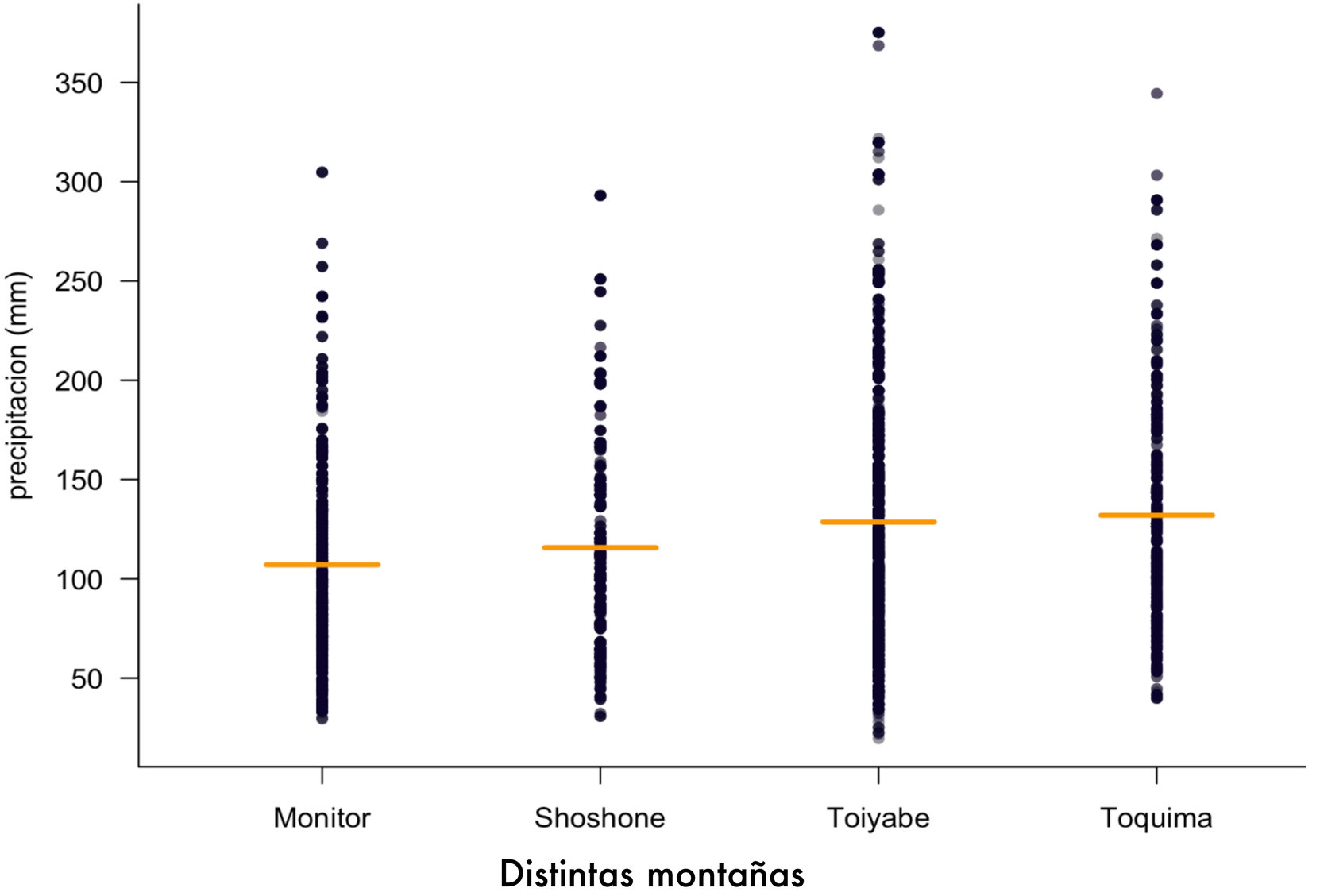
El manera de interpretar/ilustrar las estimaciones de un modelo depende de tu pregunta.

Qué tan bien se ajusta el modelo a los datos?	usa R2
Hay apoyo estadística para una asociación?	usa p-values
Una asociación estadísticamente significativa tiene sentido?	mira los coeficientes (tamaño y signo)

ANOVA


precipitacion (mm)
350
300
250
200
150
100
50
Distintas montañas

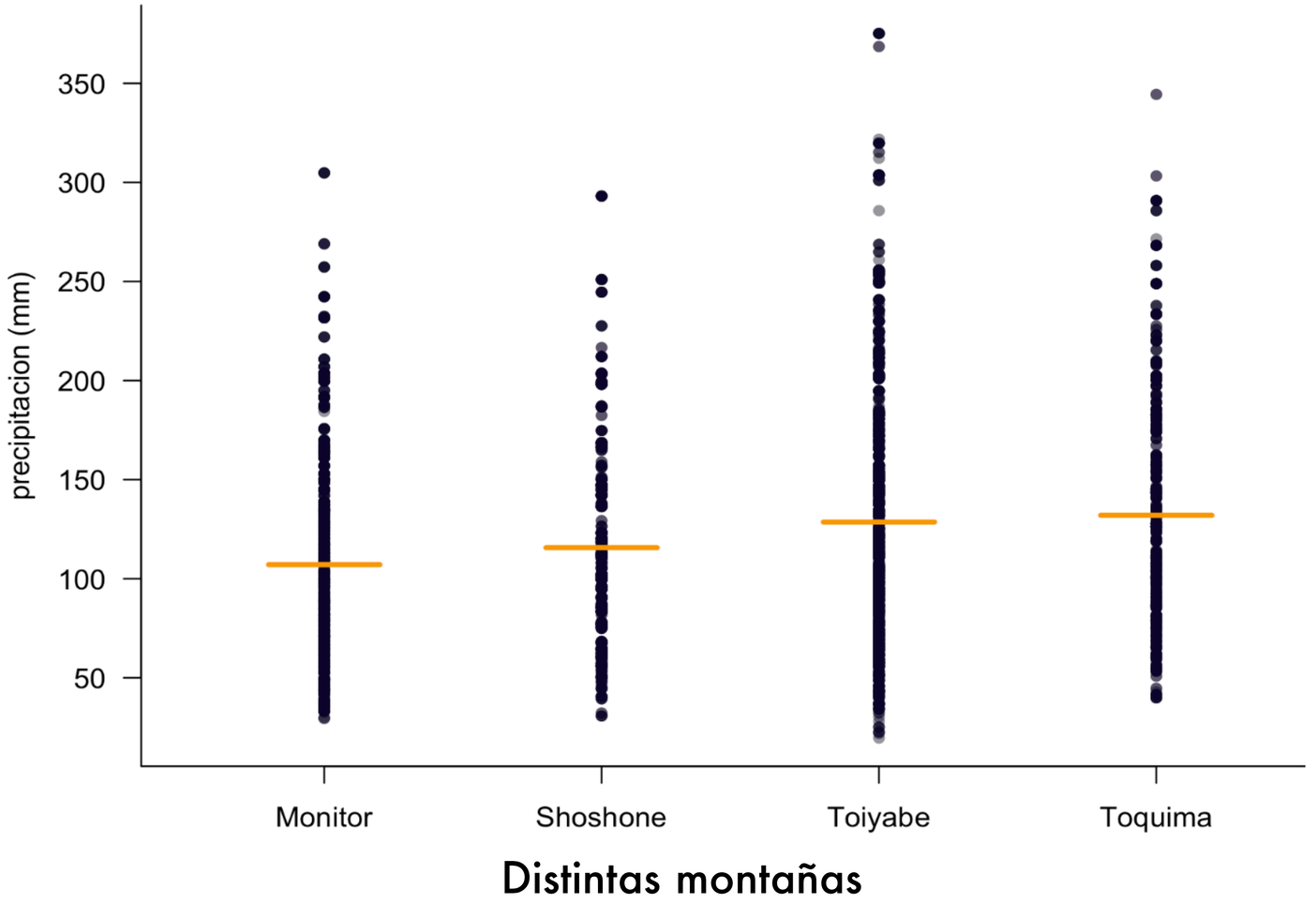
ANOVA


precipitacion (mm)
350
300
250
200
150
100
50
Monitor
Shoshone
Toiyabe
Toquima
Distintas montañas

Qué caracteriza esta ejemplo?

variable repuesta continua

variable predictora categorica



Supuestos: identicos 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

practica en R

Para usar variables categóricas o multiples variables continuas, es necesario entender la matemática.

$$y_i = \alpha + \beta^{\mathrm{T}}\ \mathbf{x_i} + \epsilon_i$$

$$\beta = \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_k \end{pmatrix};\ x_i = \begin{pmatrix} x_{i,1} \\ \vdots \\ x_{i,k} \end{pmatrix}$$

$$\beta^{\mathrm{T}} = \begin{pmatrix} \beta_1 & \dots & \beta_k \end{pmatrix}$$

Como programamos esto?

$$\beta^T = \begin{pmatrix} \beta_1 & \beta_2 & \beta_3 \end{pmatrix}$$

Transformamos las observaciones para clasificarlas en 0 o 1:

$$x_i = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}$$

$$\beta^T x_i = 0 + \beta_2 + 0$$

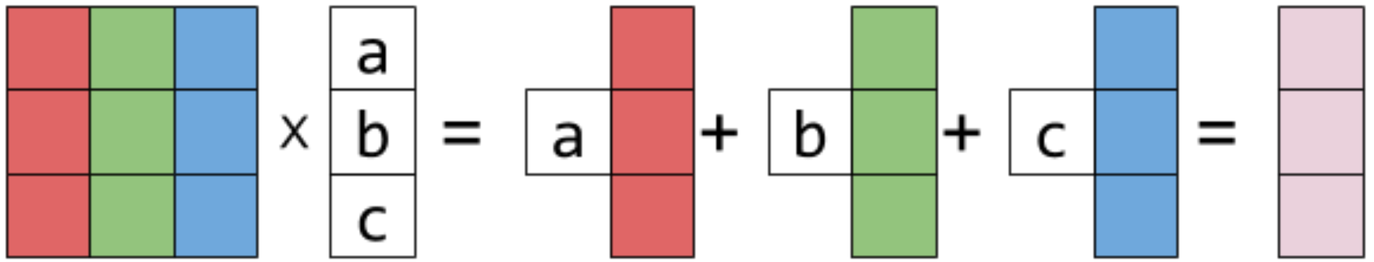
R te lo hace:

```
model.matrix( ~ discrete_predictor)
```

```
##       (Intercept) mountain_rangeShoshone mountain_rangeToiyabe
## [1,]            1                       0                      0
## [2,]            1                       1                      0
## [3,]            1                       0                      0
## [4,]            1                       0                      0
## [5,]            1                       0                      1
##       mountain_rangeToquima
## [1,]                      0
## [2,]                      0
## [3,]                      0
## [4,]                      1
## [5,]                      0
```

Con mas variables:

$$\beta^T = \begin{pmatrix} \beta_1 & \beta_2 & \beta_3 \end{pmatrix}$$



$$x_i = \begin{pmatrix} x_{i,1} \\ x_{i,2} \\ x_{i,3} \end{pmatrix}$$

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

Con mas variables:

```
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
```

Es necesaria pensar en la escala de los variables para que puedas comparar el tamaño del efecto.

```
# 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

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

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

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

Multiples variables predictoras

Suposiciones son lo mismo, ademas suponemos que los variables son relativamente independiente.

Si no lo son, lo llamamos "multicollinearity".

Es difícil para el modelo de distinguir entre los efectos de cada variable.

Multiples variables predictoras

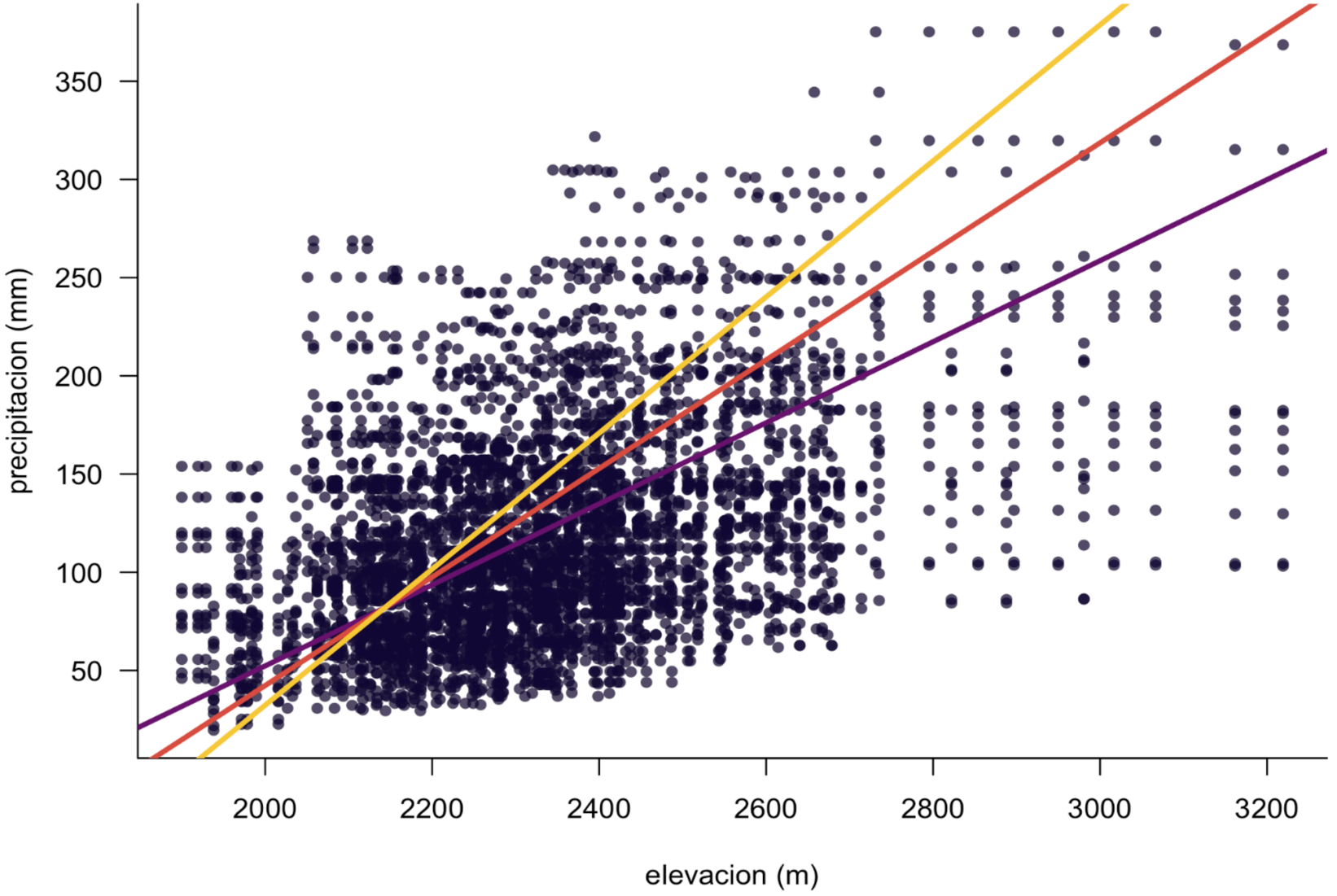
Suposiciones son lo mismo, ademas suponemos que los variables son relativamente independiente.

Si no lo son, lo llamamos "multicollinearity".

Es difícil para el modelo de distinguir entre los efectos de cada variable.

Se arregla por elegir variables con cuidado, y sacar variables colineales.

Interacciones



practica en R

Interacciones

- Es dificil 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 dificil interpretar coeficientes
- El efecto de cada parameter depende del valor estimado de otros parameters

```r
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
##
```

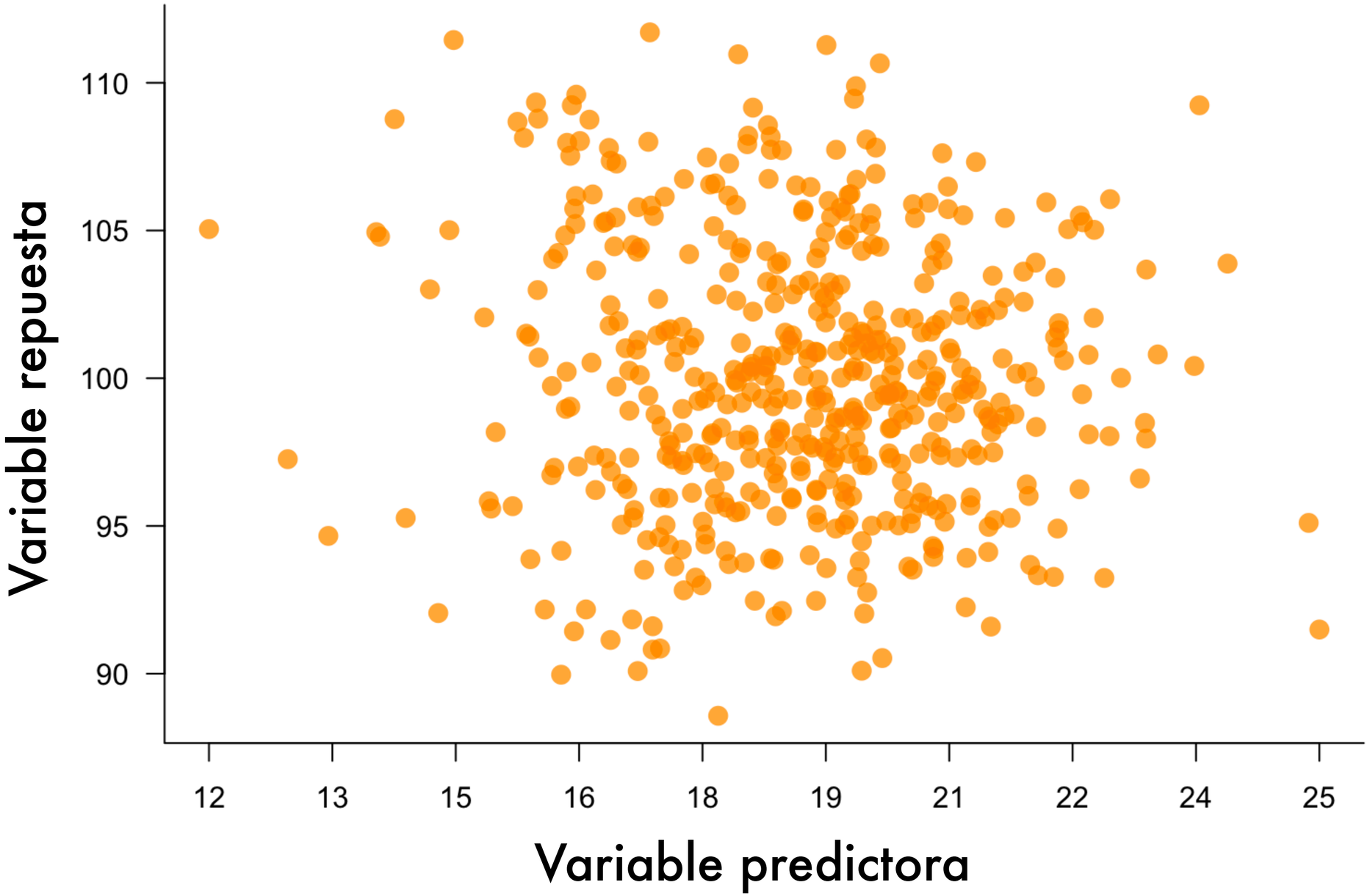
Interacciones

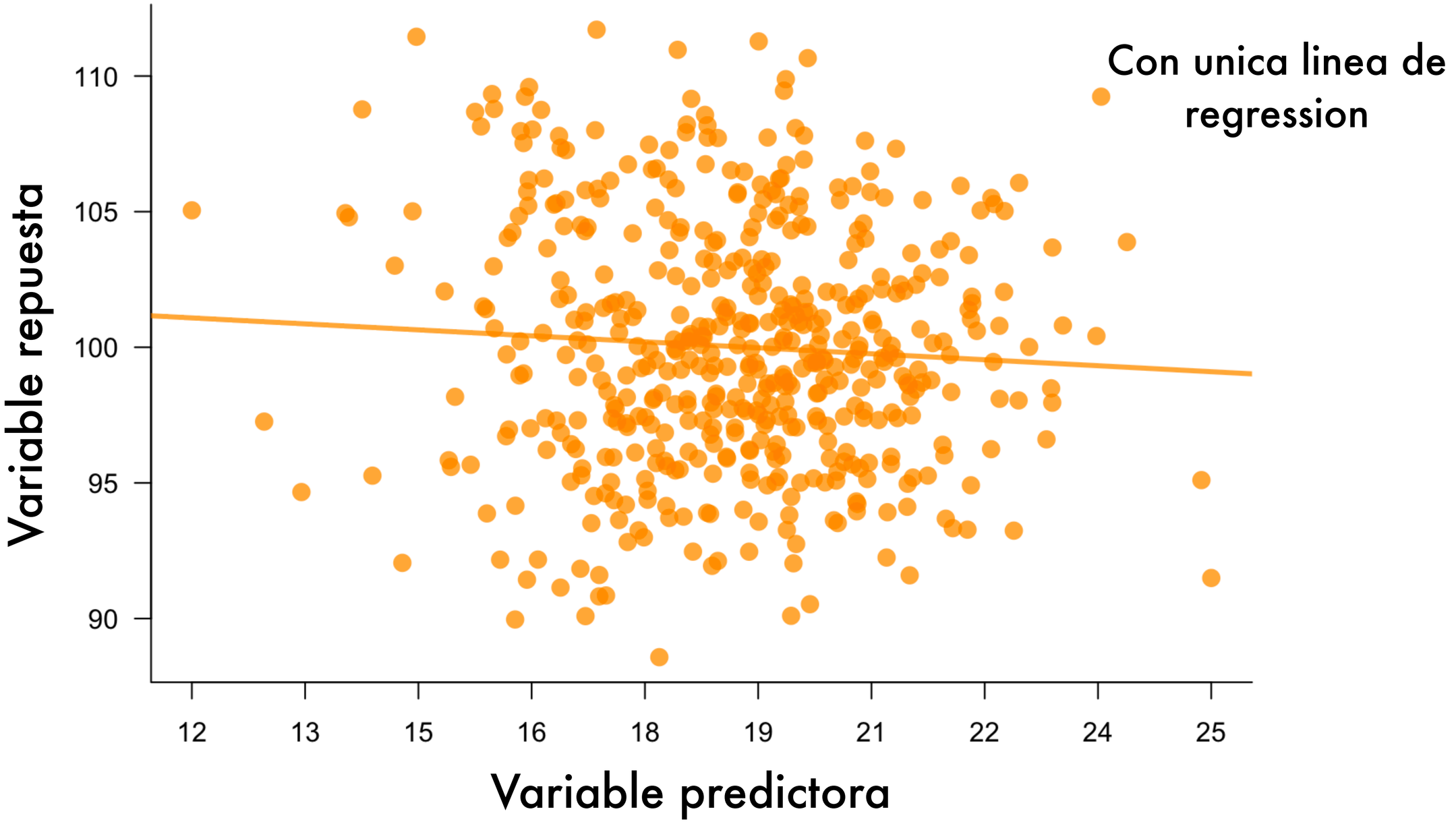
- Es dificil interpretar coeficientes
- El efecto de cada parameter depende del valor estimado de otros parameters
- Particularmente difícil si los dos variables son continuos

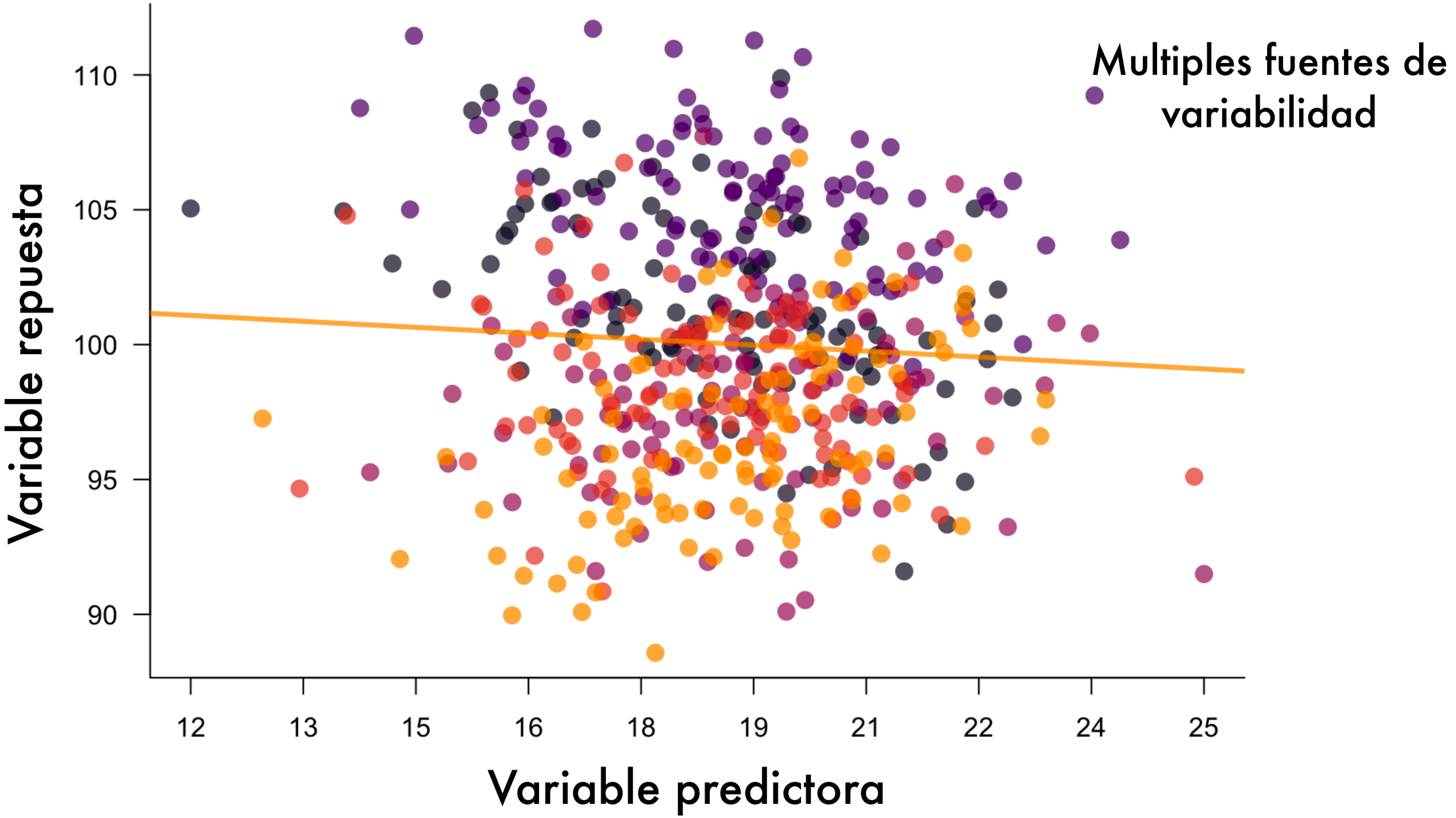
```r
mod <- lm(precipitation ~ elevation * mountain_range)
summary(mod)
```

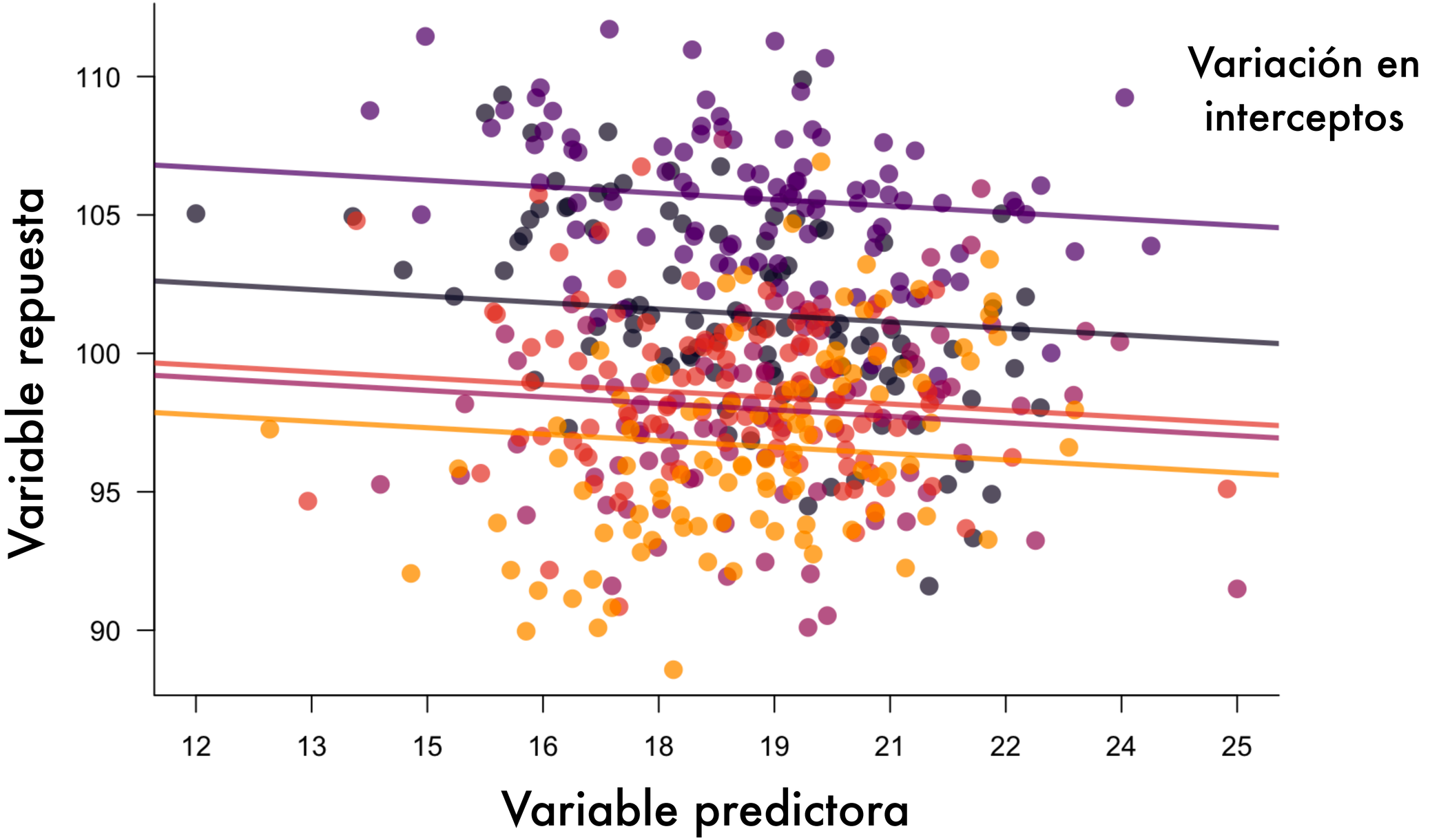
```
##
## Call:
## lm(formula = precipitation ~ elevation * mountain_range)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -120.75  -35.97   -9.48   30.03  202.73
##
## Coefficients:
##                                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     -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
##
```

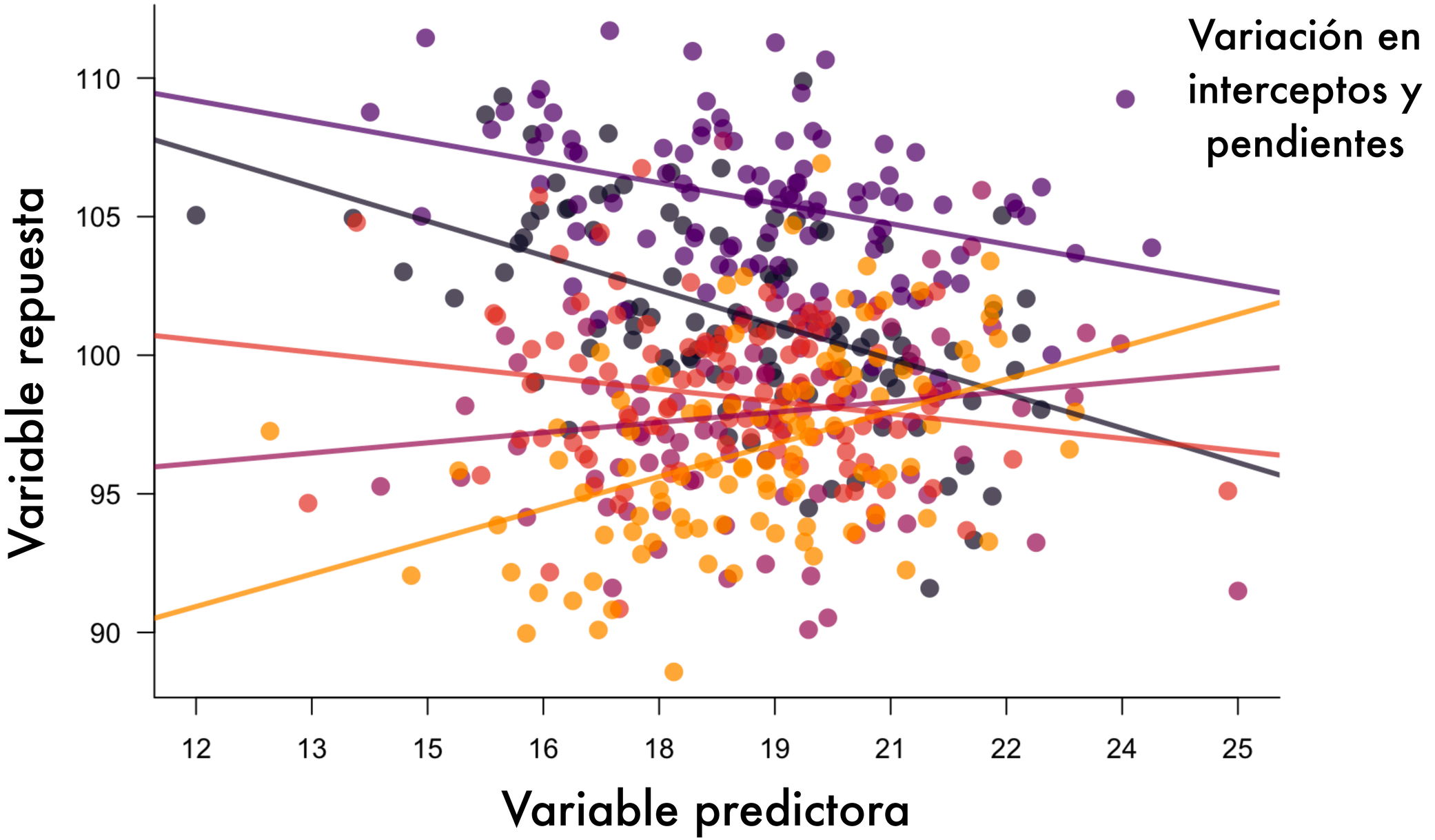
III. Modelos mixtos


Variable repuesta
110
105
100
95
90
12
13
15
16
18
19
21
22
24
25
Variable predictora

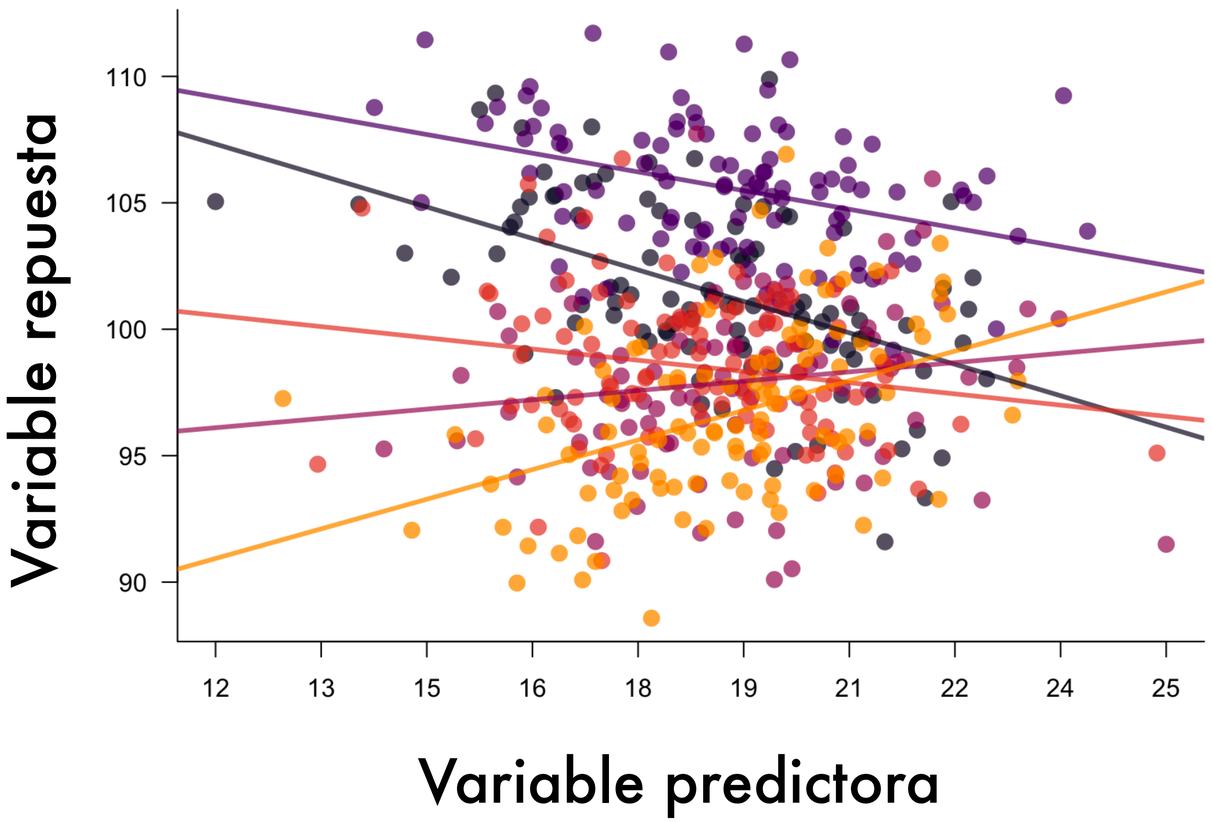

Con unica linea de regression
Variable repuesta
110
105
100
95
90
12
13
15
16
18
19
21
22
24
25
Variable predictora


Multiples fuentes de variabilidad
Variable repuesta
110
105
100
95
90
12
13
15
16
18
19
21
22
24
25
Variable predictora


Variación en interceptos
Variable repuesta
110
105
100
95
90
12
13
15
16
18
19
21
22
24
25
Variable predictora

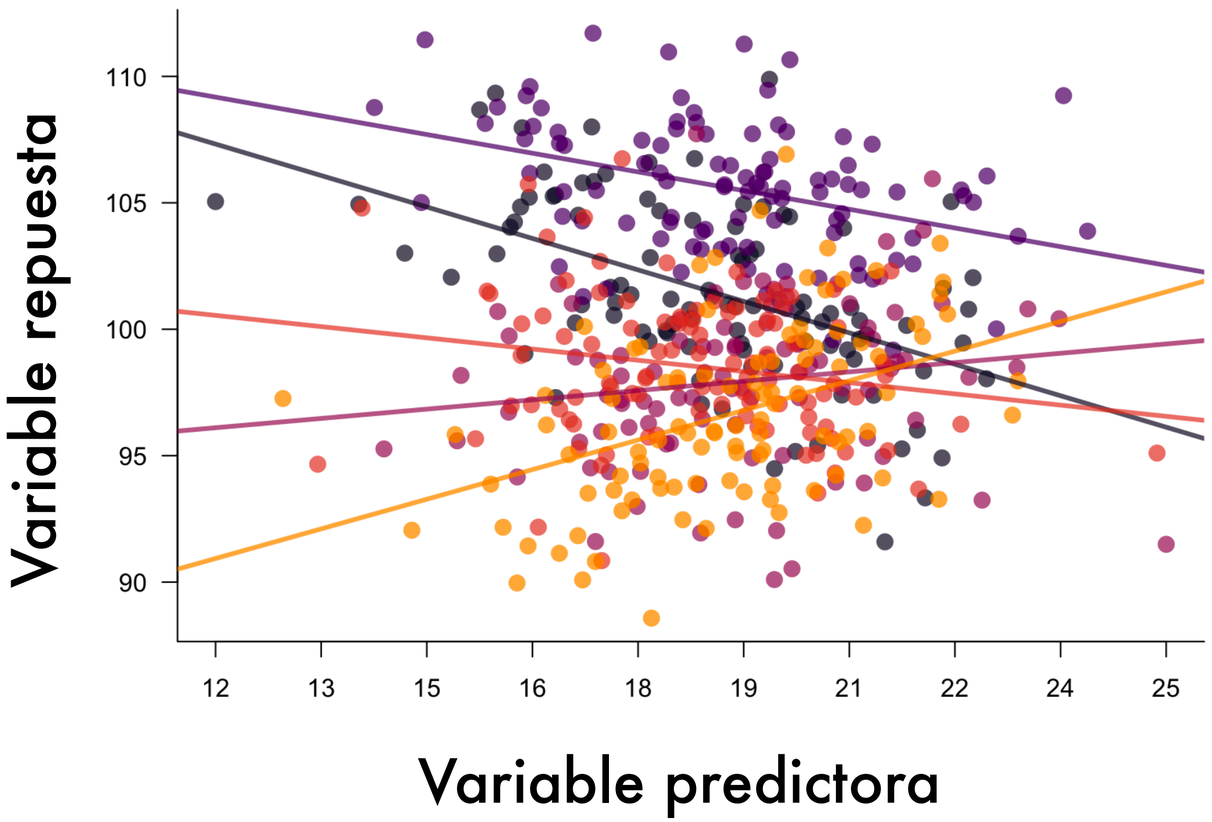

Variación en interceptos y pendientes
Variable repuesta
110
105
100
95
90
12
13
15
16
18
19
21
22
24
25
Variable predictora

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



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

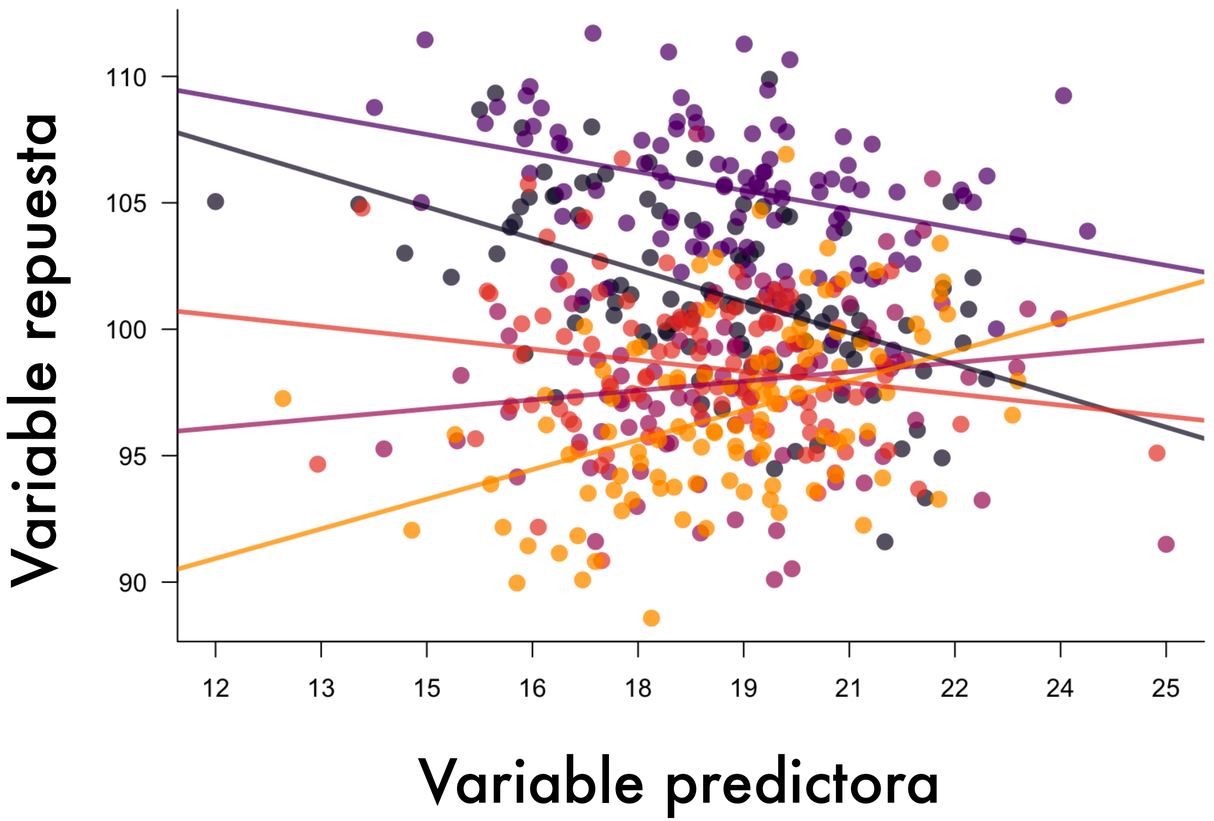
- Con variables categóricas, los grupos varian en su relación con la variable repuesta en una manera *que nos interesa* y *que sean independientes*.



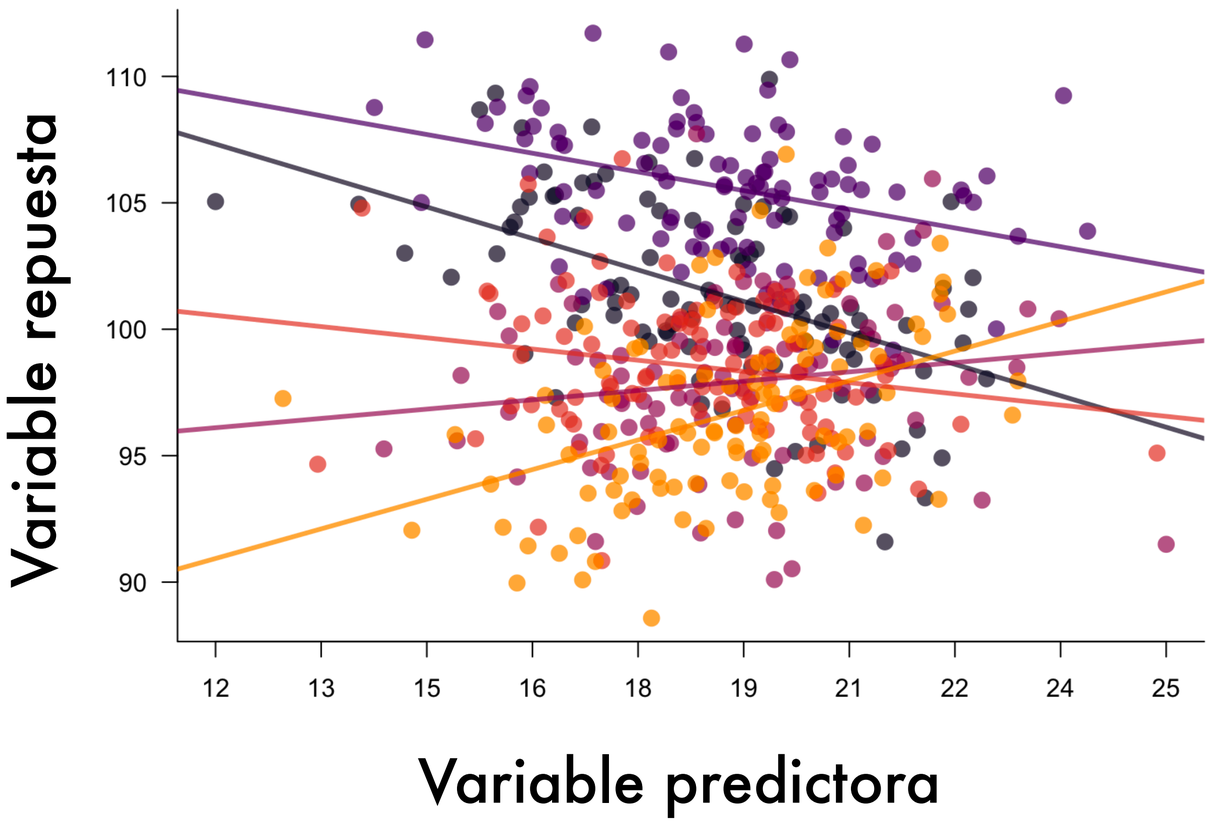
- Como vimos con variables categóricas, grupos pueden variar en su relación con la variable repuesta.

- Con variables categóricas, los grupos varian en su relación con la variable repuesta en una manera *que nos interesa* y *que sean independientes*.

- Pero a veces la variabilidad entre grupos no es informativa, y no son independiente.

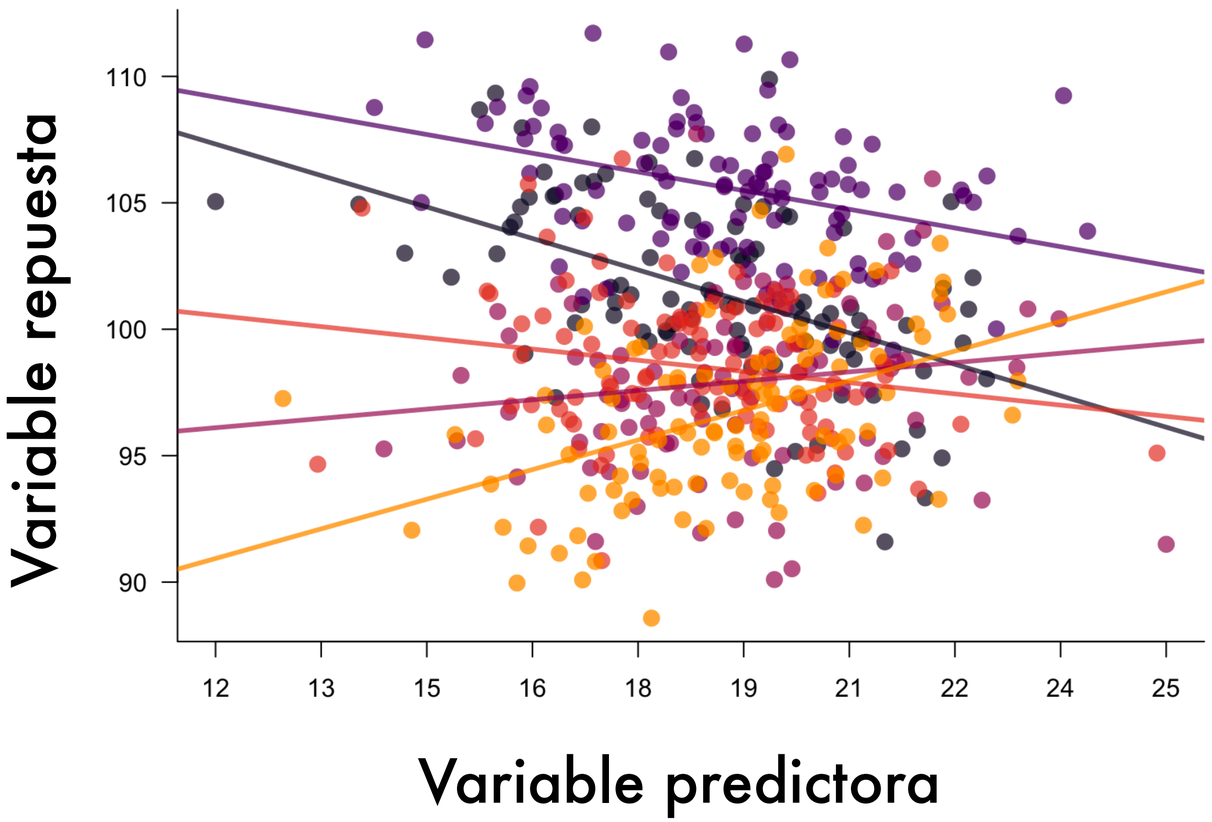


- Este tipo de variabilidad presente un problema porque puede introducir ruido al modelo o romper el suposición de independencia.



- Este tipo de variabilidad presente un problema porque puede introducir ruido al modelo o romper el suposición de independencia.

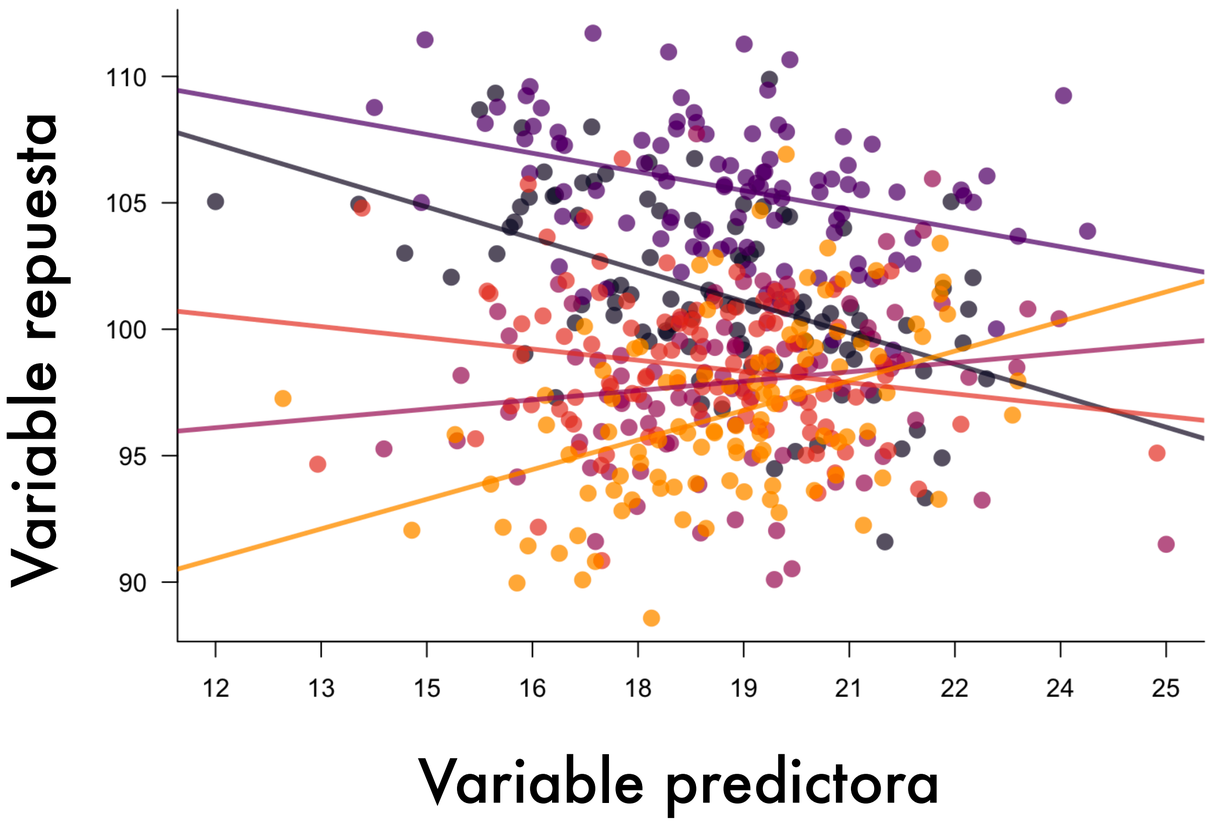
- En contraste al efectos fijos, este tipo de variabilidad se llama efectos aleatorios.



- Este tipo de variabilidad presente un problema porque puede introducir ruido al modelo o romper el suposición de independencia.

- En contraste al efectos fijos, este tipo de variabilidad se llama efectos aleatorios.

- Por ejemplo, muestreos replicados o bloques espacial.



En R usando paquete *lme4*

En R usando paquete *lme4*

```
# Cargar paquete lme4
library(lme4)

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

En R usando paquete *lme4*

```r
# 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))
```

En R usando paquete *lme4*

```r
# 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))

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

Modelos mixtos

- Normalmente no interpretamos los efectos aleatorios. Si quieres interpretarlo, puede que sea mejor como efecto fijo.

- Recomendamos decidir sus efectos aleatorios a priori por dependen el diseño experimental y no de la hipótesis.

- Aunque normalmente no interpretamos los valores del efectos aleatories, puede ser util examinar componentes de variabilidad.

Modelo conceptual

Formar la pregunta / escribir el modelo

Diseño experimental

Colección de datos

Armar el modelo

Creer resultados

Modelo conceptual

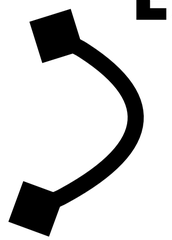
Formar la pregunta

Diseño experimental

Colección de datos

Armar el modelo

Creer resultados

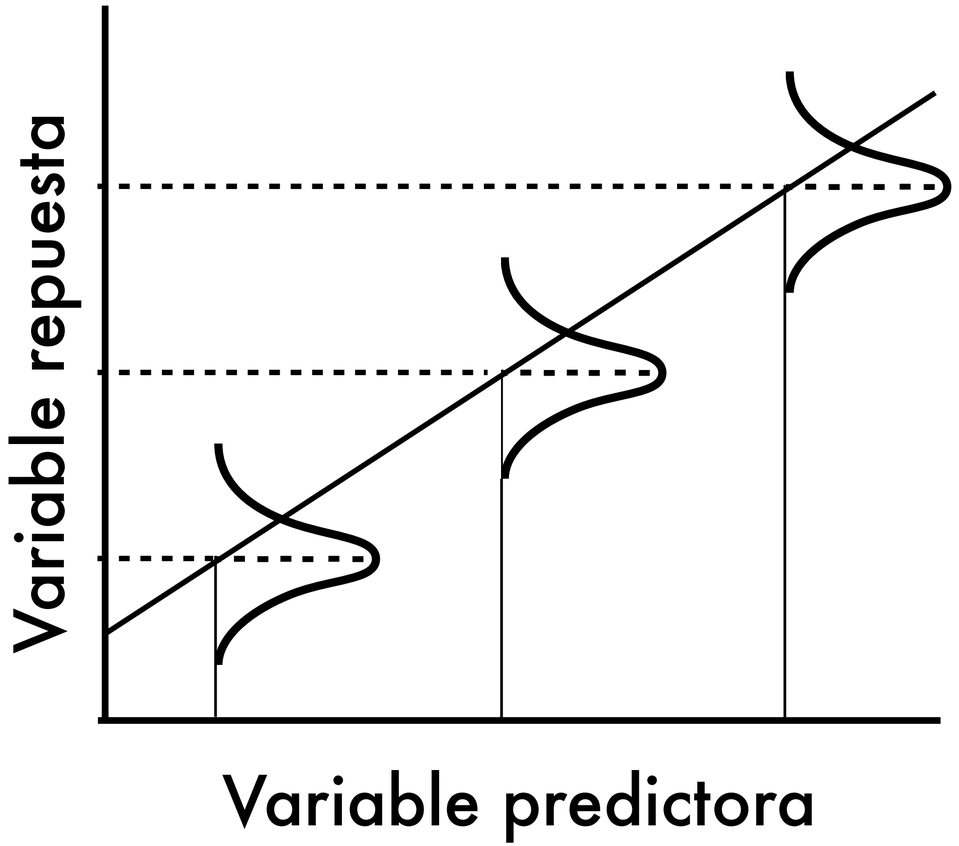


Escribir el modelo

IV. Modelos lineales generalizados

Recuerden suposición 2 de modelos lineales:

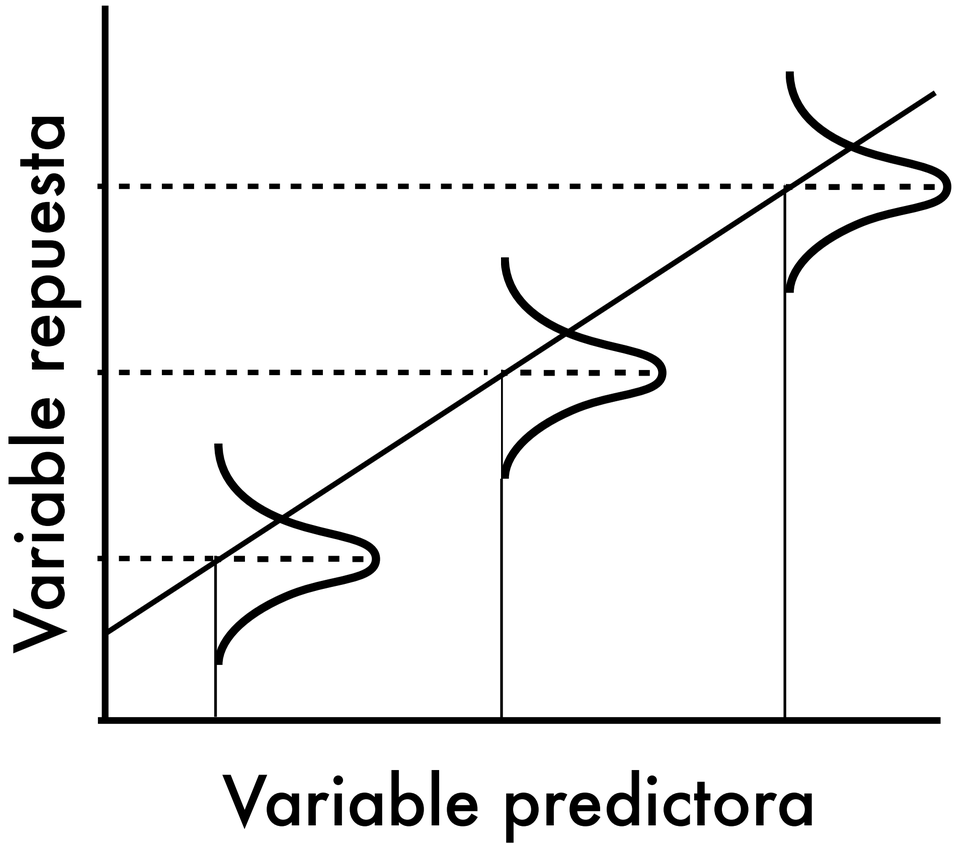
Residuales son con distribución normal



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

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

El matemática también puede escribir asi:



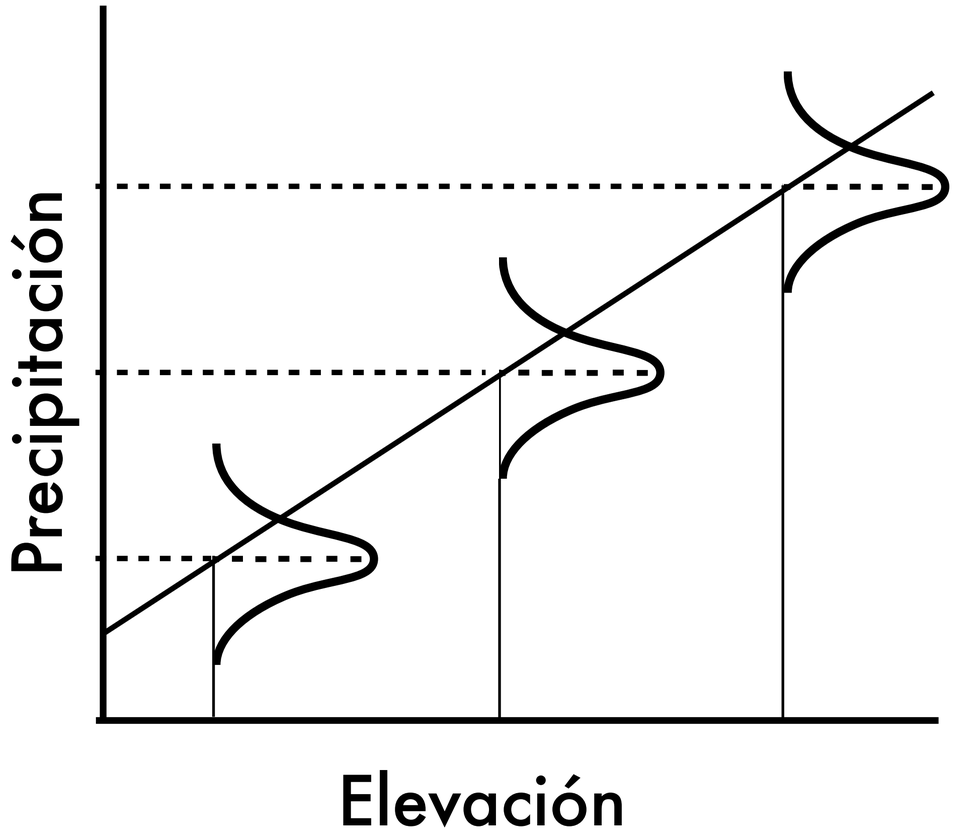
$$y_i = \alpha + \beta * x_i + \epsilon_i$$

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

==

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

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



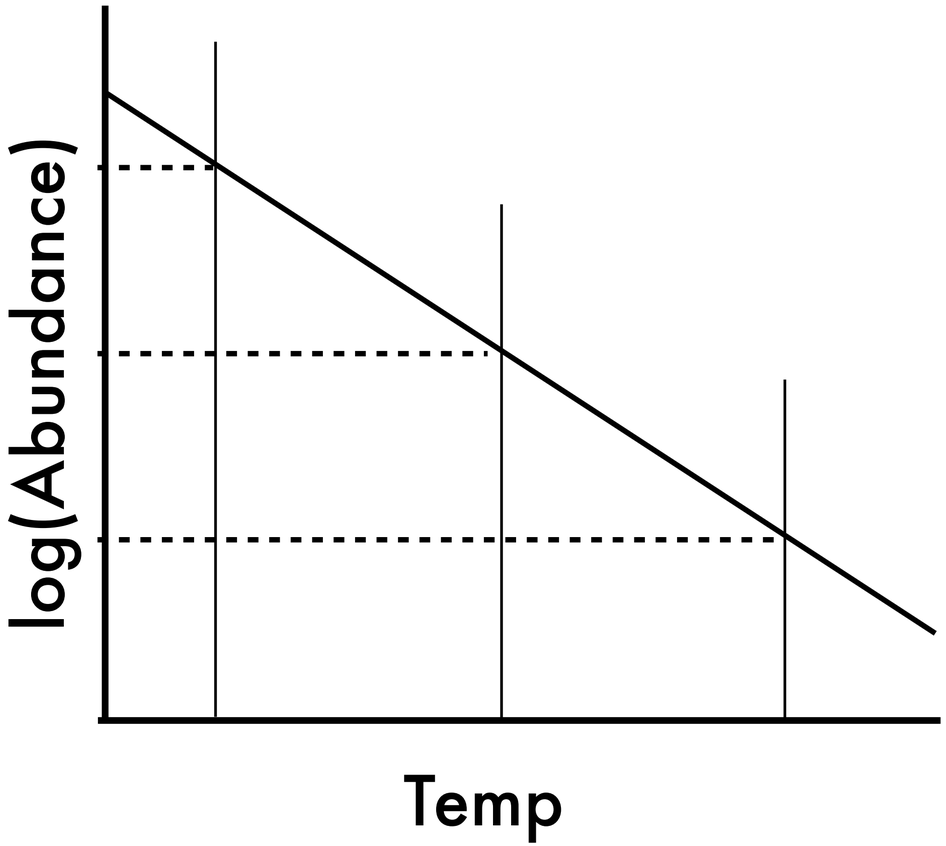
$$Y = \alpha + \beta * X + \epsilon$$

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

$$Y \sim N(\mu_t, \sigma^2)$$

$$\mu_t = \alpha + \beta * X$$

How does temperature of hollow affect ringtail possum abundance?



$$A \sim Poisson(\lambda)$$

$$log(\lambda) = \alpha + \beta * temp$$

Exercise 2: Write a data generative linear model

Simulating data

Model

Model-defined data

Model

Model-defined data

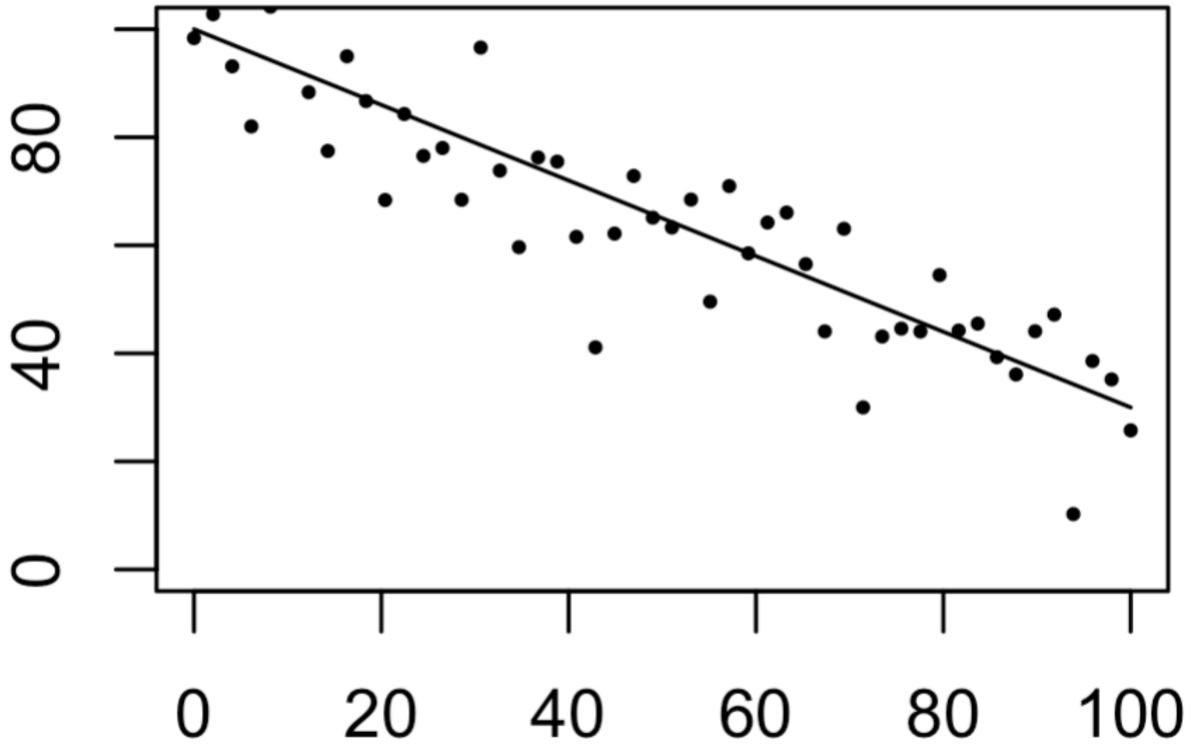
$$M_t \sim N(\mu_t, \sigma^2)$$

$$\mu_t = \alpha + \beta * time$$

R practical: simulating data

$$y = a + bx + \epsilon$$

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



What's it all about?

$$y = a + bx + \epsilon$$

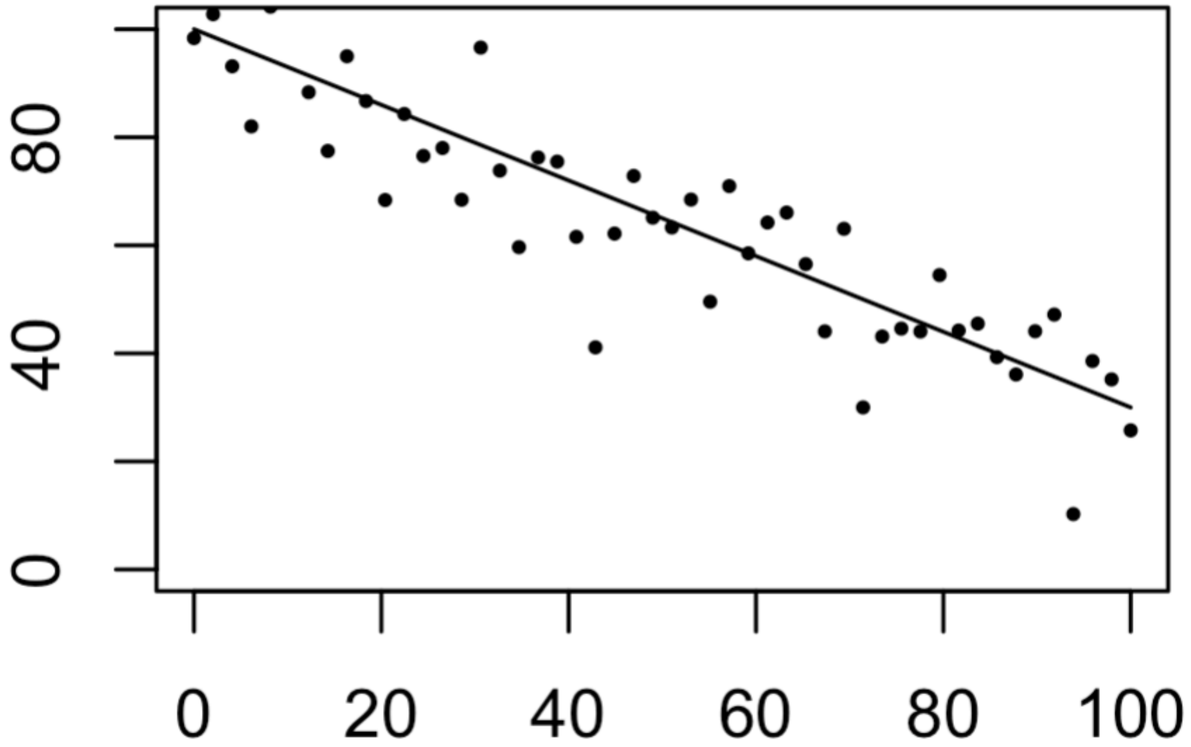
$$\epsilon \sim N(0, \sigma^2)$$

MLE

$$b = -0.67$$

Posterior mean

$$b = -0.67$$



Why Bayes

Flexibility in building models > can test any hypothesis, not just the null

Principled way of building your model > think about what you know & the implications of what you think you know

Your inferences will fail often, giving you a way to diagnose the issue

Scales up, same tools and workflow for range of model processes > from t test to bayesian network analysis

More subjective > you make more choices.

Statistical models exist independent of the method of estimating parameters

What's it all about?

Statistical models exist independent of the method of estimating parameters

There are no Frequentist models or Bayesian models

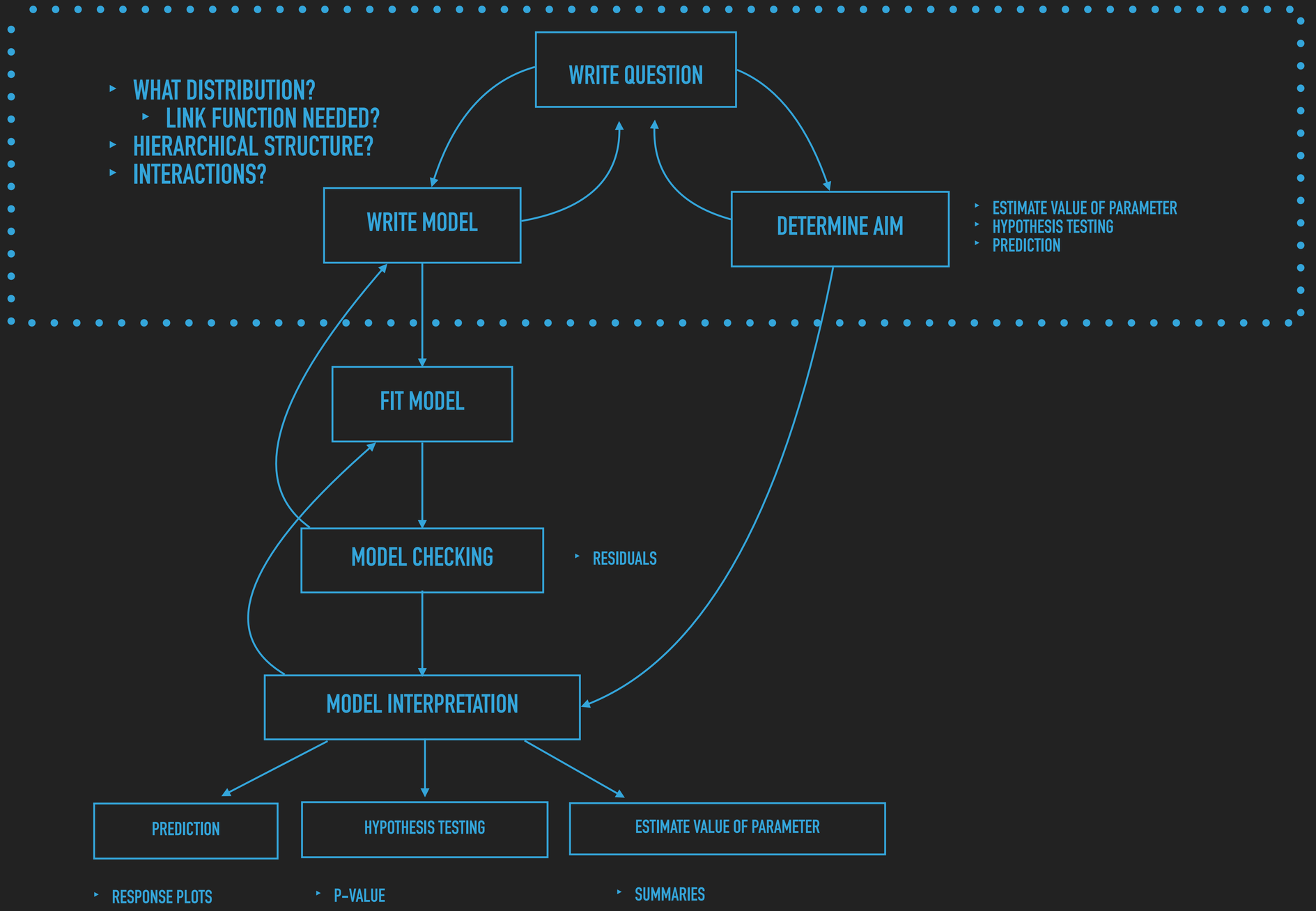
What's it all about?

Statistical models exist independent of the method of estimating parameters

There are no Frequentist models or Bayesian models

May choose to analyse a model in a Bayesian way

I. Thinking about models


WRITE QUESTION
WHAT DISTRIBUTION?
LINK FUNCTION NEEDED?
HIERARCHICAL STRUCTURE?
INTERACTIONS?
WRITE MODEL
DETERMINE AIM
ESTIMATE VALUE OF PARAMETER
HYPOTHESIS TESTING
PREDICTION
FIT MODEL
MODEL CHECKING
RESIDUALS
MODEL INTERPRETATION
PREDICTION
HYPOTHESIS TESTING
ESTIMATE VALUE OF PARAMETER
RESPONSE PLOTS
P-VALUE
SUMMARIES

Exercise 1: Research workflow

Conceptual model

Prepare question

Experimental design

Data collection

Write and build model

Create model outputs

Conceptual model

Prepare question

Experimental design

Data collection

Write and build model

Create model outputs

Thinking about models

== Using a model to understand our data.

Conceptual model

Prepare question

Experimental design

Data collection

Write and build model

Create model outputs

Conceptual model

Prepare question / Write model

Experimental design

Data collection

Build model

Create model outputs

Conceptual model

Prepare question / Write model

Experimental design

Data collection

Build model

Create model outputs

We want to collect data to test our model.

Using a model to understand our data is asking:

"What is the probability of this data, given a certain model?"

Using a model to understand our data is asking:

"What is the probability of this data, given a certain model?"

Collecting data to test our model is asking:

"What is the probability of this model, given observed data?"

A well-specified research question defines a statistical model.

AND

A statistical model underpins a research question.

Thinking about models

Model

Model-defined
data

Thinking about models

Model

Model-defined data

?=

Experiment

Collected data

VI. Model checking

Addressing your question

Estimating value of parameter:
Forest plot, summarise, mean and sd of posterior of parameter.

Hypothesis testing:
Bayesian p-value. “what is the posterior probability that X has an effect on Y.”

Prediction:
calculate, effect/response plot to new data. Cross-validation - testing on new data.

Review

Special thanks to

Nick Golding
Marc Kery
Gerry Ryan
Richard McElreath

III. Thinking about probability

Using a model to understand our data is asking:

"What is the probability of this data, given a certain model?"

Using a model to understand our data is asking:

"What is the probability of this data, given a certain model?"

p(data | parameter)

Thinking about probability

Using a model to understand our data is asking:

"What is the probability of this data, given a certain model?"

p(data | parameter)

Collecting data to test our model is asking:

"What is the probability of this model, given observed data?"

Thinking about probability

Using a model to understand our data is asking:

"What is the probability of this data, given a certain model?"

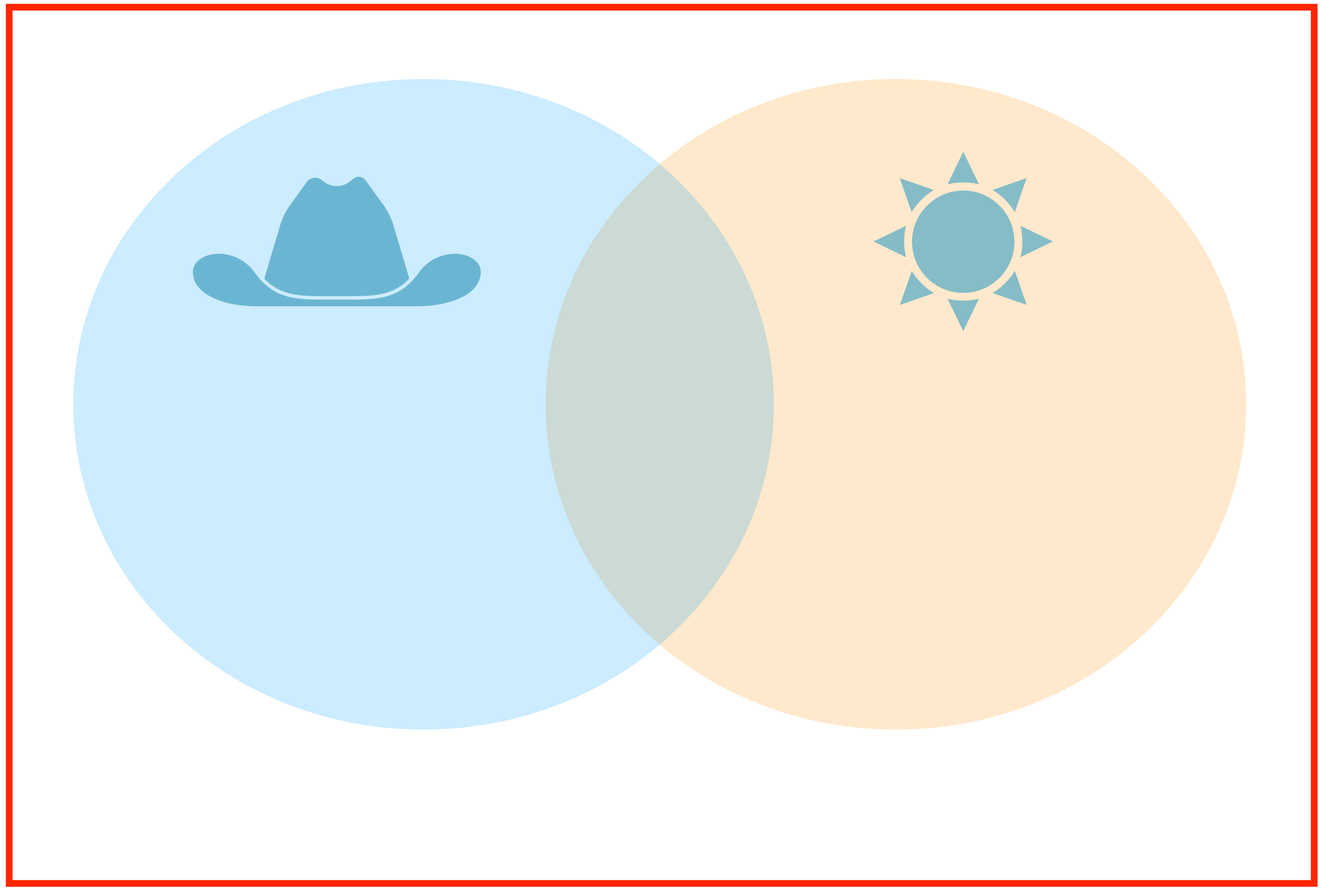
p(data | parameter)

Collecting data to test our model is asking:

"What is the probability of this model, given observed data?"

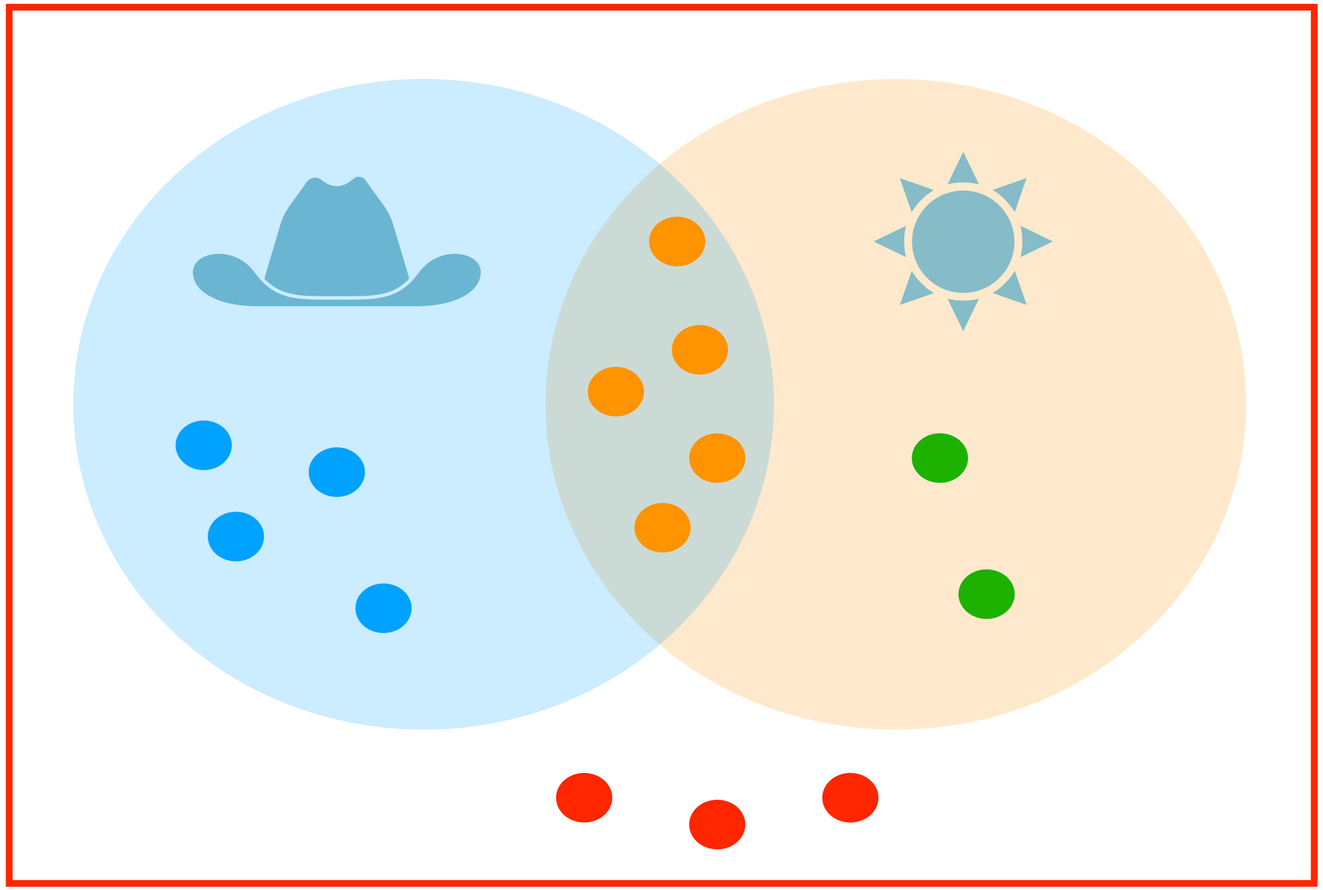
p(parameter | data)

Thinking about probability

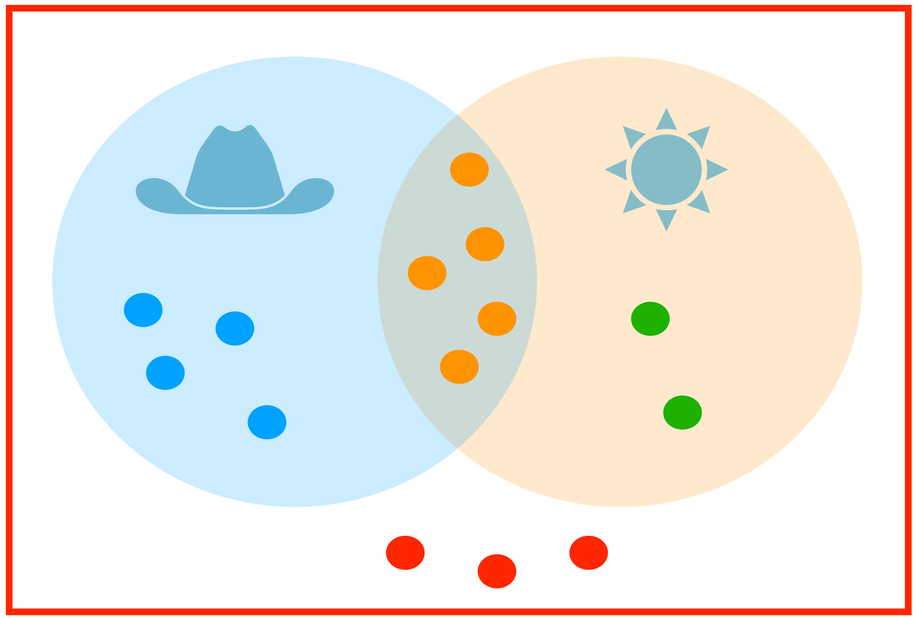


Check out: https://www.youtube.com/watch?v=9wCnvr7Xw4E&ab_channel=StatQuestwithJoshStarmer

Thinking about probability

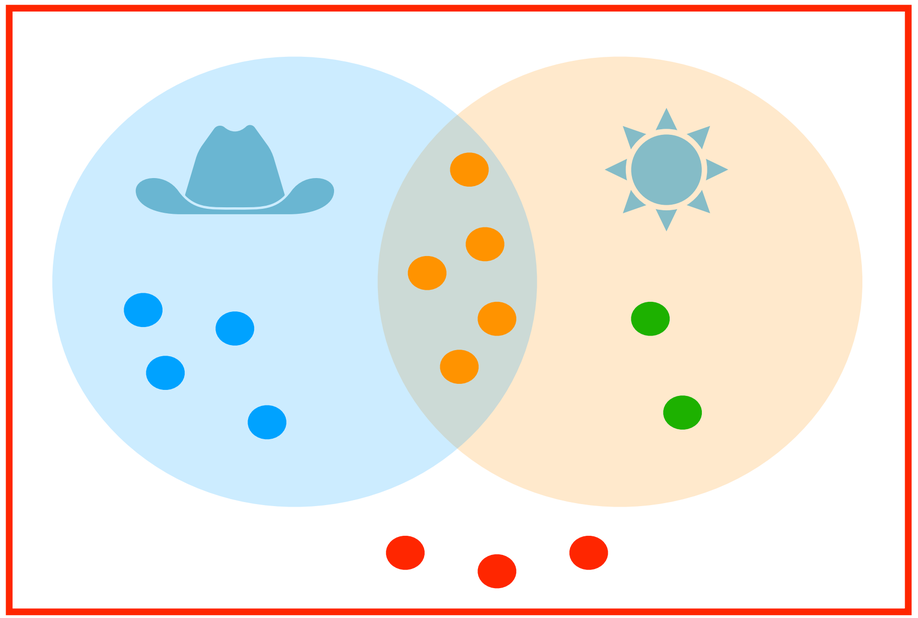


Thinking about probability



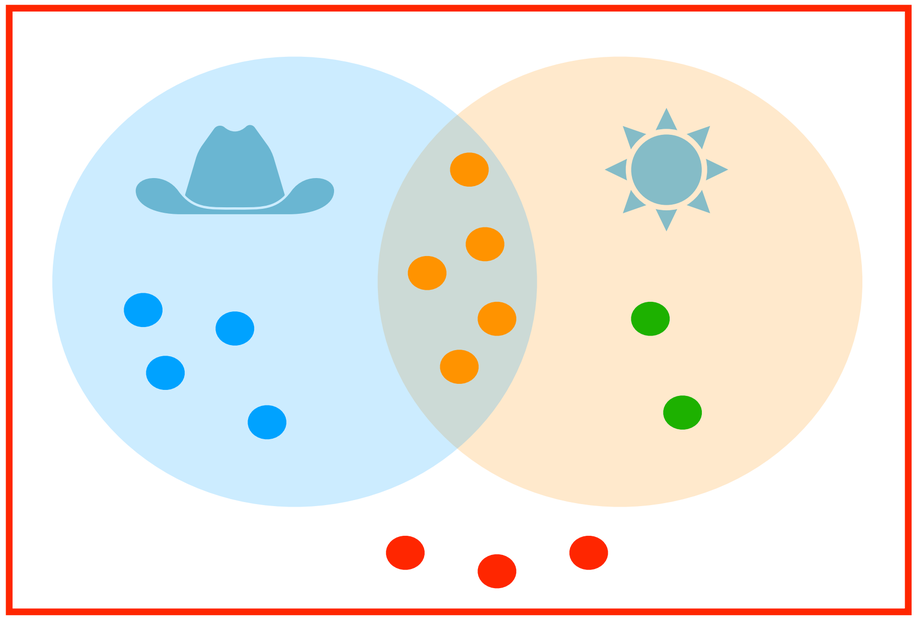
	No sun	Sun	Total
Wore hat	4/14		
Did not wear hat			
Total			

Thinking about probability



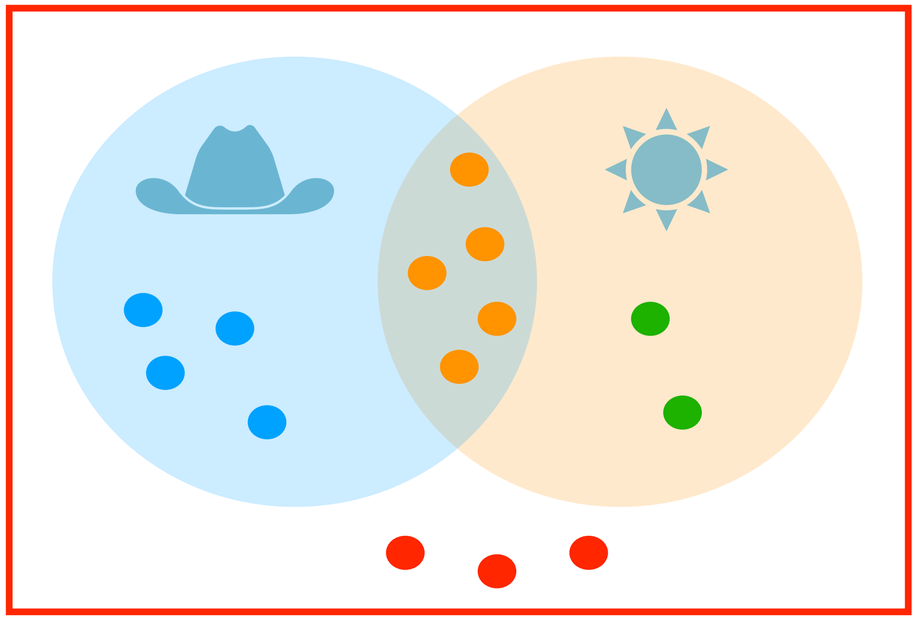
	No sun	Sun	Total
Wore hat	4/14	5/14	9/14
Did not wear hat			
Total			

Thinking about probability



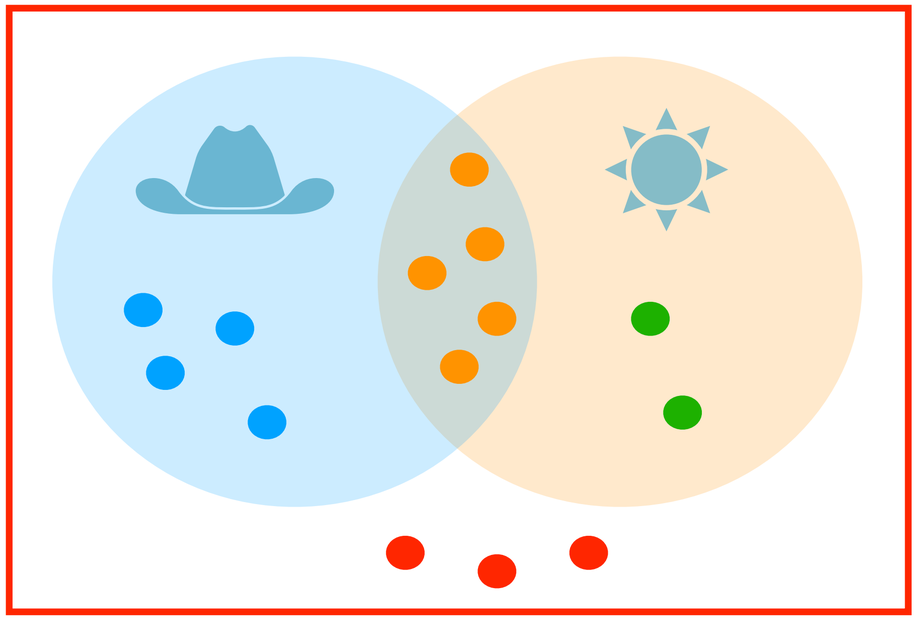
	No sun	Sun	Total
Wore hat	4/14	5/14	9/14
Did not wear hat	3/14		
Total			

Thinking about probability



	No sun	Sun	Total
Wore hat	4/14	5/14	9/14
Did not wear hat	3/14	2/14	5/14
Total	7/14	7/14	

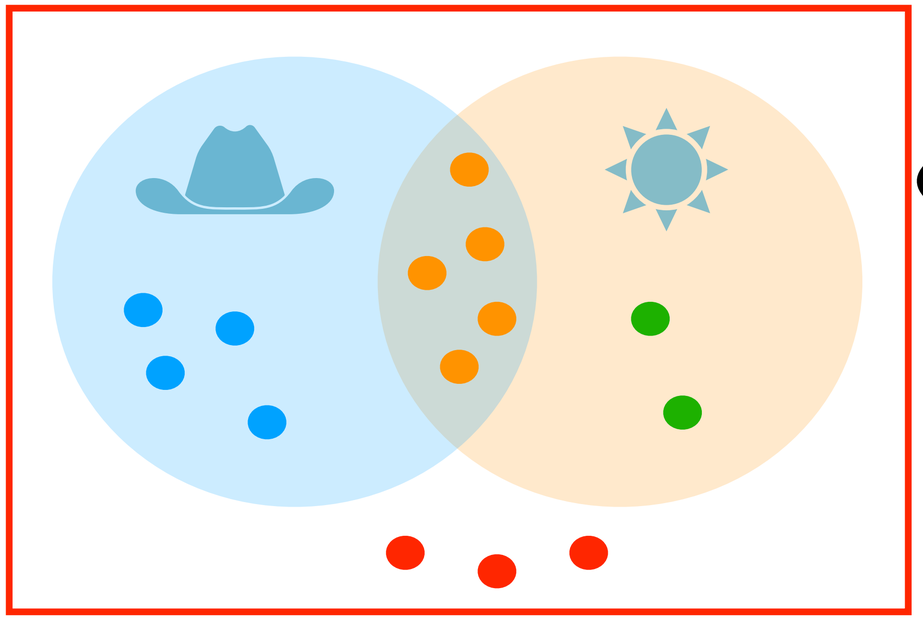
Thinking about probability



	No sun	Sun	Total
Wore hat	4/14	5/14	9/14
Did not wear hat	3/14	2/14	5/14
Total	7/14	7/14	

p(sun & hat | hat) =

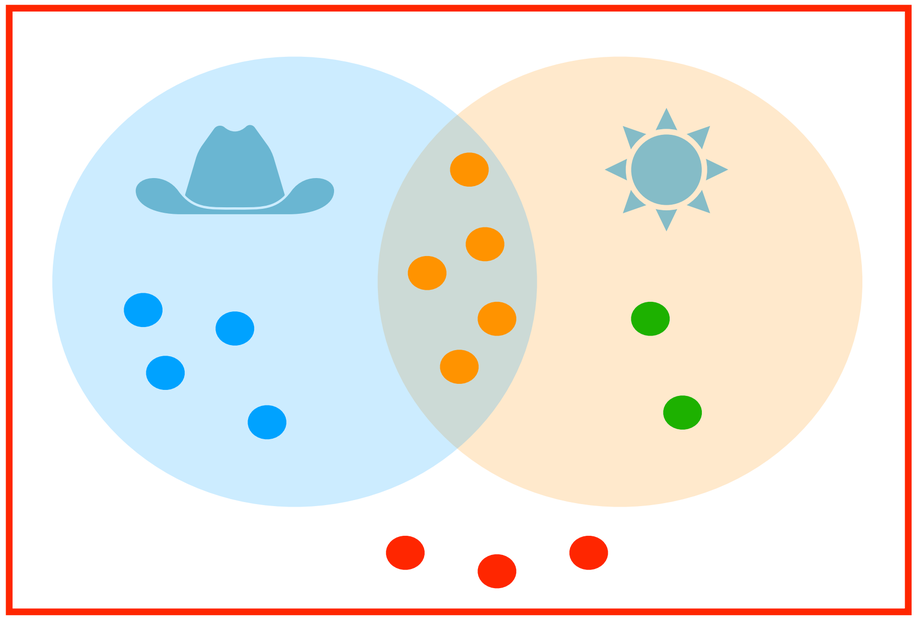
Thinking about probability



	No sun	Sun	Total
Wore hat	4/14	5/14	9/14
Did not wear hat	3/14	2/14	5/14
Total	7/14	7/14	

$$p(\text{sun \& hat} \mid \text{hat}) = \frac{\qquad\qquad}{p(\text{hat})}$$

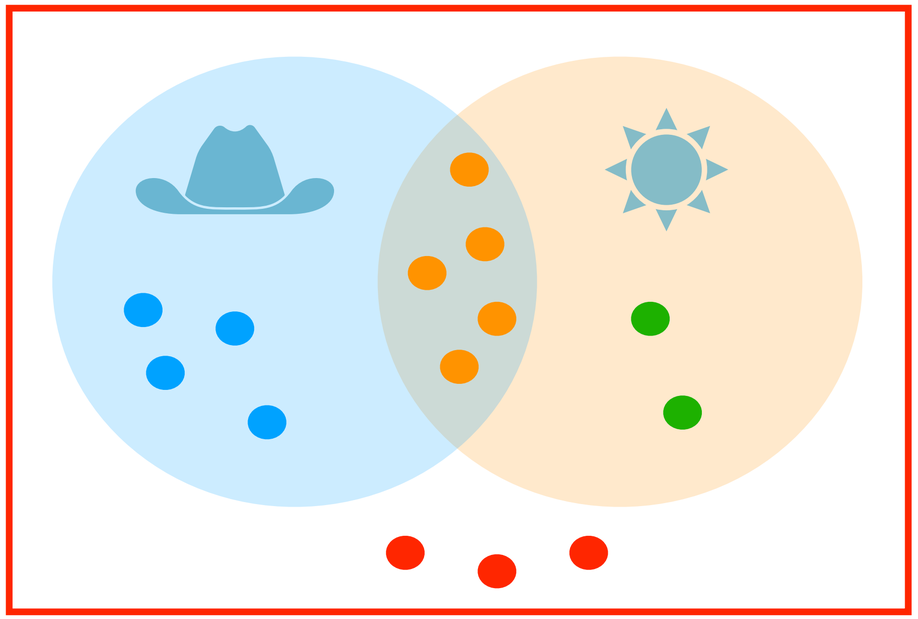
Thinking about probability



	No sun	Sun	Total
Wore hat	4/14	5/14	9/14
Did not wear hat	3/14	2/14	5/14
Total	7/14	7/14	

$$p(\text{sun \& hat} \mid \text{hat}) = \frac{p(\text{sun \& hat})}{p(\text{hat})}$$

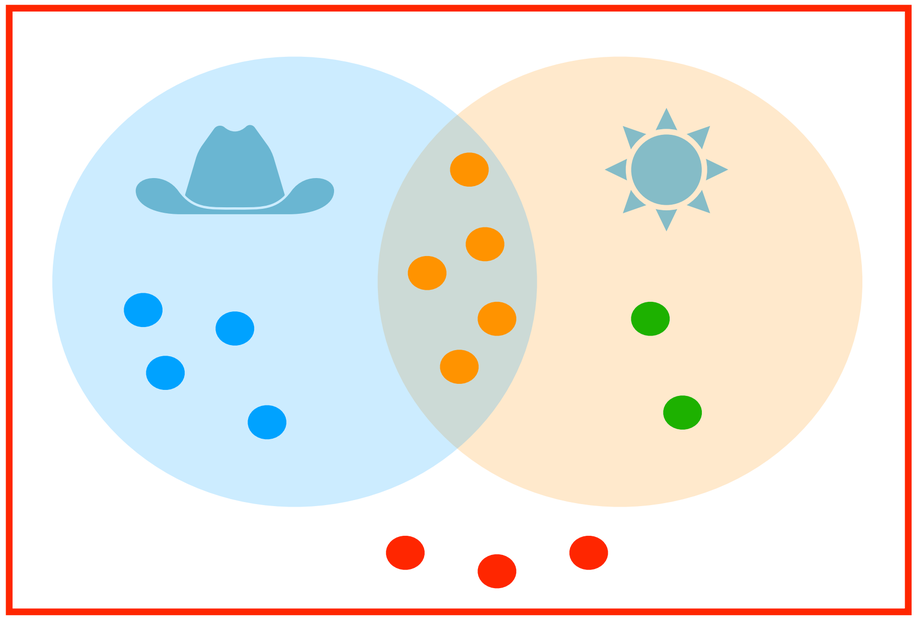
Thinking about probability



	No sun	Sun	Total
Wore hat	4/14	5/14	9/14
Did not wear hat	3/14	2/14	5/14
Total	7/14	7/14	

$$p(\text{sun \& hat} \mid \text{hat}) = \frac{5/14}{9/14} = .55$$

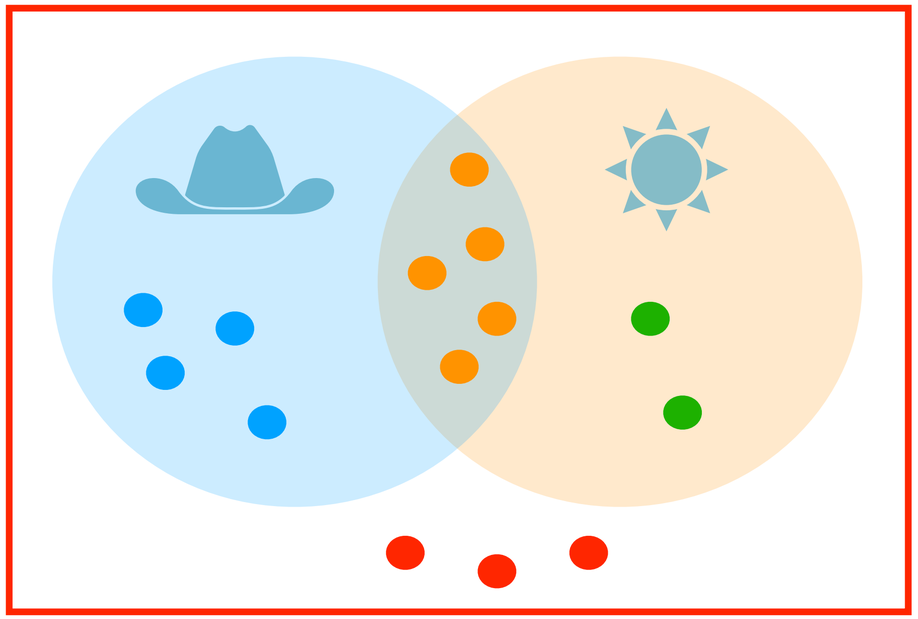
Thinking about probability



	No sun	Sun	Total
Wore hat	4/14	5/14	9/14
Did not wear hat	3/14	2/14	5/14
Total	7/14	7/14	

$$p(\text{sun} \,\&\, \text{hat} \mid \text{sun}) = \frac{\qquad\qquad}{p(\text{sun})}$$

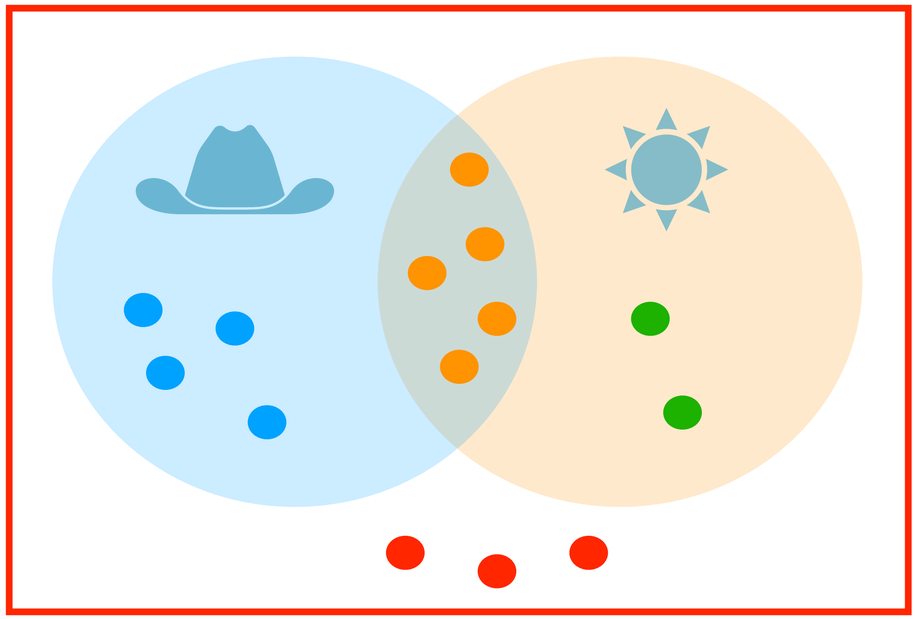
Thinking about probability



	No sun	Sun	Total
Wore hat	4/14	5/14	9/14
Did not wear hat	3/14	2/14	5/14
Total	7/14	7/14	

$$p(\text{sun} \,\&\, \text{hat} \mid \text{sun}) = \frac{p(\text{sun} \,\&\, \text{hat})}{p(\text{sun})}$$

Thinking about probability



	No sun	Sun	Total
Wore hat	4/14	5/14	9/14
Did not wear hat	3/14	2/14	5/14
Total	7/14	7/14	

$$p(\text{sun \& hat} \mid \text{sun}) = \frac{5/14}{7/14} = .71$$

Thinking about probability

$$p(\text{sun \& hat} \mid \text{hat}) = \frac{p(\text{sun \& hat})}{p(\text{hat})}$$

$$p(\text{sun \& hat} \mid \text{sun}) = \frac{p(\text{sun \& hat})}{p(\text{sun})}$$

$$p(\text{sun \& hat} \mid \text{hat}) = \frac{p(\text{sun \& hat})}{p(\text{hat})}$$

$$p(\text{sun \& hat} \mid \text{sun}) = \frac{p(\text{sun \& hat})}{p(\text{sun})}$$

We don't usually know the probabilities of both events,
so we ask, is it possible to estimate the conditional p, without that data?

Thinking about probability

$$p(\text{sun} \,\&\, \text{hat} \mid \text{hat}) * p(\text{hat}) = p(\text{sun} \,\&\, \text{hat})$$

$$p(\text{sun} \,\&\, \text{hat} \mid \text{sun}) * p(\text{sun}) = p(\text{sun} \,\&\, \text{hat})$$

Thinking about probability

$$p(\text{sun} \,\&\, \text{hat} \mid \text{hat}) * p(\text{hat}) = \quad p(\text{sun} \,\&\, \text{hat} \mid \text{sun}) * p(\text{sun})$$

$$p(\text{sun} \,\&\, \text{hat} \mid \text{hat}) * p(\text{hat}) = p(\text{sun} \,\&\, \text{hat} \mid \text{sun}) * p(\text{sun})$$

$$p(\text{sun} \,\&\, \text{hat} \mid \text{hat}) = \frac{p(\text{sun} \,\&\, \text{hat} \mid \text{sun}) * p(\text{sun})}{p(\text{hat})}$$

Thinking about probability

$$p(\text{sun} \,\&\, \text{hat} \mid \text{hat}) * p(\text{hat}) = p(\text{sun} \,\&\, \text{hat} \mid \text{sun}) * p(\text{sun})$$

$$p(\text{sun} \,\&\, \text{hat} \mid \text{hat}) = \frac{p(\text{sun} \,\&\, \text{hat} \mid \text{sun}) * p(\text{sun})}{p(\text{hat})}$$

Thinking about probability

$$p(\text{sun} \,\&\, \text{hat} \mid \text{hat}) = \frac{p(\text{sun} \,\&\, \text{hat} \mid \text{sun}) * p(\text{sun})}{p(\text{hat})}$$

== Bayes Theorem! 🥳🥳🥳

Thinking about probability

$$p(\text{sun} \,\&\, \text{hat} \mid \text{hat}) = \frac{p(\text{sun} \,\&\, \text{hat} \mid \text{sun}) * p(\text{sun})}{p(\text{hat})}$$

Thinking about probability

$$p(\text{sun} \mid \text{hat}) = \frac{p(\text{hat} \mid \text{sun}) * p(\text{sun})}{p(\text{hat})}$$

Thinking about probability

evidence/
data

hypothesis/
parameter

Thinking about probability

$$p(\text{parameter} \mid \text{data}) = \frac{p(\text{data} \mid \text{parameter}) * p(\text{parameter})}{p(\text{data})}$$

Thinking about probability

$$p(\text{parameter} \mid \text{data}) = \frac{p(\text{data} \mid \text{parameter}) * p(\text{parameter})}{p(\text{data})}$$

IV. Bayesian analysis

Bayesian analysis

$$p(\text{parameter} \mid \text{data}) = \frac{p(\text{data} \mid \text{parameter}) * p(\text{parameter})}{p(\text{data})}$$

$$p(\text{parameter} \mid \text{data}) = \frac{\overbrace{p(\text{data} \mid \text{parameter})}^{\textbf{Likelihood}} * p(\text{parameter})}{p(\text{data})}$$

What is the probability of observing this data, given this parameter value?

Posterior **Likelihood**

$$p(\text{parameter} \mid \text{data}) = \frac{p(\text{data} \mid \text{parameter}) * p(\text{parameter})}{p(\text{data})}$$

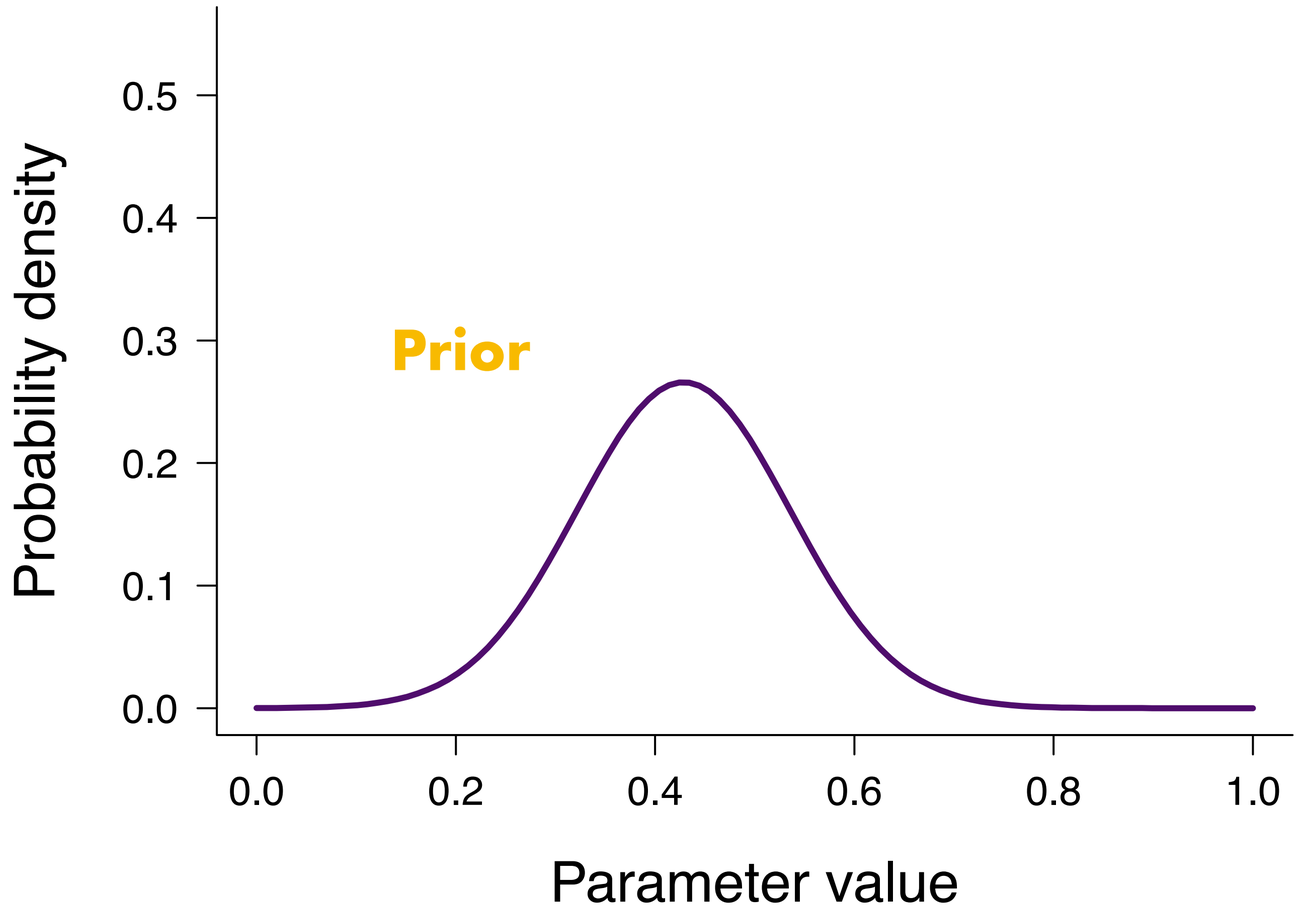
What is the probability of a given parameter value, given the data we have observed?

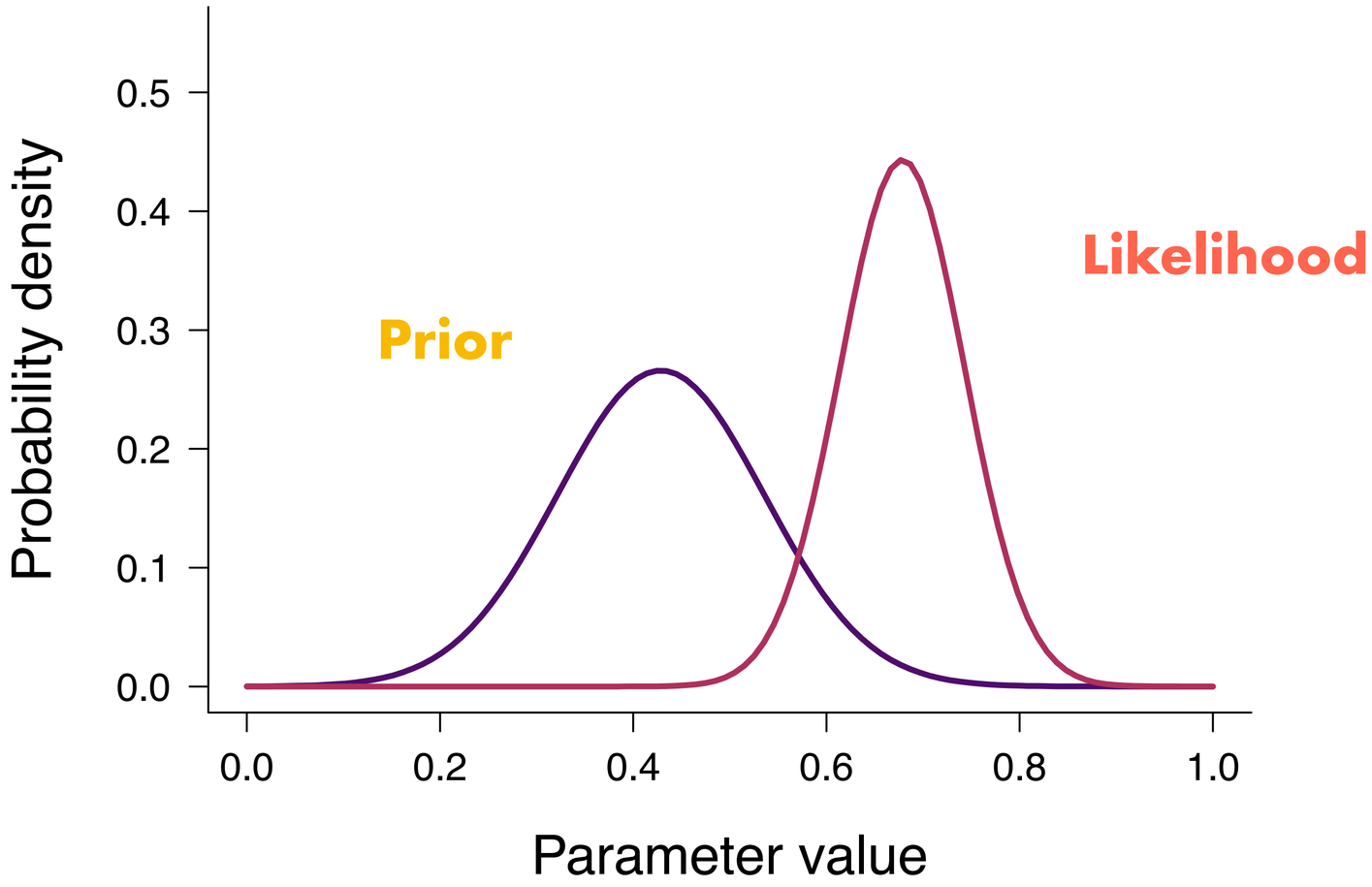
Bayesian analysis

$$\underbrace{p(\text{parameter} \mid \text{data})}_{\text{Posterior}} = \frac{\overbrace{p(\text{data} \mid \text{parameter})}^{\text{Likelihood}} * \overbrace{p(\text{parameter})}^{\text{Prior}}}{p(\text{data})}$$

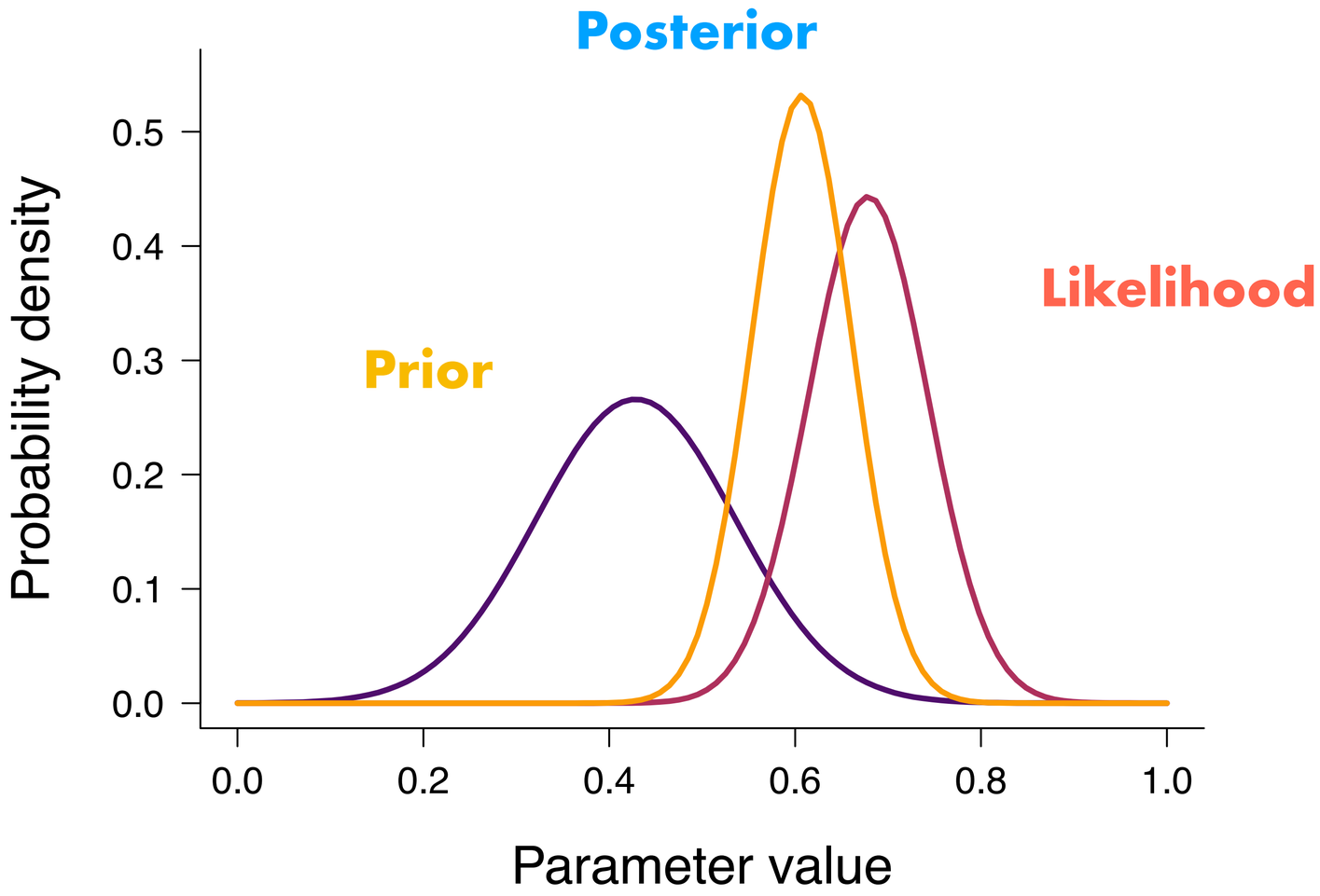
What is our expectation for the probability of the parameter?

We have three quantities of interest




Probability density
0.5
0.4
0.3
0.2
0.1
0.0
Prior
Likelihood
0.0
0.2
0.4
0.6
0.8
1.0
Parameter value

Bayesian analysis



Use prior knowledge to define *prior distributions*

Observe data and calculate the *likelihood*

Apply Bayes' theorem to estimate *posterior distributions*

Exercise 3: Explore impact of likelihood and prior on posterior probability with online tool

Defining a prior

Defining a prior

- vague/minimally informative

Defining a prior

- vague/minimally informative
- subjective/expert opinion

Defining a prior

- vague/minimally informative
- subjective/expert opinion
- estimate from previous data

Defining a prior

- vague/minimally informative
- subjective/expert opinion
- estimate from previous data

DON’T define your prior based on examining your data

Exercise 4: Defining a prior

Instead of a single value estimate for a parameter, we have an entire probability distribution across all unknowns.

$$y = a + bx + \epsilon$$

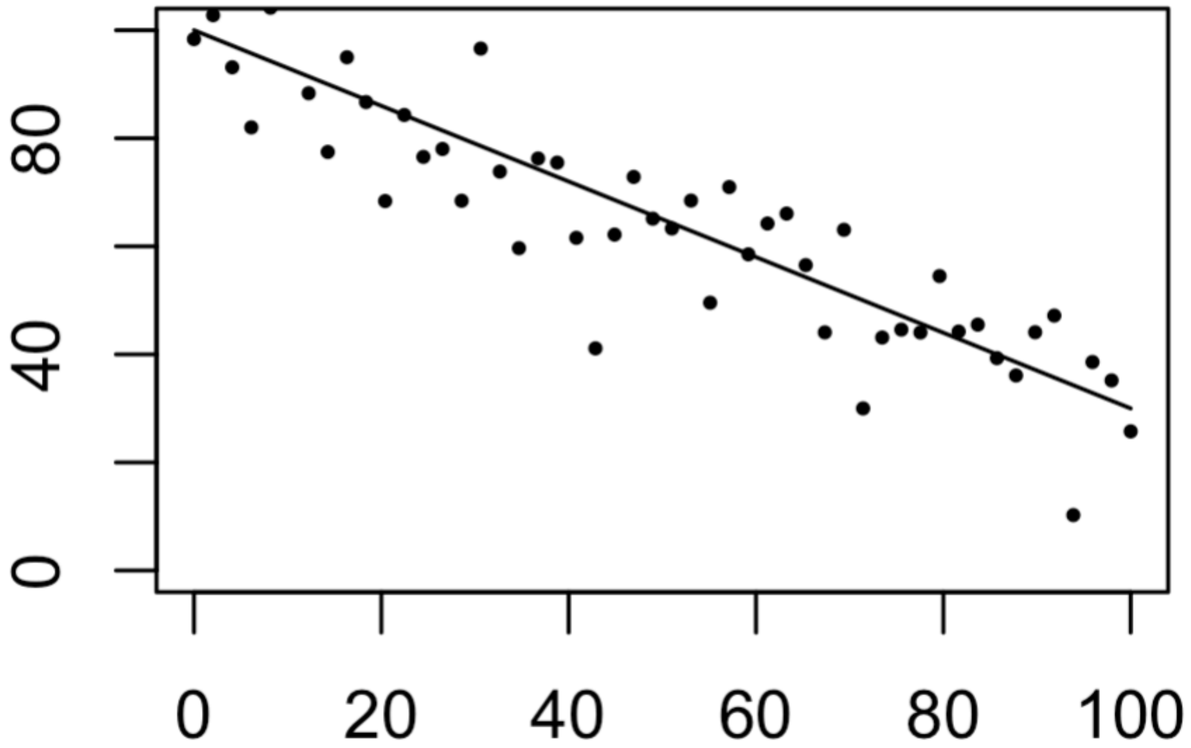
$$\epsilon \sim N(0, \sigma^2)$$

MLE

$$b = -0.67$$

Posterior mean

$$b = -0.67$$



Bayesian inference is intuitive

Frequentist inference is confusing

Bayesian inference is intuitive

Frequentist inference is confusing

Exercise 5: Confidence intervals

Frequentist confidence interval

"Were this procedure to be repeated on numerous samples, the fraction of calculated confidence intervals (which would differ for each sample) that encompass the true population parameter would tend toward 90%."

Frequentist confidence interval

"Were this procedure to be repeated on numerous samples, the fraction of calculated confidence intervals (which would differ for each sample) that encompass the true population parameter would tend toward 90%."

Bayesian credible interval

"An interval within which an unobserved parameter value falls with a particular probability."

Frequentist approach:

- No place for “prior beliefs”
- Inference should only depend on the data (likelihood)
- Probability is the same as frequency
- Point estimate

Frequentist approach:

- No place for “prior beliefs”
- Inference should only depend on the data (likelihood)
- Probability is the same as frequency
- Point estimate

Bayesian approach:

- Inference depends on prior knowledge and available data
- Probability is subjective; it is a degree of belief
- It is more intuitive! “I am 95% certain that…”

V. Introduction to brms

$$\underbrace{p(\text{parameter} \mid \text{data})}_{\text{Posterior}} = \frac{\overbrace{p(\text{data} \mid \text{parameter})}^{\text{Likelihood}} * \overbrace{p(\text{parameter})}^{\text{Prior}}}{p(\text{data})}$$

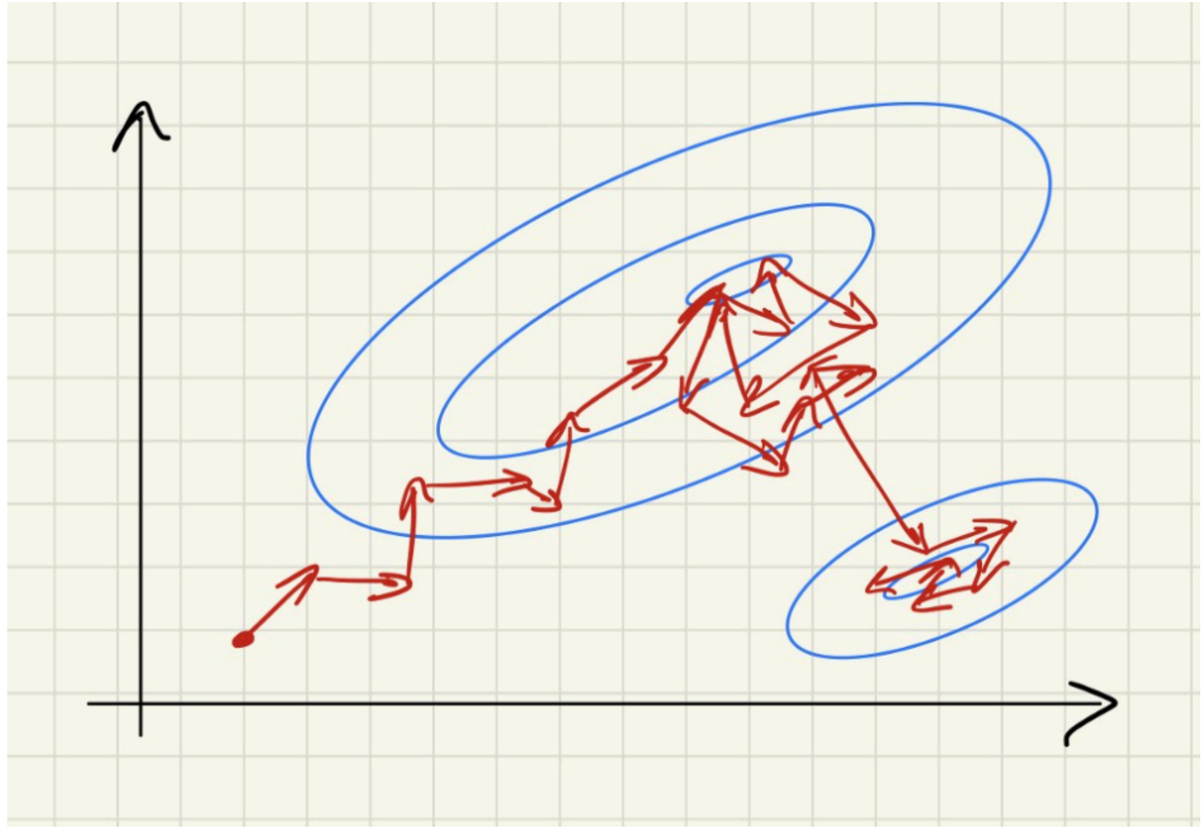
$$\underbrace{p(\text{parameter} \mid \text{data})}_{\text{Posterior}} = \frac{\overbrace{p(\text{data} \mid \text{parameter})}^{\text{Likelihood}} * \overbrace{p(\text{parameter})}^{\text{Prior}}}{😒}$$

Posterior **Likelihood** **Prior**

$$p(\text{parameter} \mid \text{data}) \propto p(\text{data} \mid \text{parameter}) * p(\text{parameter})$$

We can approximate the posterior by drawing a large random sample from the distribution using Markov chain Monte Carlo (MCMC)

1. Sample a starting point uniformly from prior distribution
2. Calculate probability density function (pdf) at that point.
3. Propose a new sample by stepping away according to some state transition function.
4. Calculate new pdf.
5. If new pdf is higher than old pdf, accept the step.
6. If less, … reject add current position as new sample and take new step.



Animated MCMC

‘brms()’ is a software package that leverages lme4-like syntax to make implementation of Stan functionality more accessible.

‘Stan’ is a software package that comes with a programming language to implement MCMC.

'Stan' is a software package that comes with a programming language to implement MCMC.

Stan uses Hamiltonian Monte Carlo and No-U-Turn Sampler (NUTS) algorithms to implement the MCMC sampling.

Chains:

Iterations:

Warmup:

Thin:

Draws:

Chains: Number of Markov chains

Iterations:

Warmup:

Thin:

Draws:

Chains: Number of Markov chains

Iterations: Number of steps per chain

Warmup:

Thin:

Draws:

Chains: Number of Markov chains

Iterations: Number of steps per chain

Warmup: First walks around parameter space that you throw away as the chain searches for the right area.

Thin:

Draws:

Chains: Number of Markov chains

Iterations: Number of steps per chain

Warmup: First walks around parameter space that you throw away as the chain searches for the right area.

Thin: Prevents correlation between steps by removing steps at this rate.

Draws:

Chains: Number of Markov chains

Iterations: Number of steps per chain

Warmup: First walks around parameter space that you throw away as the chain searches for the right area.

Thin: Prevents correlation between steps by removing steps at this rate.

Draws: (Iterations-warmup) x chains

Convergence: Do the samples in the chains converge in to the same maxima of the posterior distribution

1. Whether each chain converges on an estimate

2. Whether all chains converge on the same estimate

Convergence: Do the samples in the chains converge in to the same maxima of the posterior distribution

- Traceplot
- Rhat
- Effective sample size

Convergence: Do the samples in the chains converge in to the same maxima of the posterior distribution

- Traceplot
- Rhat
- Effective sample size

R practical: introduction to brms() & diagnosing convergence