Hola

Enfoque Estadístico del Aprendizaje

Inferencia bayesiana

Bernardo L. Español · 24 de Octubre

Introducción

LOS DATOS

Librerías

```
library(tidyverse)
library(tidymodels)
library(brms)
```

Los datos

```
df <- read_delim("usu_individual_T126.txt", delim = ";") %>%
  rename(age = CH06, salary = P21) %>%
  filter(salary > 0, ESTADO == 1, between(age, 18, 40)) %>%
  arrange(age) %>%
  select(age, salary)
df
```

```
# A tibble: 7,534 × 2
     age salary
   <int>  <int>
 1    18 400000
 2    18 160000
 3    18 250000
 4    18 100000
 5    18 750000
# ℹ 7,529 more rows
```

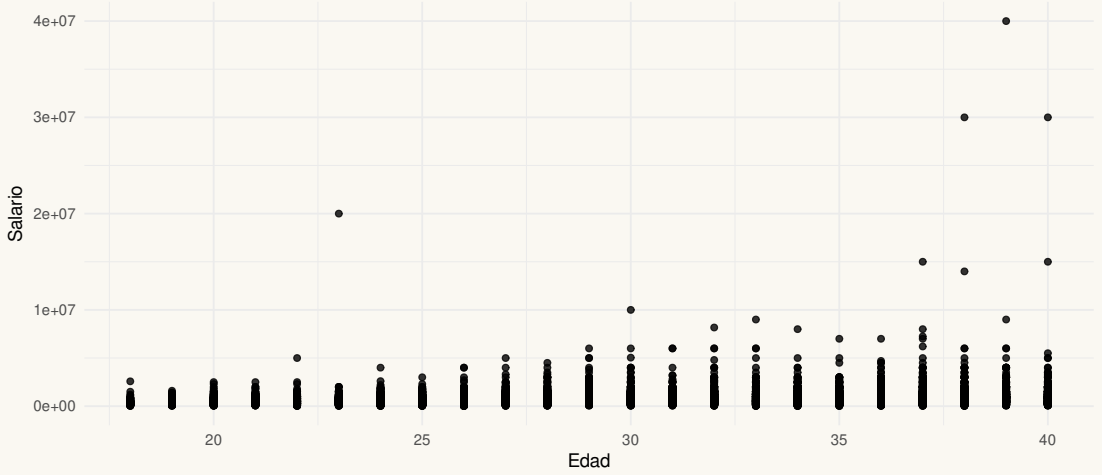
REGRESIÓN LINEAL

Salario contra edad

```
ggplot(df, aes(x = age, y = salary)) +
  geom_point(alpha = 0.8) +
  labs(x = "Edad", y = "Salario") +
  theme_minimal()
```

REGRESIÓN LINEAL

Salario contra edad



El ajuste lineal

```
mod_lm <- lm(formula = salary ~ age, data = df)
tidy(mod_lm)
```

term	estimate	std.error	statistic	p.value
(Intercept)	−32380	61710	−0.5247	5.998e−01
age	30830	1967	15.6800	1.574e−54

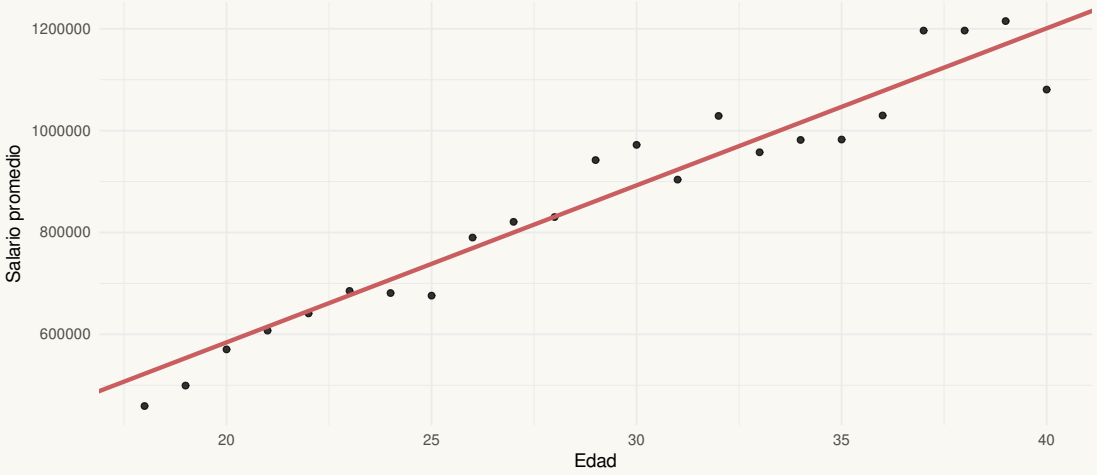
REGRESIÓN LINEAL

El ajuste lineal

```
df_mean <- df %>%
  group_by(age) %>%
  summarise(salary_mean = mean(salary))

ggplot(df_mean, aes(x = age, y = salary_mean)) +
  geom_point(alpha = 0.8) +
  geom_abline(
    intercept = coef(mod_lm)[1],
    slope = coef(mod_lm)[2],
    linewidth = 1.2, col = "#C95E61"
  ) +
  theme_minimal()
```

REGRESIÓN LINEAL

El ajuste lineal



La predicción de `lm()`

Con `lm()`, la predicción para una edad es un valor y un intervalo:

```
predict(mod_lm, newdata = data.frame(age = 30),
        interval = "prediction", level = 0.68)
```

```
       fit       lwr     upr
1 892568.7 -132182.7 1917320
```

Inferencia bayesiana

DOS ENFOQUES

El enfoque clásico

- Los parámetros β_0, β_1 son fijos y desconocidos.

DOS ENFOQUES

El enfoque clásico

- Los parámetros β_0, β_1 son fijos y desconocidos.
- Los estimadores $\hat{\beta}_0$, $\hat{\beta}_1$ son variables aleatorias.

DOS ENFOQUES

El enfoque clásico

- Los parámetros β_0, β_1 son fijos y desconocidos.
- Los estimadores $\hat{\beta}_0$, $\hat{\beta}_1$ son variables aleatorias.
- Obtenemos estimaciones de los estimadores, no de los parámetros.

DOS ENFOQUES

El enfoque clásico

- Los parámetros β_0, β_1 son fijos y desconocidos.
- Los estimadores $\hat{\beta}_0$, $\hat{\beta}_1$ son variables aleatorias.
- Obtenemos estimaciones de los estimadores, no de los parámetros.
- Calculamos los errores a partir de la distribución del estadístico.

DOS ENFOQUES

El enfoque clásico

- Los parámetros β_0, β_1 son fijos y desconocidos.
- Los estimadores $\hat{\beta}_0$, $\hat{\beta}_1$ son variables aleatorias.
- Obtenemos estimaciones de los estimadores, no de los parámetros.
- Calculamos los errores a partir de la distribución del estadístico.
- Pide pocos supuestos. Los importantes son que los errores tengan esperanza nula y no estén correlacionados con X_i.

DOS ENFOQUES

El enfoque bayesiano

Tratamos a los parámetros como variables aleatorias. Antes de mirar los datos les damos una distribución, que llamamos *prior*.

El enfoque bayesiano

Tratamos a los parámetros como variables aleatorias. Antes de mirar los datos les damos una distribución, que llamamos *prior*.

Para el modelo lineal, típicamente:

$$y_i \sim \text{Normal}(\beta_0 + \beta_1 x_i, \sigma)$$

$$\beta_0 \sim \text{Normal}(0, \tau_0), \quad \beta_1 \sim \text{Normal}(0, \tau_1), \quad \sigma \sim \text{Exponencial}(\lambda)$$

Los valores de τ_0, τ_1 y λ los fija el investigador, e indican cuánta libertad les damos a los parámetros.

El enfoque bayesiano

Tratamos a los parámetros como variables aleatorias. Antes de mirar los datos les damos una distribución, que llamamos *prior*.

Para el modelo lineal, típicamente:

$$y_i \sim \text{Normal}(\beta_0 + \beta_1 x_i, \sigma)$$

$$\beta_0 \sim \text{Normal}(0, \tau_0), \quad \beta_1 \sim \text{Normal}(0, \tau_1), \quad \sigma \sim \text{Exponencial}(\lambda)$$

Los valores de τ_0, τ_1 y λ los fija el investigador, e indican cuánta libertad les damos a los parámetros.

Lo que obtenemos es la distribución a posteriori de cada parámetro, en lugar de un valor.

El teorema de Bayes

$$\Pr(\theta \mid y) = \frac{\Pr(y \mid \theta)\Pr(\theta)}{\Pr(y)}$$

El teorema de Bayes

$$\Pr(\theta \mid y) = \frac{\Pr(y \mid \theta)\Pr(\theta)}{\Pr(y)} \propto \Pr(y \mid \theta)\Pr(\theta)$$

El teorema de Bayes

$$\Pr(\theta \mid y) = \frac{\Pr(y \mid \theta)\Pr(\theta)}{\Pr(y)} \propto \Pr(y \mid \theta)\Pr(\theta)$$

$$\underbrace{\Pr(\theta \mid y)}_{\text{posterior}} \propto \underbrace{\Pr(y \mid \theta)}_{\text{verosimilitud}} \underbrace{\Pr(\theta)}_{\text{prior}}$$

Ventajas del enfoque bayesiano

- Expresa la incertidumbre de forma natural, con una distribución para cada parámetro.

DOS ENFOQUES

Ventajas del enfoque bayesiano

- Expresa la incertidumbre de forma natural, con una distribución para cada parámetro.
- Permite incorporar información previa a través del prior.

DOS ENFOQUES

Ventajas del enfoque bayesiano

- Expresa la incertidumbre de forma natural, con una distribución para cada parámetro.
- Permite incorporar información previa a través del prior.
- Da más flexibilidad para armar el modelo, como cambiar la verosimilitud o agregar efectos por sujeto.

DOS ENFOQUES

Ventajas del enfoque bayesiano

- Expresa la incertidumbre de forma natural, con una distribución para cada parámetro.
- Permite incorporar información previa a través del prior.
- Da más flexibilidad para armar el modelo, como cambiar la verosimilitud o agregar efectos por sujeto.
- Propaga la incertidumbre de forma natural, por ejemplo de los parámetros a la predicción.

Desventajas del enfoque bayesiano

- Estimar los parámetros es lento. Un ajuste puede tardar minutos donde `lm()` es inmediato.

Desventajas del enfoque bayesiano

- Estimar los parámetros es lento. Un ajuste puede tardar minutos donde `lm()` es inmediato.
- Poner un prior puede sentirse como hacer trampa, sobre todo con pocos datos, que es cuando más pesa.

Un ejemplo con un parámetro

UN EJEMPLO CON UN PARÁMETRO

¿Qué fracción de la gente trabaja?

Llamamos θ a la fracción de personas de 18 a 40 años que trabaja. Para estimarla, tomamos 20 personas al azar, como si no conociéramos la muestra completa. En esta submuestra trabajan 11 personas, un 55%.

UN EJEMPLO CON UN PARÁMETRO

¿Qué fracción de la gente trabaja?

Llamamos θ a la fracción de personas de 18 a 40 años que trabaja. Para estimarla, tomamos 20 personas al azar, como si no conociéramos la muestra completa. En esta submuestra trabajan 11 personas, un 55%.

Con el enfoque clásico, estimamos θ con esa proporción, 0.55.

UN EJEMPLO CON UN PARÁMETRO

¿Qué fracción de la gente trabaja?

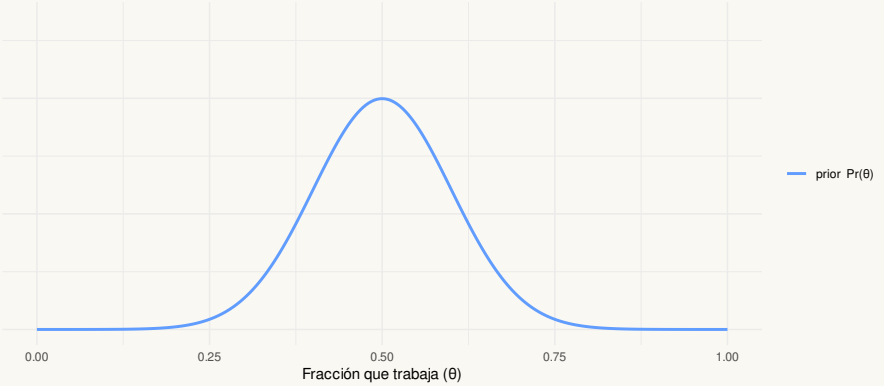
Llamamos θ a la fracción de personas de 18 a 40 años que trabaja. Para estimarla, tomamos 20 personas al azar, como si no conociéramos la muestra completa. En esta submuestra trabajan 11 personas, un 55%.

Con el enfoque clásico, estimamos θ con esa proporción, 0.55.

Con el enfoque bayesiano, primero decimos qué creemos sobre θ antes de mirar la encuesta. Después corregimos esa creencia con los datos.

UN EJEMPLO CON UN PARÁMETRO

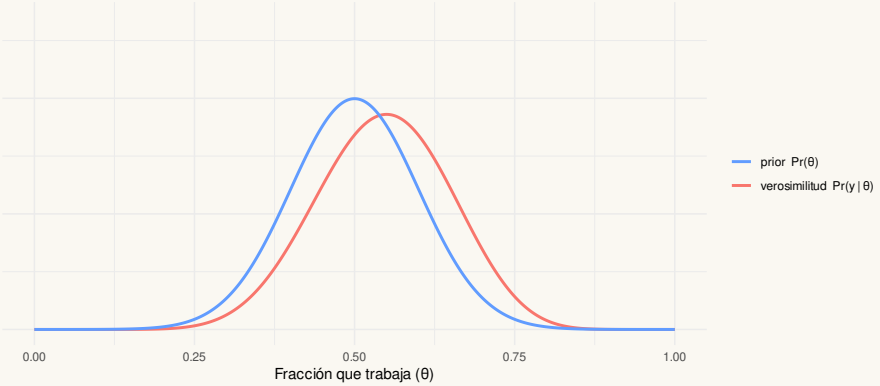
El prior



Antes de mirar la encuesta, creemos que trabaja más o menos la mitad,
y que es raro que trabaje menos del 30% o más del 70%.
Lo escribimos como una normal con media 0.5 y desvío 0.1.

UN EJEMPLO CON UN PARÁMETRO

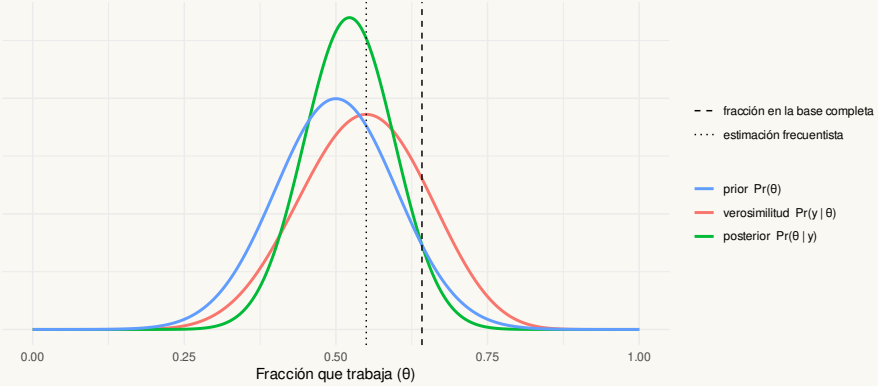
La verosimilitud



La verosimilitud dice, para cada valor de θ, qué tan probable es que trabajen 11 de 20 personas.

Si trabajara el 10% de la gente, sería casi imposible. El máximo está en 0.55, la proporción de la muestra.

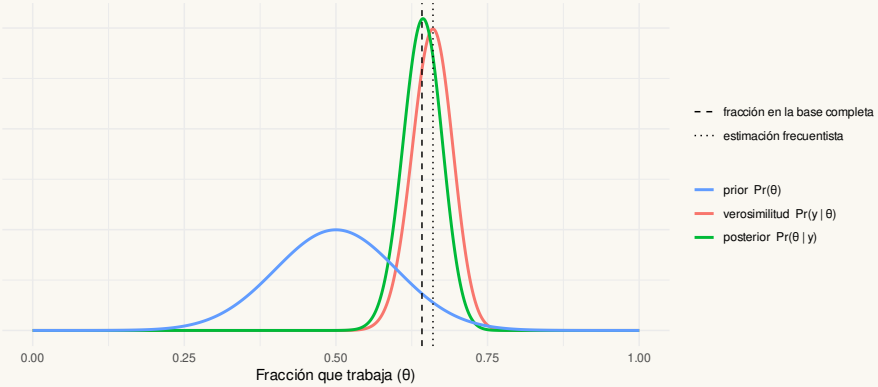
El posterior



El posterior es el producto del prior y la verosimilitud. Queda entre los dos y es más angosto que ambos, porque junta la información de las dos curvas.

UN EJEMPLO CON UN PARÁMETRO

Con 200 personas



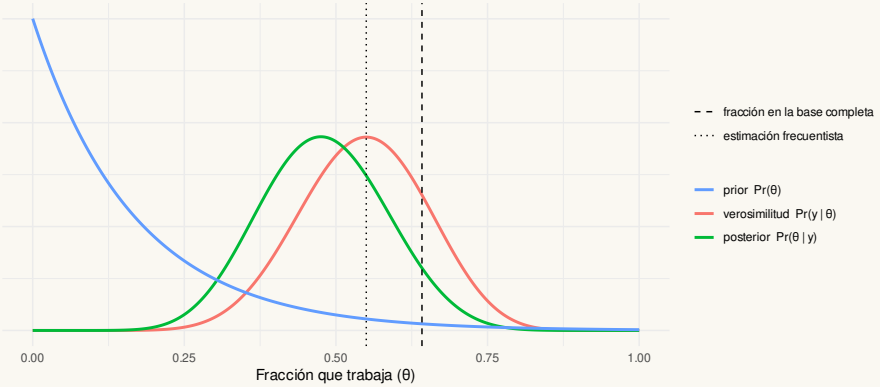
Encuestamos a 200 personas y trabajan 132, el 66%.

La verosimilitud se angosta y el posterior queda casi encima de ella.

El prior pesa mucho menos que con 20 personas.

UN EJEMPLO CON UN PARÁMETRO

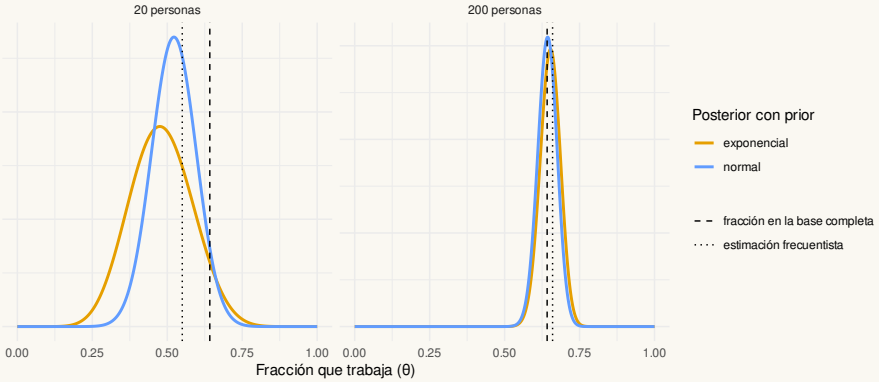
Un prior exponencial



Volvemos a las 20 personas, pero ahora creemos que trabaja poca gente. Usamos un prior exponencial, que es máximo en 0 y decae. El posterior se corre hacia la izquierda de la verosimilitud.

UN EJEMPLO CON UN PARÁMETRO

Los posteriores con cada prior



Con 20 personas, el prior cambia bastante el posterior: su media pasa de 0.52 a 0.48.

Con 200, los dos posteriores casi coinciden y quedan sobre la fracción de la base.

Modelo lineal bayesiano

Cómo se calcula el posterior

Para dibujar las curvas del ejemplo, calculamos el prior y la verosimilitud en 1000 valores de θ entre 0 y 1, y los multiplicamos. Con un parámetro, eso alcanza.

Cómo se calcula el posterior

Para dibujar las curvas del ejemplo, calculamos el prior y la verosimilitud en 1000 valores de θ entre 0 y 1, y los multiplicamos. Con un parámetro, eso alcanza.

Con tres parámetros, como en la regresión (β_0, β_1 y σ), la misma grilla tendría 1000^3 puntos.

Cómo se calcula el posterior

Para dibujar las curvas del ejemplo, calculamos el prior y la verosimilitud en 1000 valores de θ entre 0 y 1, y los multiplicamos. Con un parámetro, eso alcanza.

Con tres parámetros, como en la regresión (β_0, β_1 y σ), la misma grilla tendría 1000^3 puntos.

Los métodos de Markov Chain Monte Carlo (MCMC) generan muestras del posterior evaluando el producto solo donde el posterior es grande. `brm()` usa Stan, un programa de MCMC. Las *chains* y los *draws* de la salida de `brm()` son esas muestras.

El modelo en Stan

Stan describe el modelo en bloques. Para la regresión del salario contra la edad:

```
data {
  int<lower=1> N;          // cantidad de personas
  vector[N] age;           // predictor
  vector[N] salary;        // respuesta
  real<lower=0> tau0;      // escalas de los priors, elegidas por nosotros
  real<lower=0> tau1;
  real<lower=0> lambda;
}
parameters {
  real b0;                 // ordenada
  real b1;                 // pendiente
  real<lower=0> sigma;     // desvío del ruido
}
model {
  b0 ~ normal(0, tau0);    // priors
  b1 ~ normal(0, tau1);
  sigma ~ exponential(lambda);
  salary ~ normal(b0 + b1 * age, sigma);  // verosimilitud
}
```

$$\beta_0 \sim \text{Normal}(0, \tau_0)$$
$$\beta_1 \sim \text{Normal}(0, \tau_1)$$
$$\sigma \sim \text{Exponencial}(\lambda)$$
$$y_i \sim \text{Normal}(\beta_0 + \beta_1 x_i, \sigma)$$

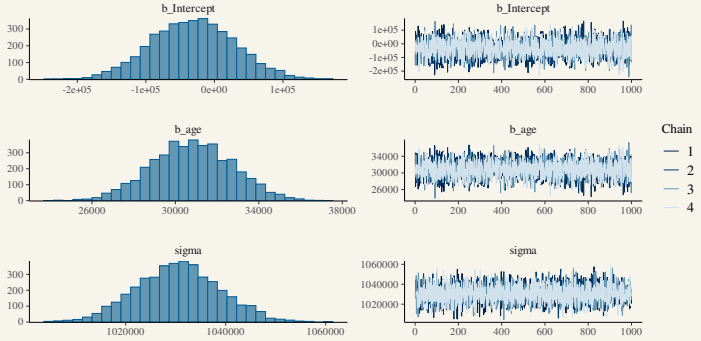
brm()

```
mod_brm_default <- brm(salary ~ age, data = df, refresh = 0)
mod_brm_default
```

```
 Family: gaussian
  Links: mu = identity
Formula: salary ~ age
   Data: df (Number of observations: 7534)
  Draws: 4 chains, each with iter = 2000; warmup = 1000; thin = 1;
         total post-warmup draws = 4000

Regression Coefficients:
          Estimate Est.Error   l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
Intercept -31188.64  60989.90 -150091.31 89600.89 1.00     3713     2887
age        30790.33   1938.87   27007.13 34521.05 1.00     3693     2933

Further Distributional Parameters:
       Estimate Est.Error   l-95% CI   u-95% CI Rhat Bulk_ESS Tail_ESS
sigma 1030501.77   8369.46 1014585.40 1046827.92 1.00     3645     2984
```

El posterior de los parámetros

```
plot(mod_brm_default)
```



Preguntas al posterior

Con los *draws* respondemos preguntas sobre los parámetros contando cuántos cumplen una condición. Por ejemplo, ¿qué probabilidad hay de que el salario suba más de 30000 por año de edad?

```
draws <- as_draws_df(mod_brm_default)
```

Preguntas al posterior

Con los *draws* respondemos preguntas sobre los parámetros contando cuántos cumplen una condición. Por ejemplo, ¿qué probabilidad hay de que el salario suba más de 30000 por año de edad?

```
draws <- as_draws_df(mod_brm_default)

mean(draws$b_age > 30000)
```

```
[1] 0.6525
```

Preguntas al posterior

Con los *draws* respondemos preguntas sobre los parámetros contando cuántos cumplen una condición. Por ejemplo, ¿qué probabilidad hay de que el salario suba más de 30000 por año de edad?

```
draws <- as_draws_df(mod_brm_default)

mean(draws$b_age > 30000)
```

```
[1] 0.6525
```

Si queremos el intervalo de coonfianza del 95%, tomamos los cuantiles 2.5% y 97.5% de los *draws*:

```
quantile(draws$b_age, c(0.025, 0.975))
```

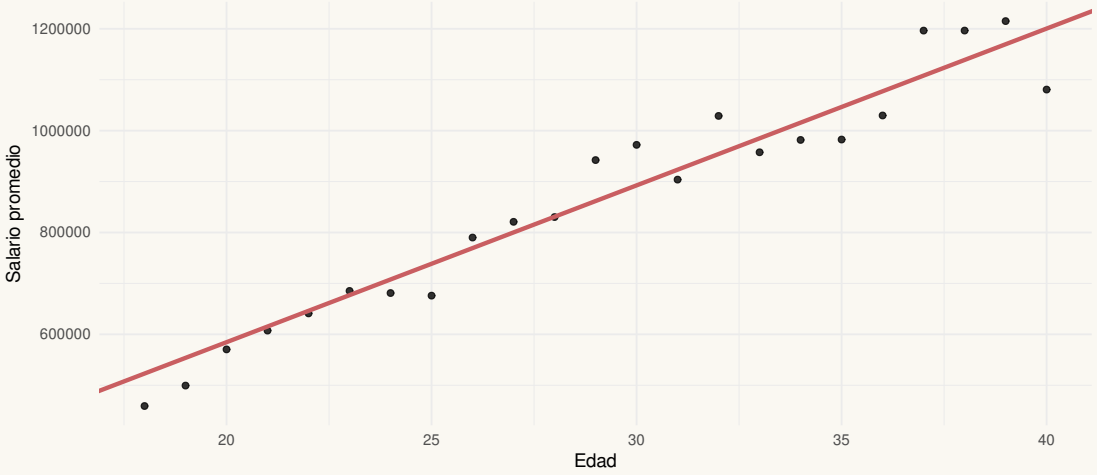
```
    2.5%    97.5% 
27007.13 34521.05 
```

Predicción para los valores medios

```
b <- fixef(mod_brm_default)

ggplot(df_mean, aes(x = age, y = salary_mean)) +
  geom_point(alpha = 0.8) +
  geom_abline(
    intercept = b["Intercept", "Estimate"],
    slope     = b["age", "Estimate"],
    col = "#C95E61",
    linewidth = 1.2
  ) +
  theme_minimal()
```

MODELO LINEAL BAYESIANO

Predicción para los valores medios



conditional_effects()

También podemos usar `brms::conditional_effects()` para graficar la predicción:

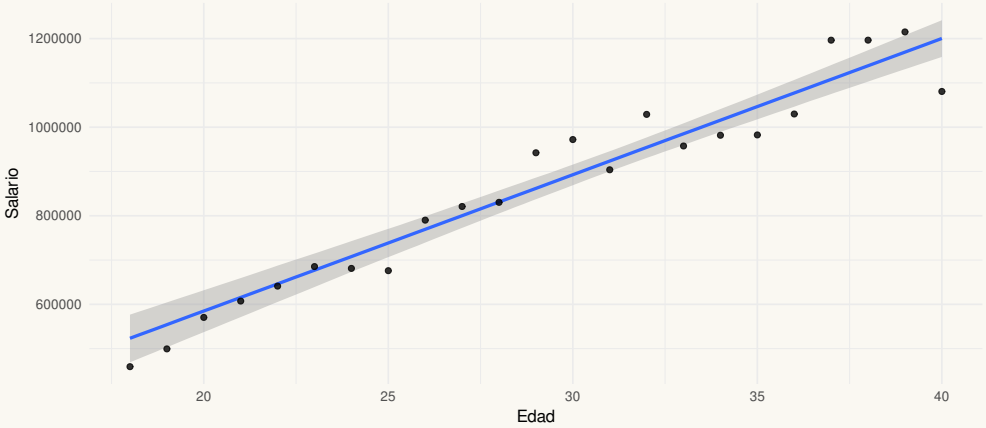
```
ce <- conditional_effects(mod_brm_default)
p  <- plot(ce, plot = FALSE)[[1]]
```

conditional_effects()

También podemos usar `brms::conditional_effects()` para graficar la predicción:

```
ce <- conditional_effects(mod_brm_default)
p  <- plot(ce, plot = FALSE)[[1]]

p +
  geom_point(data = df_mean, aes(x = age, y = salary_mean),
             alpha = 0.8, inherit.aes = FALSE) +
  theme_minimal()
```

conditional_effects()



La banda es la incertidumbre de la recta media, que viene de β_0 y β_1.

Los priors

```
get_prior(mod_brm_default)
```

```
                         prior     class coef group resp dpar nlpar lb ub tag       source
                        (flat)         b                                           default
                        (flat)         b  age                                (vectorized)
 student_t(3, 8e+05, 539666.4) Intercept                                       default
     student_t(3, 0, 539666.4)     sigma                              0         default
```

Los priors

```
get_prior(mod_brm_default)
```

```
                        prior     class coef group resp dpar nlpar lb ub tag       source
                       (flat)         b                                           default
                       (flat)         b  age                                (vectorized)
 student_t(3, 8e+05, 539666.4) Intercept                                       default
     student_t(3, 0, 539666.4)     sigma                               0       default
```

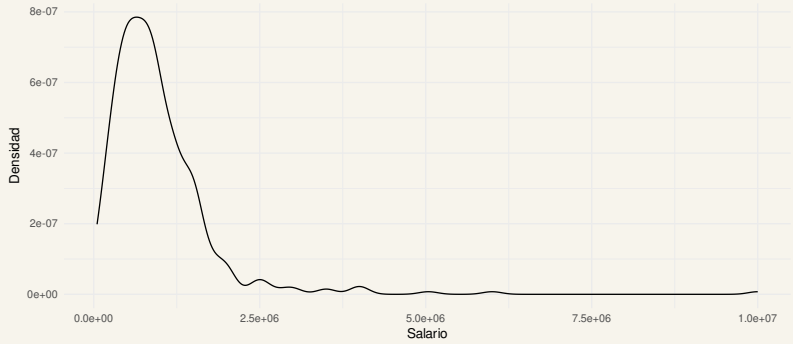
Por defecto, `brms` usa otros priors que los que vimos antes. Para la pendiente usa un prior plano, con todos los valores igual de probables, y para el intercepto y para σ, una t de Student cuya escala sale de los datos.

Predicción

PREDICCIÓN

Los salarios a los 30 años

```
age_eval <- 30

ggplot(filter(df, age == age_eval), aes(x = salary)) +
  geom_density() +
  theme_minimal()
```



PREDICCIÓN

posterior_predict()

Para predecir salarios a una edad usamos `brms::posterior_predict()`, y los superponemos a los salarios observados a esa edad:

```
plot_age_posterior_predict <- function(model, age_eval) {
  salary_draws <- posterior_predict(
    model,
    newdata = data.frame(age = age_eval),
    ndraws = 4000
  )[, 1]
```

PREDICCIÓN

posterior_predict()

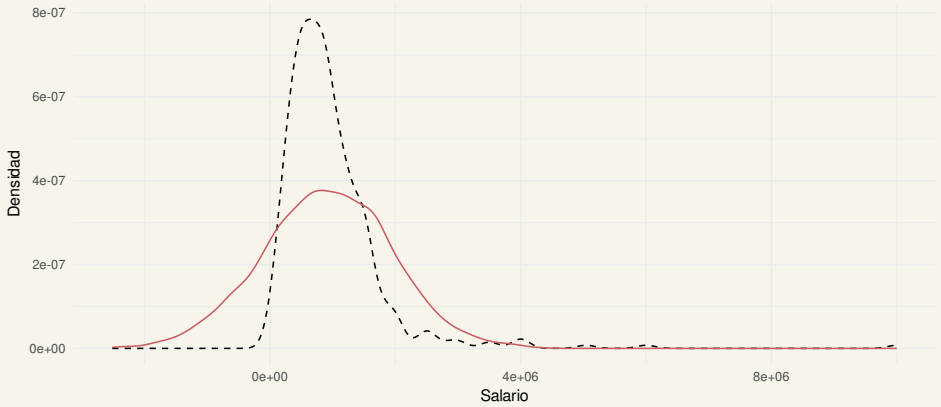
Para predecir salarios a una edad usamos `brms::posterior_predict()`, y los superponemos a los salarios observados a esa edad:

```
plot_age_posterior_predict <- function(model, age_eval) {
  salary_draws <- posterior_predict(
    model,
    newdata = data.frame(age = age_eval),
    ndraws = 4000
  )[, 1]

  ggplot() +
    geom_density(data = filter(df, age == age_eval), aes(x = salary),
                 linetype = "dashed") +
    geom_density(data = data.frame(salary = salary_draws), aes(x = salary),
                 col = "#C95E61") +
    theme_minimal()
}
```

PREDICCIÓN

posterior_predict()

```
plot_age_posterior_predict(mod_brm_default, 30)
```

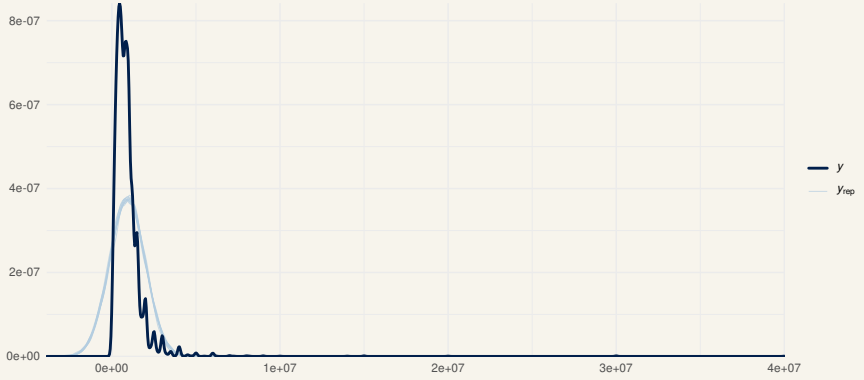


En rojo, la predicción del modelo. En trazos, los salarios observados a los 30 años.

PREDICCIÓN

Predicción para todas las edades

```
pp_check(mod_brm_default)
```



Otra verosimilitud

OTRA VEROSIMILITUD

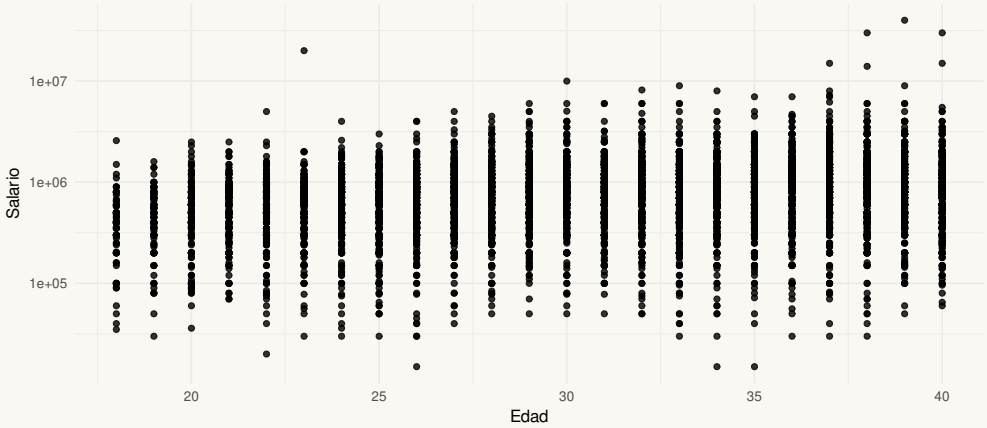
Salario contra edad, en escala log

Los salarios son positivos y asimétricos. Veamos los datos en escala log:

```
ggplot(df, aes(x = age, y = salary)) +
  geom_point(alpha = 0.8) +
  scale_y_log10() +
  theme_minimal()
```

OTRA VEROSIMILITUD

Salario contra edad, en escala log



Vemos que, en escala log, la nube es mucho más simétrica.

OTRA VEROSIMILITUD

Familia y enlace

$$y \sim \text{Normal}(\mu, \sigma), \quad \mu = \beta_0 + \beta_1 x \qquad \longrightarrow \qquad y \sim \text{Gamma}(\mu, \alpha), \quad \log(\mu) = \beta_0 + \beta_1 x$$

OTRA VEROSIMILITUD

brm() con family = Gamma("log")

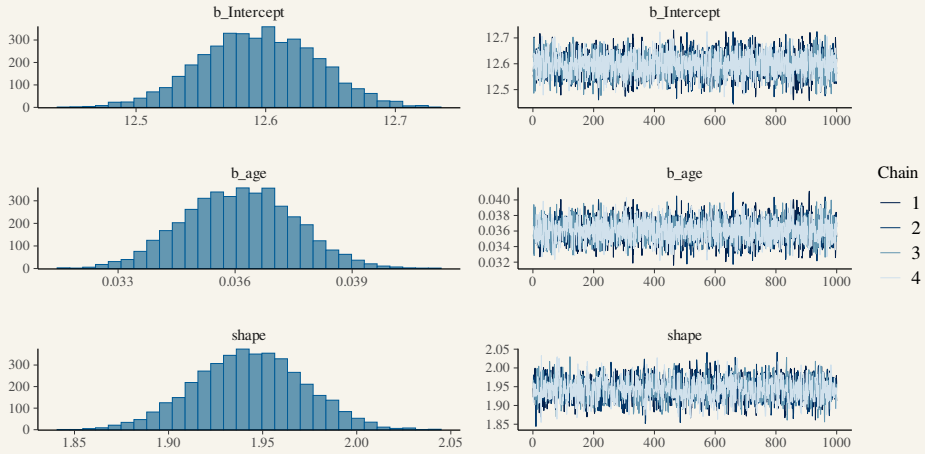
```
mod_brm_family_gamma <- brm(
  salary ~ age, data = df,
  family = Gamma("log"),
  refresh = 0
)
mod_brm_family_gamma
```

```
 Family: gamma
  Links: mu = log
Formula: salary ~ age
   Data: df (Number of observations: 7534)
  Draws: 4 chains, each with iter = 2000; warmup = 1000; thin = 1;
         total post-warmup draws = 4000

Regression Coefficients:
          Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
Intercept    12.59      0.04    12.51    12.68 1.00     4218     2817
age           0.04      0.00     0.03     0.04 1.00     4258     2486

Further Distributional Parameters:
      Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
shape     1.94      0.03     1.89     2.00 1.00     2170     2153
```

OTRA VEROSIMILITUD

El posterior de los parámetros

```
plot(mod_brm_family_gamma)
```



OTRA VEROSIMILITUD

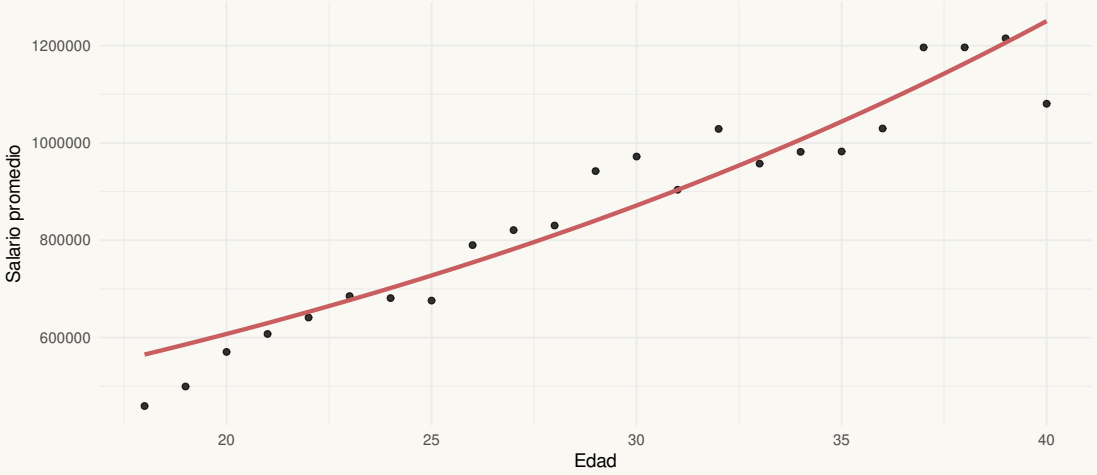
Predicción para los valores medios

El modelo es lineal en $\log(\mu)$, así que para volver a la escala del salario usamos `exp()`.

```
b <- fixef(mod_brm_family_gamma)

ggplot(df_mean, aes(x = age, y = salary_mean)) +
  geom_point(alpha = 0.8) +
  geom_function(
    fun = \(x) exp(b["Intercept", "Estimate"] +
                   b["age", "Estimate"] * x),
    col = "#C95E61",
    linewidth = 1.2
  ) +
  theme_minimal()
```

OTRA VEROSIMILITUD

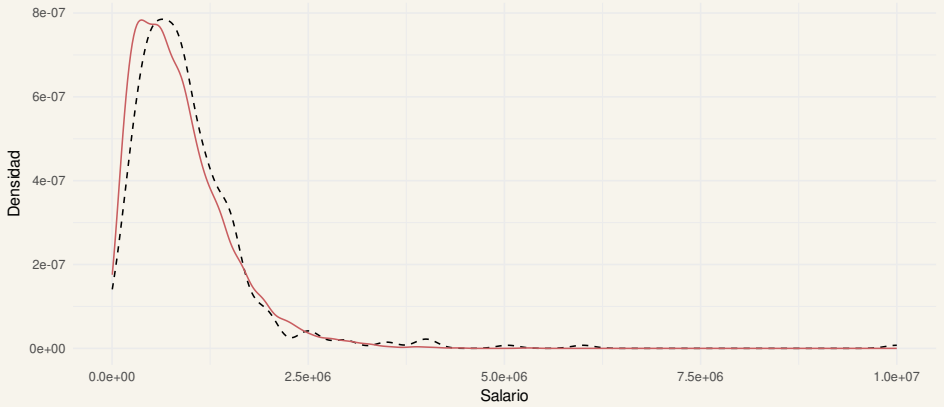
Predicción para los valores medios



OTRA VEROSIMILITUD

Predicción para una edad

```
plot_age_posterior_predict(mod_brm_family_gamma, 30)
```

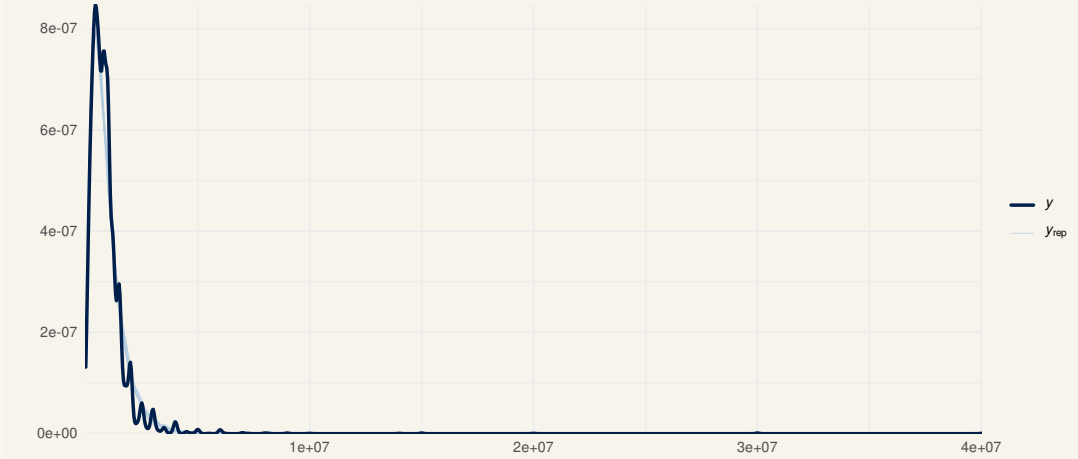


En rojo, la predicción del modelo. En trazos, los salarios observados a los 30 años.

OTRA VEROSIMILITUD

Predicción para todas las edades

```
pp_check(mod_brm_family_gamma)
```

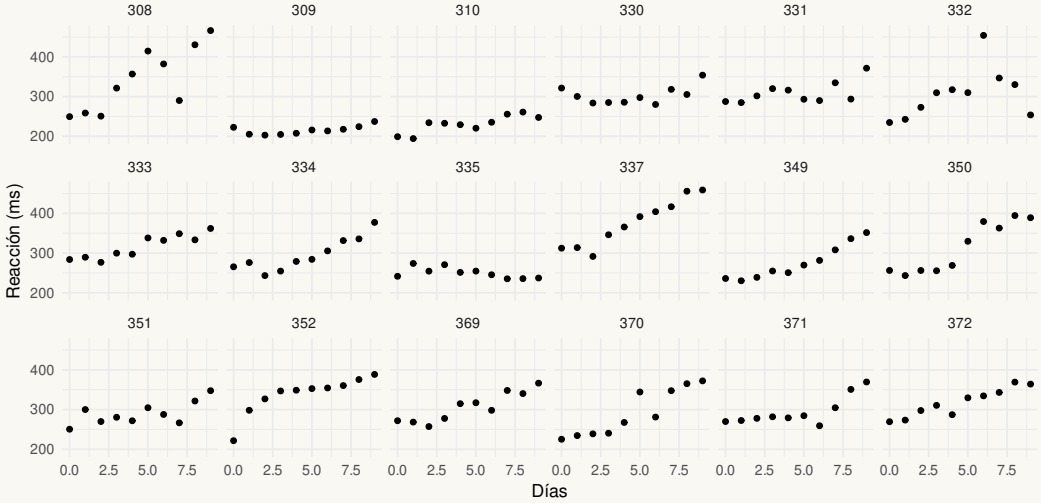


Otras ventajas del modelo bayesiano

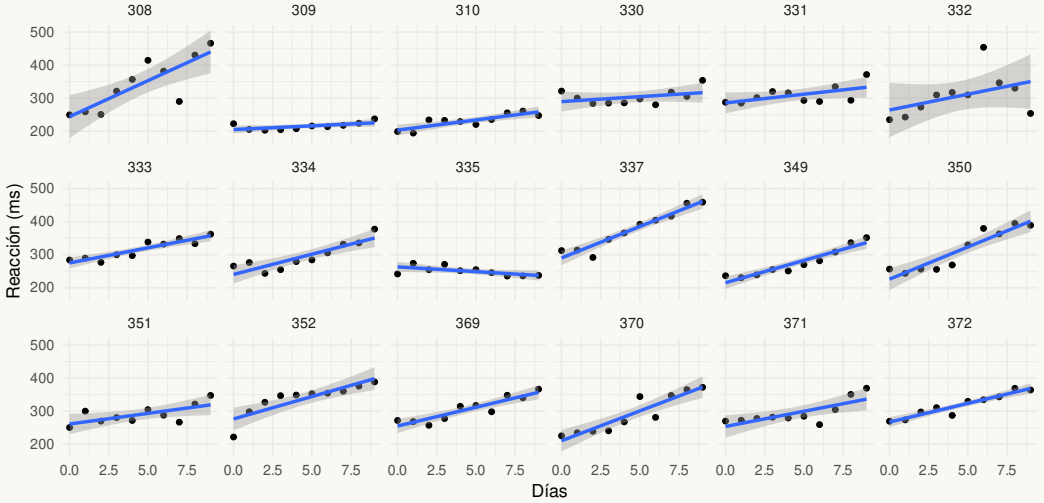
sleepstudy

```
data("sleepstudy", package = "lme4")

ggplot(sleepstudy, aes(x = Days, y = Reaction, group = Subject)) +
  geom_point(size = 1.3) +
  facet_wrap(~ Subject, nrow = 3, ncol = 6) +
  labs(x = "Días", y = "Reacción (ms)") +
  theme_minimal()
```

sleepstudy

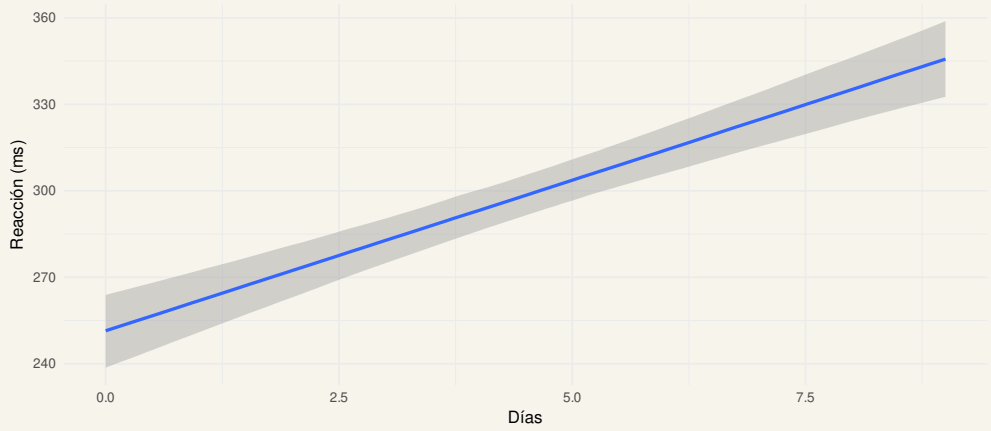

308
309
310
330
331
332
333
334
335
337
349
350
351
352
369
370
371
372
Reacción (ms)
200
300
400
0.0
2.5
5.0
7.5
Días

sleepstudy


308
309
310
330
331
332
333
334
335
337
349
350
351
352
369
370
371
372
Reacción (ms)
Días
500
400
300
200
0.0
2.5
5.0
7.5

Un modelo para todos

```
mod <- brm(Reaction ~ Days, data = sleepstudy, refresh = 0)
conditional_effects(mod)
```



El modelo por sujeto

En el primer modelo, todas las mediciones comparten la misma recta:

$$y_i \sim \text{Normal}(\beta_0 + \beta_1 x_i, \sigma)$$

El modelo por sujeto

En el primer modelo, todas las mediciones comparten la misma recta:

$$y_i \sim \text{Normal}(\beta_0 + \beta_1 x_i, \sigma)$$

En el modelo por sujeto, la medición i del sujeto j sigue la recta media más un apartamiento propio:

$$y_{ij} \sim \text{Normal}\big(\mu_{ij}, \sigma\big), \quad \mu_{ij} = \underbrace{\beta_0 + \beta_1 x_{ij}}_{\text{recta media}} + \underbrace{u_{0j} + u_{1j} x_{ij}}_{\text{apartamiento}}$$

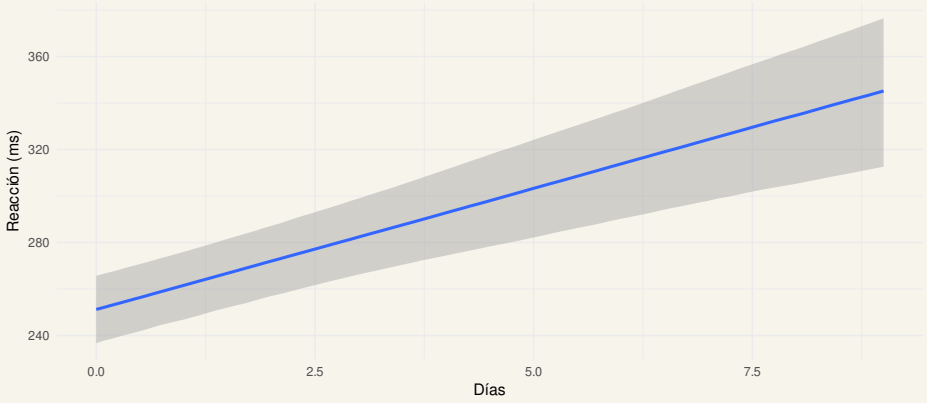
Los apartamientos de los sujetos salen de una distribución común:

$$\big(u_{0j}, u_{1j}\big) \sim \text{Normal}((0,0), \Sigma)$$

Σ dice cuánto varían las rectas de un sujeto a otro. Todos los parámetros llevan un prior.

Un modelo por sujeto

```
mod_conditioned <- brm(Reaction ~ Days + (Days | Subject),
                       data = sleepstudy, refresh = 0)
conditional_effects(mod_conditioned)
```



Esto es todo.