Tarea 8

Juan M Karawcki, Bruno PIntos

2025-11-06

9.10.2 Ejercicios aplicados

Ejercicio 9.9 (¿Qué tan húmedo es *demasiado* húmedo?: construcción del modelo)
A lo largo de este capítulo, exploramos cómo el uso de bicicletas varía con la temperatura. ¿Y la humedad? En los próximos ejercicios, vas a explorar el **modelo de regresión Normal** de viajes (Y) por **humedad** (X) usando el conjunto de datos `bikes`.
Con base en análisis previos de sistemas de bicicletas compartidas, supondremos el siguiente entendimiento previo de esta relación:

- En un día de **humedad promedio**, típicamente hay **alrededor de 5000 viajes**, aunque este promedio podría estar entre **1000** y **9000**.
- El **uso tiende a disminuir** a medida que **aumenta la humedad**. En particular, por cada **punto porcentual** adicional de humedad, el uso tiende a **disminuir en 10 viajes**, aunque esta disminución promedio podría estar entre **0** y **20**.
- El uso está **débilmente relacionado** con la humedad. Para cualquier nivel de humedad, el uso tenderá a variar con una **desviación estándar grande de 2000** viajes.

1. **Sintonizá** el modelo de regresión Normal (9.6) para que coincida con nuestro entendimiento previo. Usá notación cuidadosa para escribir la **estructura bayesiana completa** del modelo.

Variable explicativa centrada

Para que la **ordenada al origen** represente directamente el **promedio de viajes a humedad promedio**, centramos la humedad:

- Sea X_i la humedad (en puntos porcentuales) del día i.
- Definimos $x_i \equiv X_i - \bar{X}$, donde $\bar{X}$ es la humedad promedio de los días observados.

Con esta transformación, cuando $x_i = 0$ (es decir, $X_i = \bar{X}$), la media del modelo coincide con el promedio de viajes en un día de humedad promedio.

Estructura bayesiana completa

1) Modelo de datos (verosimilitud) Para cada día $i = 1, \ldots, n$,

$$Y_i \mid \mu_i, \sigma \;\sim\; \text{Normal}(\mu_i,\ \sigma), \qquad \mu_i \equiv \beta_0 + \beta_1\, x_i,$$

donde $x_i = X_i - \bar{X}$ y $\sigma > 0$ es la desviación estándar de los errores.

2) Modelo de parámetros (priores) Elegimos priores que reflejen el entendimiento previo:

- **Intercepto** (promedio de viajes a humedad promedio):

$$\beta_0 \sim \text{Normal}(5000,\ 2000)\,.$$

Justificación: Un desvío estándar de 2000 induce un intervalo central del 95% $\approx 5000 \pm 1.96 \times 2000$, es decir, [1080, 8920], coherente con el rango plausible **1000–9000**.

- **Pendiente** (variación de viajes por punto porcentual de humedad):

$$\beta_1 \sim \text{Normal}(-10,\ 5)\,.$$

Justificación: Un desvío estándar de 5 produce un 95% $\approx [-19.8,\ -0.2]$, que refleja la disminución esperada entre **0** y **20** viajes por punto porcentual (signo negativo).

- **Desviación estándar de los errores**:

$$\sigma \sim \text{Exponential}\big(\lambda = \tfrac{1}{2000}\big)\,, \qquad \sigma > 0.$$

Justificación: Esta prior tiene media $1/\lambda = 2000$, consistente con la "desviación estándar grande" indicada. (Alternativa compatible: $\sigma \sim$ Half-Normal(0, 2000).)

3) Independencias a priori Asumimos independencia a priori entre parámetros:

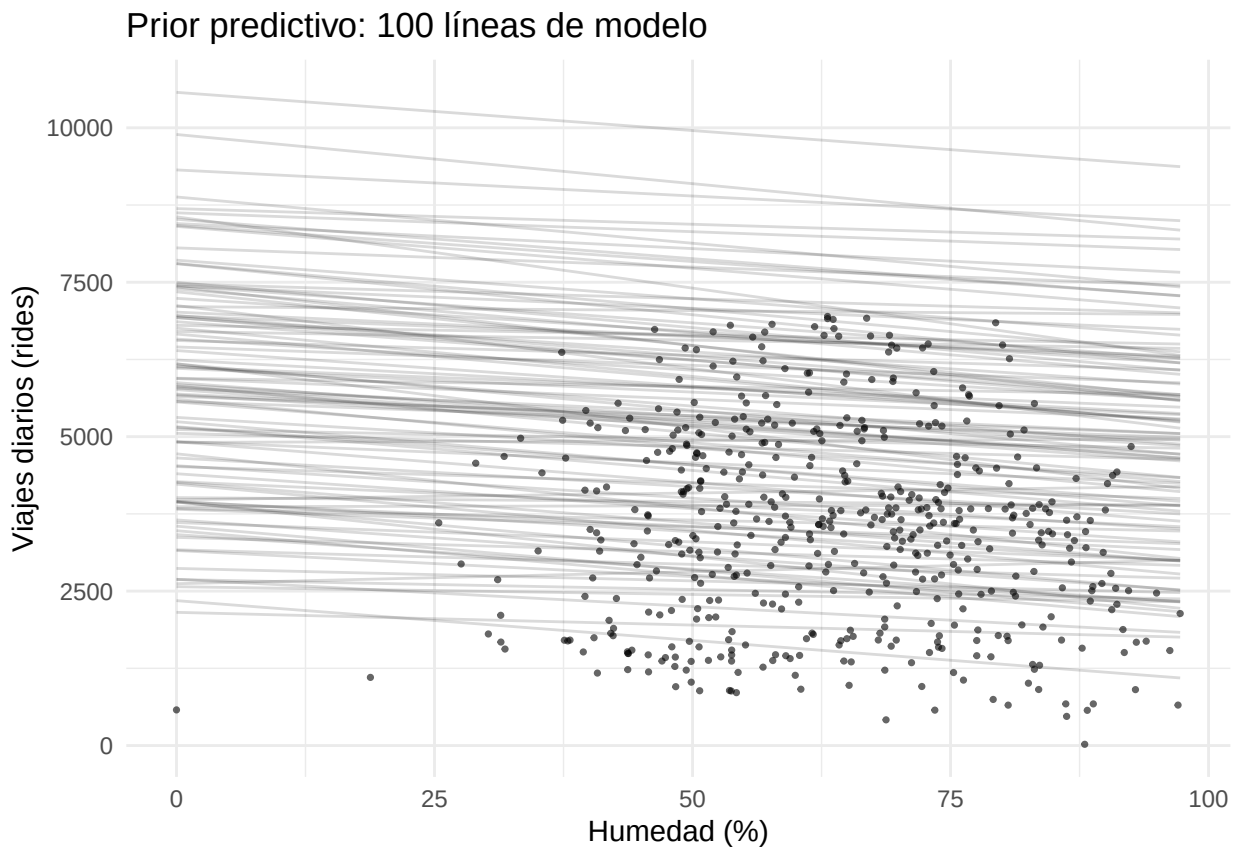
$$(\beta_0,\ \beta_1,\ \sigma) \text{ independientes a priori.}$$

2. Para explorar nuestro entendimiento previo **combinado** sobre los parámetros del modelo, **simulá el modelo previo** de regresión Normal con **5 cadenas** de **8000 iteraciones** cada una.
Pista: podés usar la misma sintaxis de `stan_glm()` que usarías para simular el posterior, pero con `prior_PD = TRUE`.

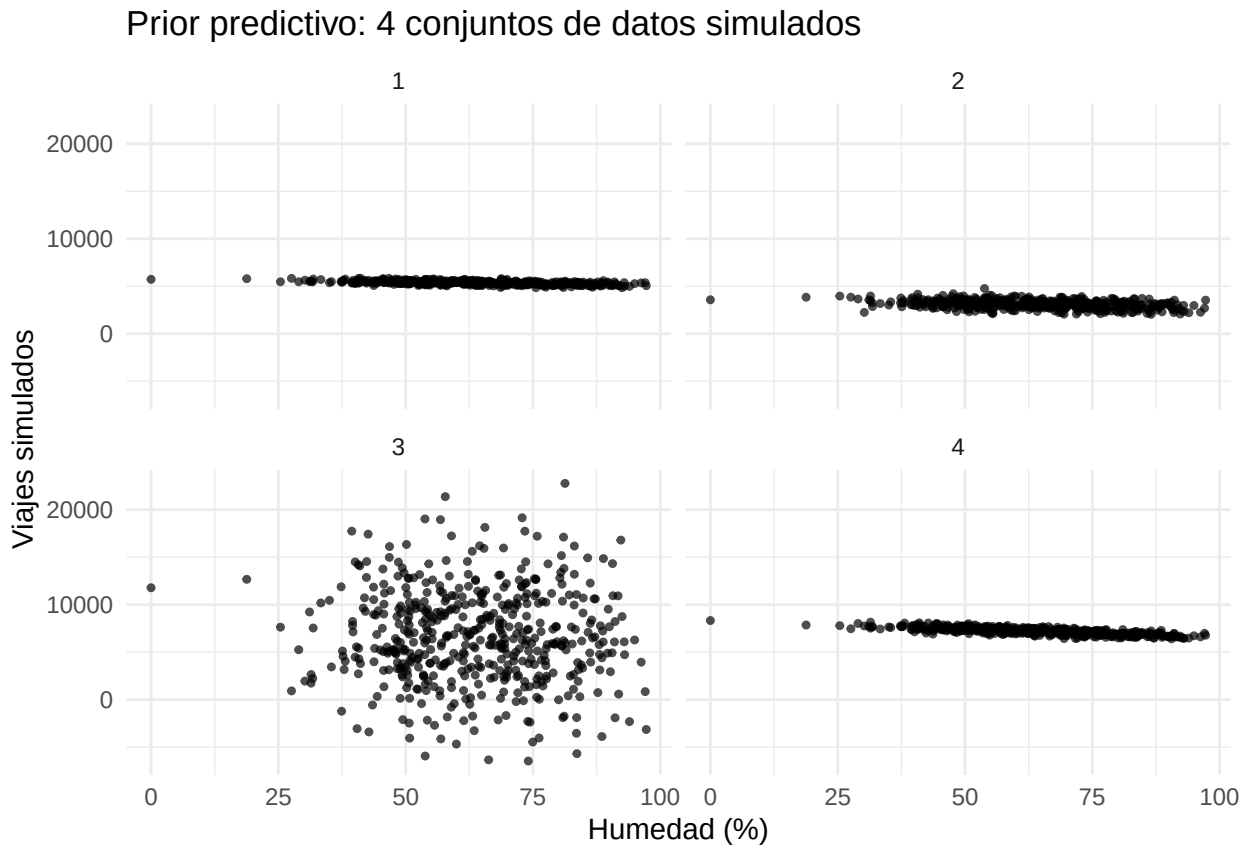
```
data(bikes)
bikes$humidity_c <- with(bikes, humidity - mean(humidity, na.rm = TRUE))
bike_model_prior <- stan_glm(rides ~ humidity_c, data = bikes,
                        family = gaussian,
                        prior_intercept = normal(5000, 2000),
                        prior = normal(location = -10, scale = 5, autoscale = FALSE),
                        prior_aux = exponential(rate = (1/2000)),
                        prior_PD = TRUE,
                        chains = 5, iter = 8000, refresh = 0, seed = 84735)
```

3. **Graficá 100 líneas de modelo previas plausibles** $(\beta_0 + \beta_1 X)$ y **4 conjuntos de datos simulados** bajo los priors.

```
# 100 líneas previas con eje en humedad original (0-100)
bikes %>%
  add_fitted_draws(bike_model_prior, n = 100) %>%
  ggplot(aes(x = humidity, y = rides)) +
  geom_line(aes(y = .value, group = .draw), alpha = 0.15) +
  geom_point(data = bikes, size = 0.6, alpha = 0.6) +
  labs(x = "Humedad (%)", y = "Viajes diarios (rides)",
       title = "Prior predictivo: 100 líneas de modelo") +
  theme_minimal(base_size = 11)
```




```
# 4 datasets simulados bajo los priors (prior predictive)
bikes %>%
  add_predicted_draws(bike_model_prior, n = 4) %>%
  ggplot(aes(x = humidity, y = .prediction)) +
  geom_point(size = 0.9, alpha = 0.7) +
  facet_wrap(~ .draw, nrow = 2) +
  labs(x = "Humedad (%)", y = "Viajes simulados",
       title = "Prior predictivo: 4 conjuntos de datos simulados") +
  theme_minimal(base_size = 11)
```



4. **Describí** nuestro entendimiento previo global de la relación entre uso y humedad.

A partir de los *priors*, nuestro entendimiento previo sobre la relación **uso–humedad** puede resumirse así:

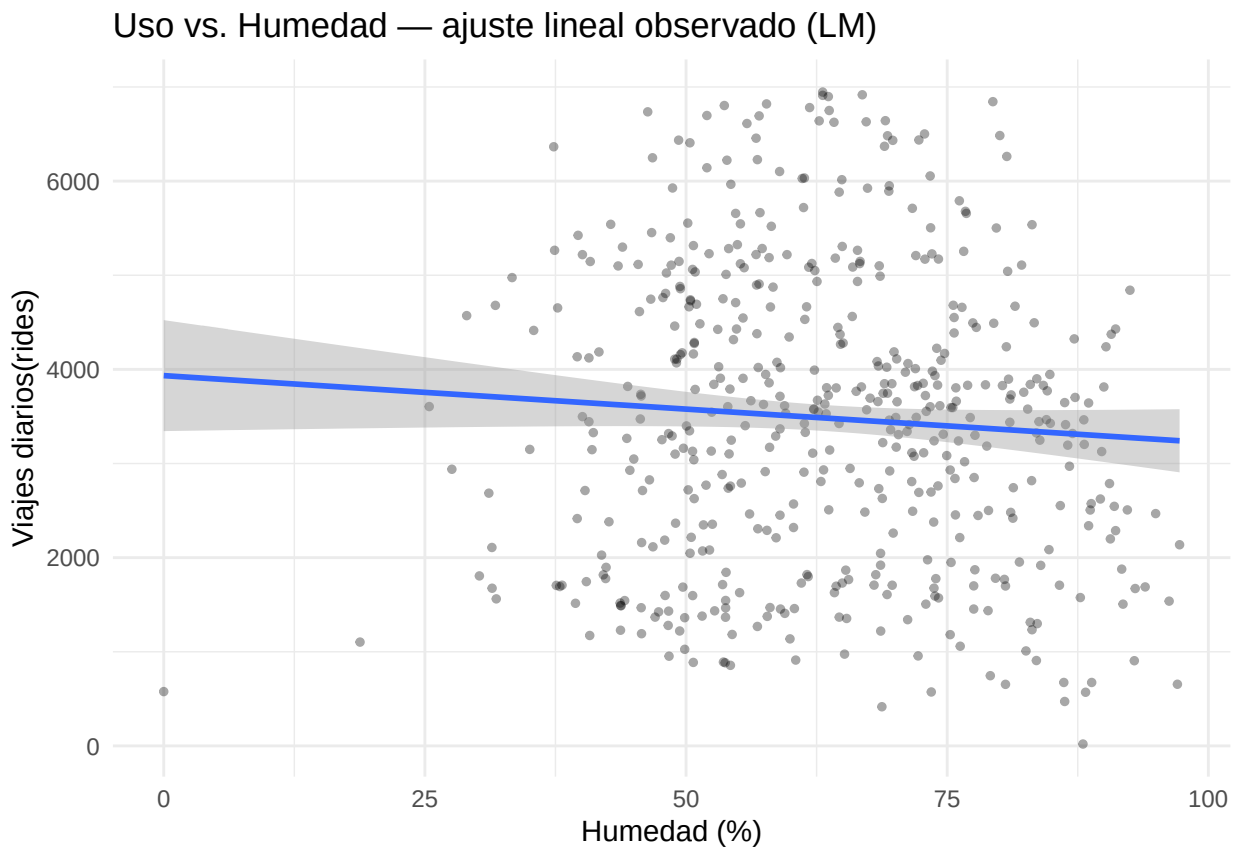
- **Nivel promedio esperado.** Para una humedad **promedio**, esperamos $\beta_0 \approx 5000$ viajes, con gran incertidumbre: En el gráfico de **100 líneas previas** esto se ve como una **banda vertical amplia** de interceptos posibles.
- **Tendencia con la humedad.** La pendiente previa $\beta_1 \sim \text{Normal}(-10, 5)$ sugiere una **disminución suave**: por cada punto porcentual de humedad.
  En todo el rango 0–100, el cambio medio total es del orden de ≈ -1000 viajes, pequeño en relación con la variabilidad del nivel y de los errores; por eso, en la figura, las líneas previas son **casi paralelas** y con **pendiente levemente negativa**.
- **Variabilidad alrededor de la recta.** La desviación estándar de los errores $\sigma \sim \text{Exponential}(\text{rate} = 1/2000)$ tiene **media 2000** pero gran dispersión, por lo que en los **4 conjuntos simulados** aparecen escenarios:
  - **Bandas estrechas** (simulaciones con σ pequeño), donde los puntos se agrupan cerca de la recta.
  - **Nubes muy dispersas** (simulaciones con σ grande), incluso con valores negativos, consecuencia natural del supuesto Normal no acotado. Esto es coherente con la idea de **relación débil** entre uso y humedad.
- **Implicación práctica del prior.**
  1) el uso típico ronda **5000 viajes** en días de humedad promedio,
  2) la **humedad** probablemente **reduce** el uso, pero **de forma moderada**, y
  3) la **variabilidad día a día** es **grande**, por lo que esperamos solapamiento considerable de los niveles de uso entre distintos valores de humedad.

En síntesis, nuestro *prior* favorece una **tendencia negativa suave** con **gran dispersión**, reflejando que la **humedad** es, como mucho, un **moderado** predictor del uso de bicicletas.

Ejercicio 9.10 (¿Qué tan húmedo es *demasiado* húmedo?: datos)
Con los priors definidos, analicemos los datos.

1. **Graficá y discutí** la relación observada entre uso y humedad en `bikes`.

```
ggplot(data = bikes, aes(x = humidity, y = rides)) +
   geom_point(alpha = 0.35, size = 1) +
   geom_smooth(method = "lm", se = TRUE, formula = y ~ x) +
   labs(x = "Humedad (%)", y = "Viajes diarios(rides)",
        title = "Uso vs. Humedad - ajuste lineal observado (LM)") +
  theme_minimal(base_size = 11)
```



2. ¿La **regresión Normal simple** parece razonable para modelar esta relación? **Explicá.**

A partir del gráfico de dispersión y del ajuste lineal observado (`rides ~ humidity`), la **regresión Normal simple** parece ser una **aproximación razonable pero limitada**.
Podemos justificarlo así:

- **Dirección y forma:**
  La tendencia observada es **negativa**, lo que concuerda con la expectativa previa de que un aumento en la humedad se asocia con una ligera disminución del uso de bicicletas.

Visualmente, la relación entre *humedad* y *viajes* es **monótona y aproximadamente lineal**, sin curvaturas fuertes, por lo que una regresión lineal puede capturar la tendencia principal.

- **Dispersión:**
  El **ruido es muy grande** comparado con el efecto medio.
  Es decir, para un mismo nivel de humedad, los viajes diarios varían varios miles de unidades, lo que indica que la humedad **no explica la mayor parte de la variación** en el uso.
  Aun así, el modelo Normal puede describir de forma adecuada esa variabilidad amplia mediante un σ grande.

- **Supuestos de Normalidad y homocedasticidad:**
  El patrón de dispersión no muestra asimetrías graves ni varianza creciente o decreciente con la humedad, por lo que **no hay evidencia fuerte contra la homocedasticidad ni contra la forma Normal de los errores**.
  La principal fuente de variación probablemente provenga de otras covariables omitidas (temperatura, día de la semana, estación, etc.), no de un cambio sistemático en la varianza con la humedad.

Ejercicio 9.11 (¿Qué tan húmedo es *demasiado* húmedo?: simulación posterior)
Ahora podemos **simular el modelo posterior** de la relación entre uso y humedad, balanceando nuestro entendimiento previo y los datos.

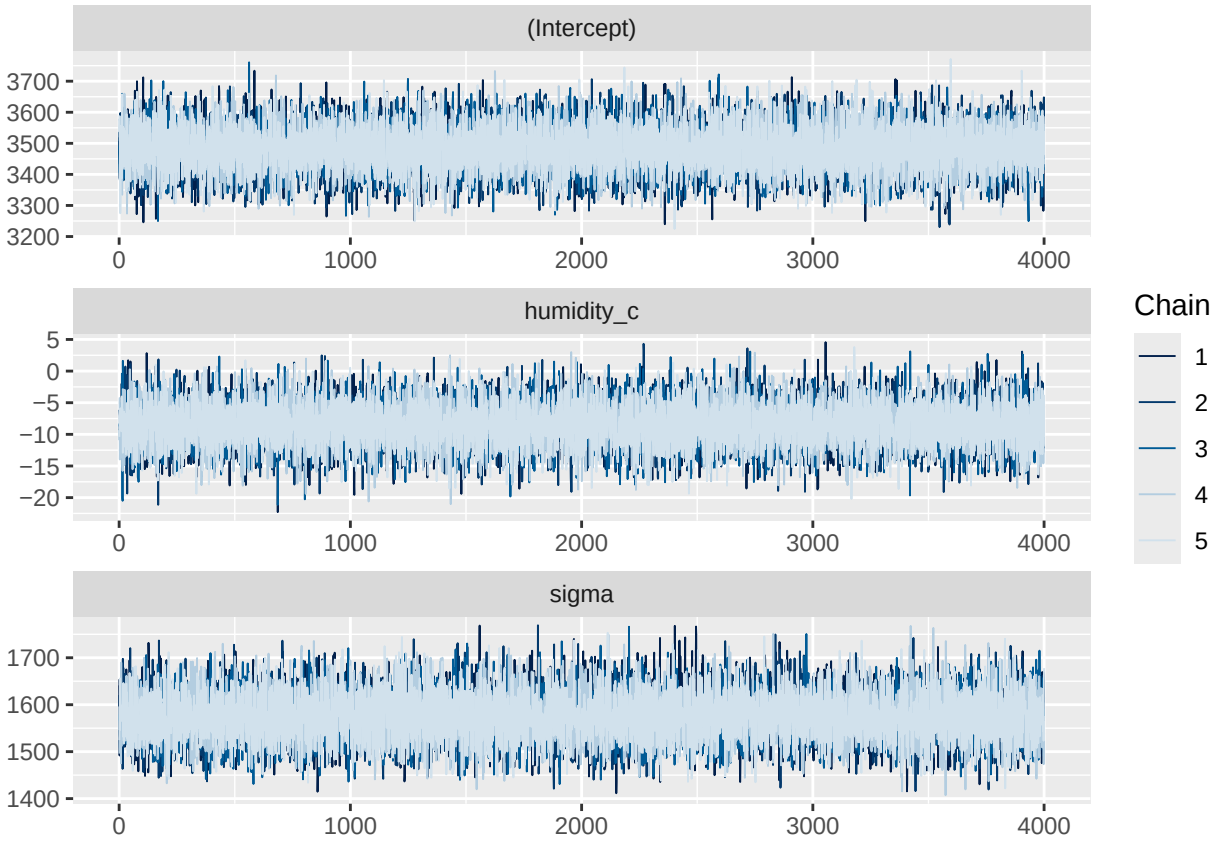
1. Usá `stan_glm()` para **simular el modelo posterior** de regresión Normal, con **5 cadenas** de **8000 iteraciones** cada una.
   Pista: podés hacerlo desde cero o **`update()`**-ar tu simulación previa del Ejercicio 9.9 usando `prior_PD = FALSE`.

```
bike_model_post <- stan_glm(rides ~ humidity_c, data = bikes,
                        family = gaussian,
                        prior_intercept = normal(5000, 2000),
                        prior = normal(location = -10, scale = 5, autoscale = FALSE),
                        prior_aux = exponential(rate = (1/2000)),
                        prior_PD = FALSE,
                        chains = 5, iter = 8000, refresh = 0, seed = 84735)
```

2. Realizá y discutí **diagnósticos MCMC** para determinar si podemos "**confiar**" en estos resultados de simulación.

```
# Convertimos el modelo a formato draws-array
draws_post <- as.array(bike_model_post)
```

```
mcmc_trace(draws_post, pars = c("(Intercept)", "humidity_c", "sigma"),
           facet_args = list(ncol = 1))
```



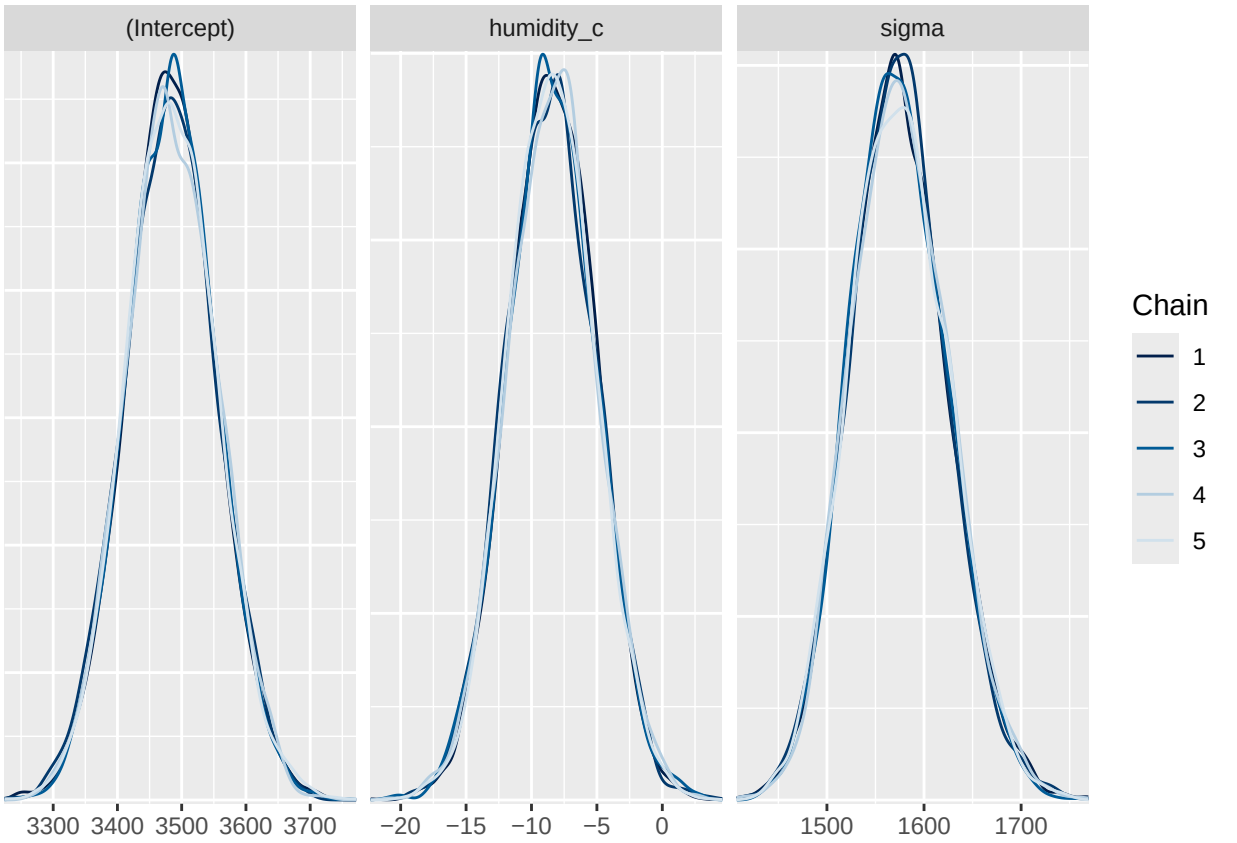
Trazas (`mcmc_trace`)

Buscamos que las cadenas:

- se mezclen bien (no queden separadas por color),
- muestren un patrón tipo "ruido aleatorio" sin tendencia temporal,
- no presenten "pegotes" ni cambios abruptos.

Si las 5 cadenas oscilan alrededor del mismo rango y parecen estacionarias, la **convergencia visual es buena**. Como es este caso.

```
mcmc_dens_overlay(draws_post, pars = c("(Intercept)", "humidity_c", "sigma"))
```



Densidades por cadena

Las curvas de las distintas cadenas deberían superponerse completamente. En este ejemplo se aprecia que se superponen en su totalidad. Si una cadena queda desplazada respecto a las demás, hay evidencia de **no convergencia** o de **mala mezcla**.

```
rhat_vals <- rhat(bike_model_post)
neff_vals <- neff_ratio(bike_model_post)
data.frame(Rhat = rhat_vals, Neff_ratio = neff_vals)
```

```
##                  Rhat Neff_ratio
## (Intercept) 0.9998643    0.95810
## humidity_c  1.0000971    1.08365
## sigma       1.0000012    0.92220
```

$\widehat{R}$ (R-hat)

Debe estar muy cerca de 1 para todos los parámetros.
Valores orientativos:

- $\widehat{R} < 1.01$: excelente convergencia.
- $1.01 \leq \widehat{R} < 1.05$: aceptable, pero conviene revisar.
- $\widehat{R} \geq 1.05$: no confiable, se recomienda aumentar iteraciones o ajustar `adapt_delta`.

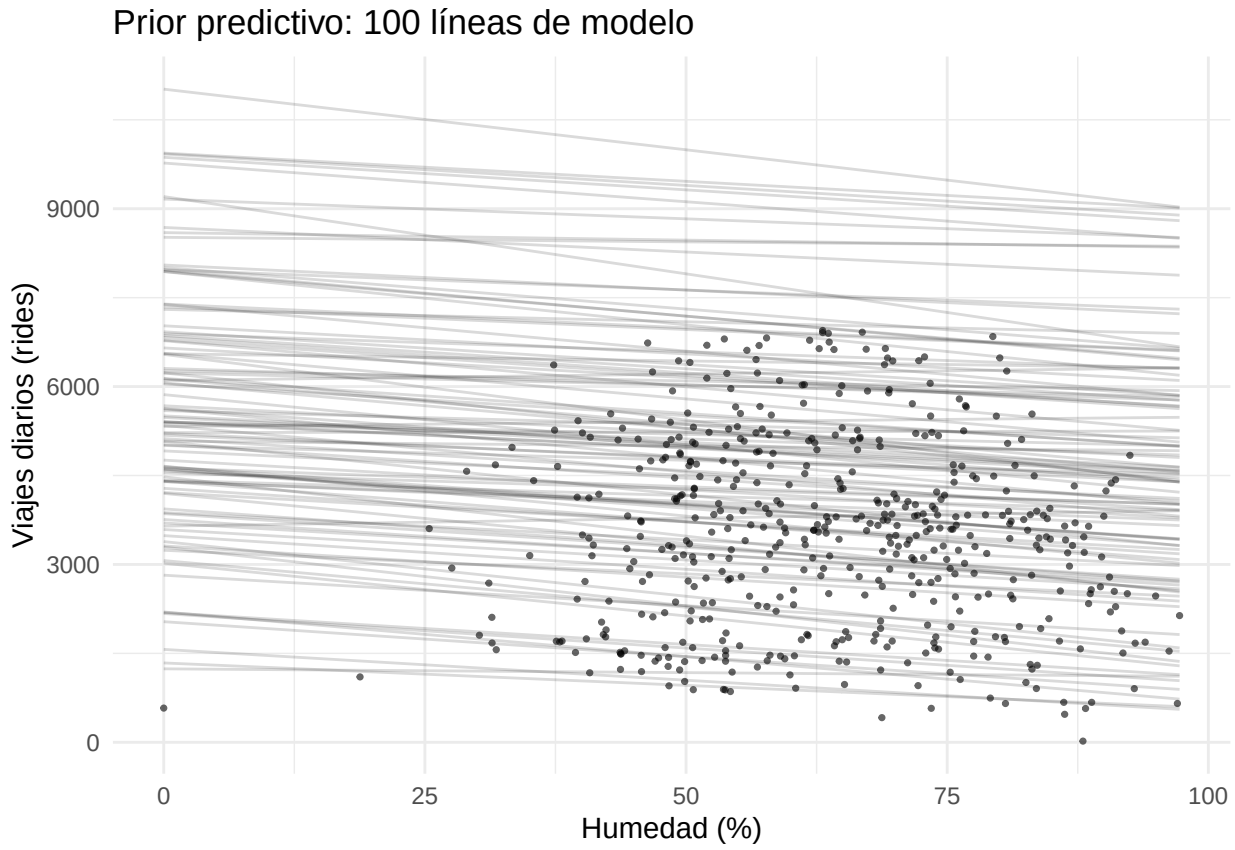
---

n_{eff} / **iteración**

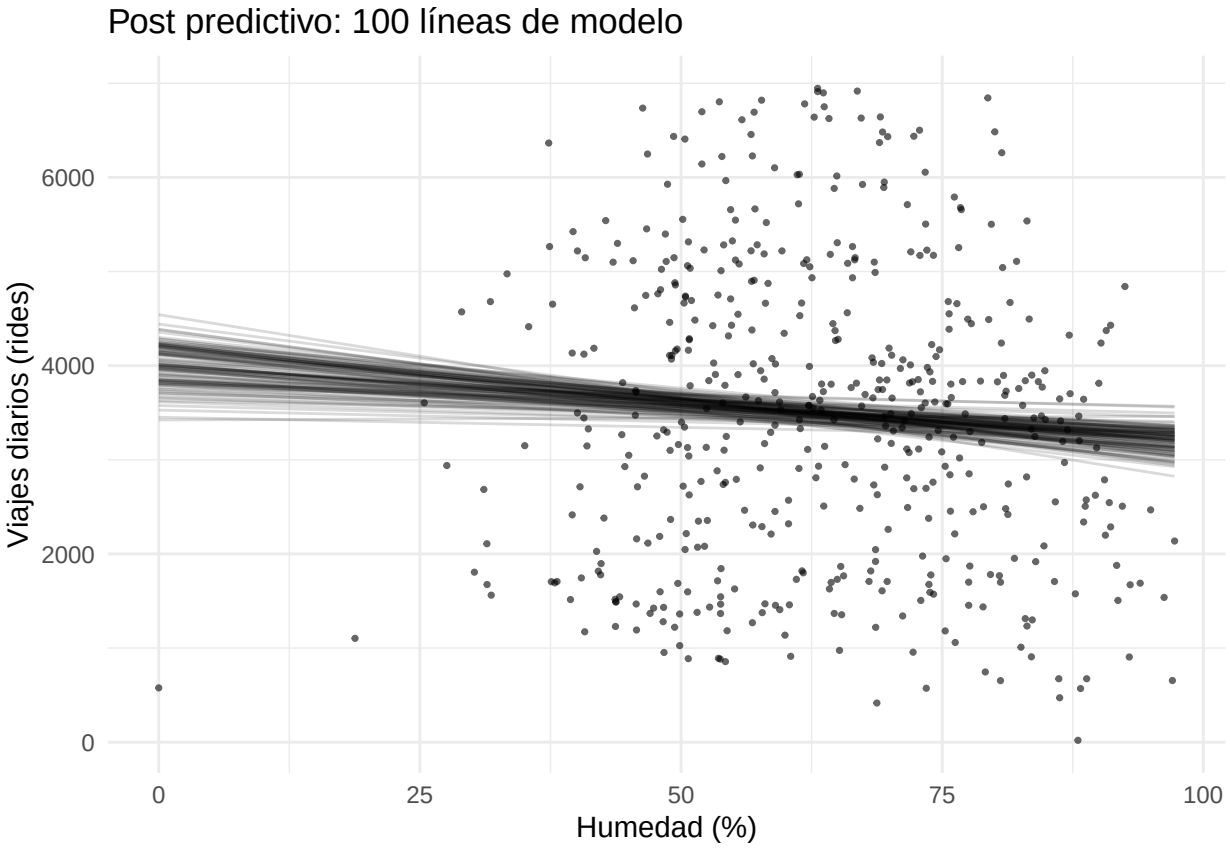
Representa cuán **independientes** son las muestras.
Una razón efectiva mayor a 0.1–0.2 suele indicar que tenemos suficientes *draws*; valores muy bajos implican **alta autocorrelación**.

3. **Graficá 100 líneas de modelo posteriores** para la relación entre uso y humedad. **Compará y contrastá** con las líneas del **modelo previo** del Ejercicio 9.9.

```
bikes %>%
  add_fitted_draws(bike_model_prior, n = 100) %>%
  ggplot(aes(x = humidity, y = rides)) +
  geom_line(aes(y = .value, group = .draw), alpha = 0.15) +
  geom_point(data = bikes, size = 0.6, alpha = 0.6) +
  labs(x = "Humedad (%)", y = "Viajes diarios (rides)",
       title = "Prior predictivo: 100 líneas de modelo") +
  theme_minimal(base_size = 11)
```




```
bikes %>%
  add_fitted_draws(bike_model_post, n = 100) %>%
  ggplot(aes(x = humidity, y = rides)) +
  geom_line(aes(y = .value, group = .draw), alpha = 0.15) +
  geom_point(data = bikes, size = 0.6, alpha = 0.6) +
  labs(x = "Humedad (%)", y = "Viajes diarios (rides)",
       title = "Post predictivo: 100 líneas de modelo") +
  theme_minimal(base_size = 11)
```



Diferencia entre *prior* y *posterior*

Modelo previo (*prior predictivo*)

- Se genera **antes de observar los datos reales**.
- Las líneas del modelo provienen únicamente de las **distribuciones a priori** asignadas a los parámetros (intercepto, pendiente y error).
- Su objetivo es comprobar si esas *priors* producen valores **plausibles dentro del rango observado** de la variable respuesta (*viajes diarios*).
- No reflejan ninguna información del conjunto de datos.

Modelo posterior (*post predictivo*)

- Se obtiene **después de incorporar los datos observados** mediante el proceso de **inferencia bayesiana**.
- Las líneas reflejan las combinaciones de parámetros que son **más probables dado los datos y las *priors***, es decir, la **distribución posterior**.
- En consecuencia, las predicciones se ajustan **mucho mejor a la tendencia real** de los puntos.

Comparación de modelos previos y posteriores En el gráfico **previo** (prior predictivo), las 100 líneas de modelo muestran una gran dispersión y pendientes muy diversas.
Esto ocurre porque las distribuciones a priori de los parámetros son amplias y no contienen información sobre la relación real entre *humedad* y *viajes diarios*.
El resultado es un conjunto de líneas que cubren un rango muy extenso de valores, muchas de ellas alejadas de los datos observados.
Este chequeo sirve para verificar que las priors sean **lo suficientemente flexibles**, sin generar predicciones absurdas (ej. valores negativos o demasiado altos).

En cambio, en el gráfico **posterior** (post predictivo), las 100 líneas se agrupan en torno a una pendiente negativa suave.
Esto indica que, tras incorporar la información de los datos, el modelo aprende una relación **ligeramente decreciente** entre la humedad y el uso de bicicletas.
La incertidumbre entre líneas se reduce notablemente, lo que sugiere que la evidencia empírica domina sobre las priors y permite una estimación más precisa.

En resumen:

- El **modelo previo** refleja la incertidumbre antes de ver los datos.
- El **modelo posterior** integra la información observada y produce predicciones más coherentes con la realidad.

Ejercicio 9.12 (¿Qué tan húmedo es *demasiado* húmedo?: interpretación posterior)
Profundicemos en nuestro entendimiento posterior de la relación entre uso y humedad.

1. Proveé un **resumen `tidy()`** de tu modelo posterior, incluyendo **intervalos creíbles del 95%**.
2. **Interpretá** el valor de la **mediana posterior** del parámetro σ.
3. **Interpretá** el **intervalo creíble posterior del 95%** para el coeficiente de humedad, β_1.
4. ¿Tenemos evidencia posterior suficiente de que hay una **asociación negativa** entre uso y humedad? **Explicá.**

Ejercicio 9.13 (¿Qué tan húmedo es *demasiado* húmedo?: predicción)
Se espera **90% de humedad** mañana en Washington, D.C.
¿Qué niveles de uso deberíamos anticipar?

1. **Sin** usar el atajo `posterior_predict()`, **simulá dos modelos posteriores**:
   - el **modelo posterior** para el **número típico de viajes** en días con **90% de humedad**; y
   - el **modelo predictivo posterior** para el **número de viajes mañana**.
2. **Construí, discutí y compará** visualizaciones con **gráficos de densidad** para los **dos** modelos posteriores de la parte (a).
3. Calculá e **interpretá** un **intervalo de predicción posterior del 80%** para el número de viajes **mañana**.
4. Usá `posterior_predict()` para **confirmar** los resultados de tu modelo predictivo posterior del uso de mañana.