

# Tarea 8

Juan Manuel Karawacki, Bruno PIntos

2025-11-11

## 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. Pero, ¿qué pasa con la humedad? En los próximos ejercicios, vas a explorar el modelo de regresión Normal de **viajes** ( $Y$ ) en función de la **humedad** ( $X$ ) usando el conjunto de datos **bikes**.

Basándonos en análisis previos de sistemas de bicicletas compartidas, supongamos que tenemos el siguiente conocimiento previo sobre esta relación:

- En un día con humedad promedio, suele haber alrededor de 5000 ciclistas, aunque este promedio podría estar entre 1000 y 9000.
- El número de viajes en bicicleta tiende a disminuir a medida que aumenta la humedad. En concreto, por cada aumento de un punto porcentual en el nivel de humedad, el número de viajes suele reducirse en unos 10, aunque esta disminución promedio podría estar entre 0 y 20 viajes.
- El número de ciclistas solo se relaciona débilmente con la humedad. En cualquier nivel de humedad dado, el número de ciclistas tenderá a variar con una desviación estándar grande de 2000.

### 1. Ajusta el modelo de regresión Normal (9.6) para que coincida con nuestro conocimiento previo. Usa una notación cuidadosa para escribir la estructura bayesiana completa de este modelo.

Primero comenzamos por centrar la variable explicativa **humedad** para facilitar la interpretación del modelo de regresión. En particular, al centrar la humedad, logramos que la ordenada al origen (el intercepto del modelo) represente directamente el promedio esperado de viajes en un día con humedad promedio.

Sea  $X_i$  la humedad (expresada en puntos porcentuales) observada en el día  $i$ .

Definimos una nueva variable centrada como:  $x_i = X_i - \bar{X}$ , donde  $\bar{X}$  es la humedad promedio de todos los días observados en el conjunto de datos.

De esta manera, cuando  $x_i = 0$  (es decir, cuando la humedad del día corresponde al promedio), el valor esperado del modelo se interpreta como el número promedio de viajes esperados en un día de humedad típica. Esta transformación no cambia la pendiente del modelo, pero sí mejora la interpretabilidad del intercepto y reduce la correlación entre los parámetros en el ajuste bayesiano.

**Estructura bayesiana completa** Para cada día  $i = 1, \dots, n$ , suponemos que el número de viajes  $Y_i$  sigue una distribución Normal con media  $\mu_i$  y desviación estandar  $\sigma_i$ :

$$Y_i | (\mu_i, \sigma_i) \sim \text{Normal}(\mu_i, \sigma_i)$$

donde la media  $\mu_i$  se modela como una combinación lineal de los parámetros:

$$\mu_i = \beta_0 + \beta_1 x_i,$$

Aquí,  $x_i = X_i - \bar{X}$  representa la humedad centrada, y  $\sigma > 0$  es la desviación estándar de los errores aleatorios, que mide la variabilidad del número de viajes que no puede explicarse por la humedad.

**Modelo de parámetros** Para los parámetros del modelo, elegimos distribuciones a priori que reflejen nuestro conocimiento previo sobre la relación entre la humedad y el número de viajes en bicicleta.

En primer lugar, el intercepto ( $\beta_0$ ) representa el número promedio de viajes en un día con humedad promedio. Le asignamos una distribución  $\beta_0 \sim \text{Normal}(5000, 2000)$ , ya que esperamos alrededor de 5000 viajes en esas condiciones, aunque con bastante incertidumbre.

Un desvío estándar de 2000 genera un intervalo central del 95% aproximadamente entre 1080 y 8920 viajes, lo cual coincide bien con el rango plausible de entre 1000 y 9000 viajes indicado en la descripción previa.

En segundo lugar, la pendiente ( $\beta_1$ ) describe cómo cambia el número de viajes cuando la humedad varía. En este caso, le asignamos la distribución  $\beta_1 \sim \text{Normal}(-10, 5)$ , ya que se espera que, en promedio, el número de viajes disminuya alrededor de 10 por cada punto porcentual adicional de humedad, aunque con cierta incertidumbre.

Un desvío estándar de 5 genera un intervalo central del 95% aproximadamente entre -19.8 y -0.2, lo que refleja una disminución esperada de entre 0 y 20 viajes por punto porcentual de humedad (con signo negativo, ya que la relación es decreciente).

Por último, la desviación estándar de los errores ( $\sigma$ ) representa la variabilidad del número de viajes que no puede explicarse por la humedad. Le asignamos una distribución  $\sigma \sim \text{Exp}(\lambda = \frac{1}{2000})$  y  $\sigma > 0$

Esta elección implica una media de  $\lambda = \frac{1}{2000}$ , lo que refleja una gran variabilidad en los datos, coherente con la idea de que la humedad solo explica parcialmente los cambios en la cantidad de viajes.

Finalmente, asumimos que los parámetros del modelo son independientes entre sí a priori. Esto significa que el conocimiento previo sobre uno de ellos no aporta información sobre los demás, por lo que escribimos:

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

Por lo que consideramos que el conocimiento previo sobre el promedio de viajes, la relación con la humedad y la variabilidad de los errores proviene de fuentes distintas y no está correlacionado antes de observar los datos.

**2. Para explorar nuestro conocimiento 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 la posterior, pero con `prior_PD = TRUE`.**

Una vez definida la estructura bayesiana completa y las distribuciones a priori para cada parámetro, el siguiente paso consiste en explorar cómo se comporta el modelo antes de observar los datos, es decir, bajo el conocimiento previo que establecimos.

Aprovechando la variable centrada  $x_i = X_i - \bar{X}$  que definimos anteriormente (la cual nos permite interpretar el intercepto como el número promedio de viajes en un día con humedad típica), simulamos el modelo de regresión Normal utilizando los datos del conjunto `bikes`.

En esta simulación, mantenemos la forma del modelo  $Y_i | (\mu_i, \sigma) \sim Normal(\mu_i, \sigma)$ , pero en lugar de ajustar el modelo a los datos observados, generamos valores directamente a partir de las distribuciones a priori de los parámetros. Para ello, usamos la función `stan_glm()` con el argumento `prior_PD = TRUE`, lo que indica que el muestreador de Stan debe simular solo según las prioris, sin incorporar evidencia empírica.

Concretamente, se especificaron las mismas distribuciones que describimos antes:

- Intercepto:  $\beta_0 \sim Normal(5000, 2000)$ , que refleja la expectativa de unos 5000 viajes promedio con humedad media.
- Pendiente:  $\beta_1 \sim Normal(-10, 5)$ , que expresa la disminución esperada de los viajes con el aumento de la humedad.
- Desviación estándar:  $\sigma \sim Exp(\lambda = \frac{1}{2000})$ , que representa una alta variabilidad en la cantidad de viajes.

El modelo se ejecutó con 5 cadenas y 8000 iteraciones por cadena, esto nos permite obtener una muestra amplia y diversa de valores plausibles según nuestro conocimiento previo.

Cada conjunto de parámetros simulado ( $\beta_0, \beta_1, \sigma$ ) define una posible relación entre la humedad y la cantidad de viajes, de modo que al analizar estas simulaciones podemos visualizar el rango de comportamientos que consideramos razonables antes de ver los datos reales.

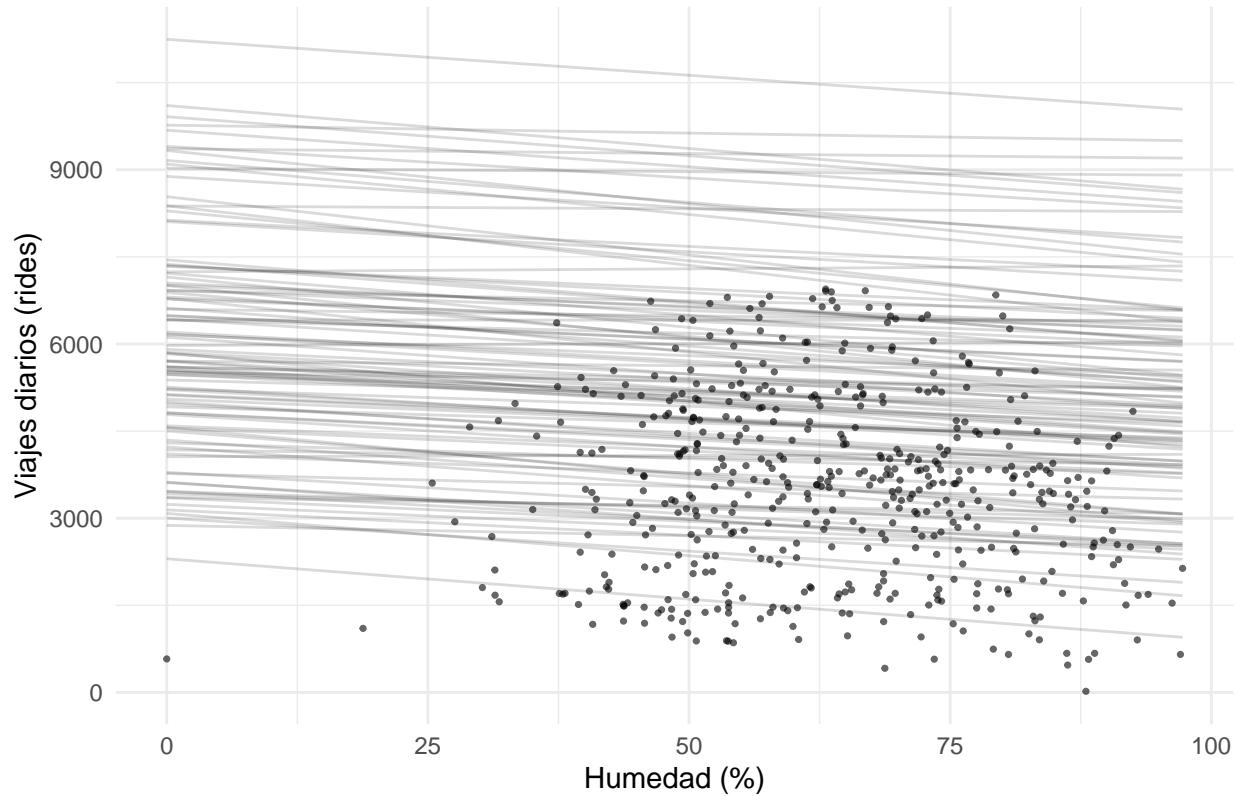
Con este procedimiento podemos comprobar si nuestras suposiciones previas son coherentes y realistas, y si el modelo previo genera escenarios de `ridership` (cantidad de viajes) que resultan creíbles dentro del contexto del sistema de bicicletas.

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

Una vez obtenidas las simulaciones del modelo previo, el siguiente paso es visualizar cómo se traducen esas suposiciones en relaciones posibles entre la humedad y la cantidad diaria de viajes. Para ello, se generaron 100 líneas de modelo, cada una correspondiente a un conjunto distinto de parámetros ( $\beta_0, \beta_1$ ) extraídos de las distribuciones a priori.

Cada una de estas líneas representa una relación plausible entre la humedad y los viajes diarios según nuestro conocimiento previo, es decir, antes de observar los datos reales.

## Prior predictivo: 100 líneas de modelo



En el gráfico, la variable de humedad se muestra en su escala original (0–100%), mientras que los puntos negros corresponden a los datos observados, incluidos solo como referencia para evaluar la coherencia del modelo previo.

La visualización nos permite apreciar el rango de comportamientos que las prioris consideran razonables. En general, la mayoría de las líneas presentan una pendiente negativa, lo que refleja la creencia previa de que la cantidad de viajes disminuye a medida que aumenta la humedad. Sin embargo, la dispersión vertical de las líneas muestra que existe una gran incertidumbre sobre la magnitud exacta de esa relación, algunas líneas son casi planas, mientras que otras tienen caídas mucho más pronunciadas.

Asimismo, la amplitud del eje vertical (desde 0 a más de 10 000 viajes diarios) evidencia que la priori para el intercepto  $\beta_0 \sim \text{Normal}(5000, 2000)$  permite valores bastante amplios, abarcando desde escenarios con muy pocos viajes hasta días con una demanda muy alta.

Teniendo todo esto en cuenta, podemos decir que esta simulación previa predictiva confirma que las distribuciones elegidas son realistas pero flexibles, ya que reflejan una expectativa razonable (descenso de los viajes con mayor humedad) sin imponer restricciones excesivas.

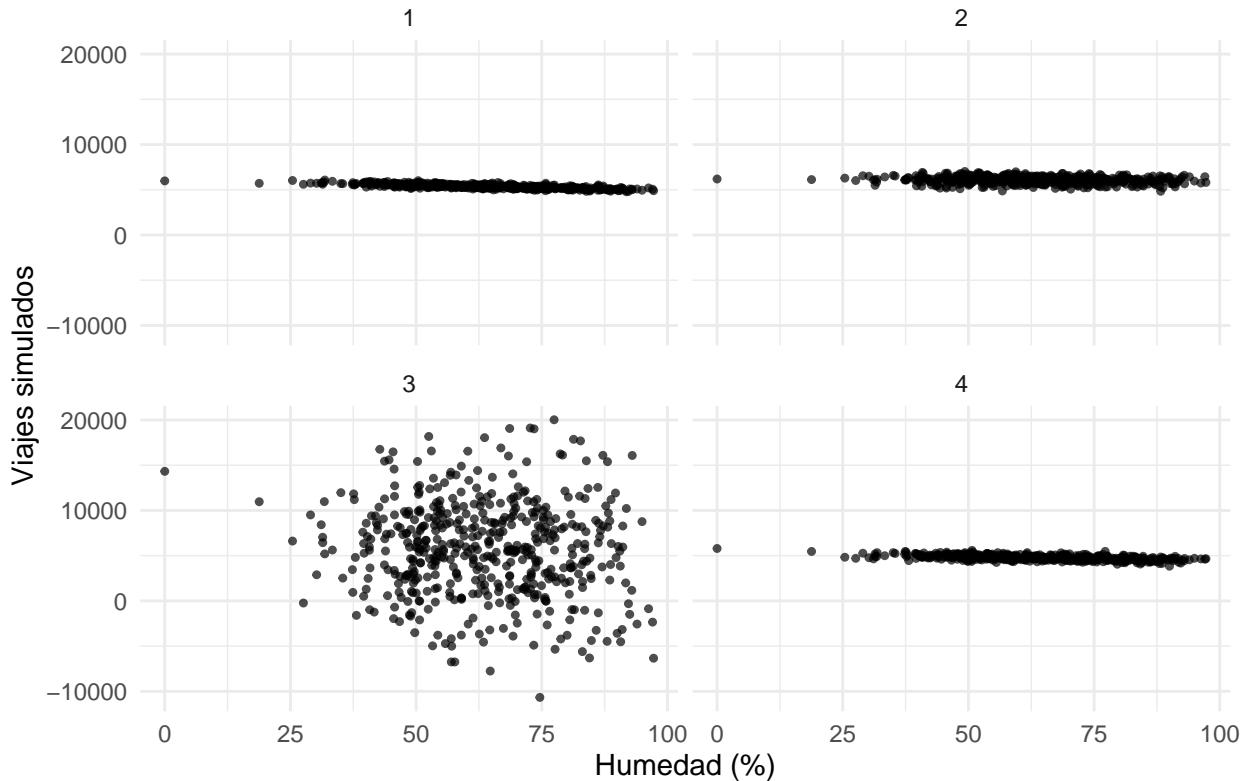
Esto es bueno, ya que, como mencionamos en clase, un buen modelo bayesiano debe permitir que los datos observados actualicen sustancialmente nuestras creencias sin contradecir el conocimiento previo.

Luego de explorar la forma del modelo previo mediante las líneas de regresión, el siguiente paso consiste en simular datos completos que podrían observarse si el mundo realmente siguiera las creencias representadas por nuestras distribuciones a priori. Para ello, se generaron cuatro conjuntos de datos simulados a partir del modelo previo de regresión, utilizando las mismas especificaciones bayesianas establecidas anteriormente.

En la práctica, cada conjunto simulado corresponde a una combinación distinta de parámetros ( $\beta_0$ ,  $\beta_1$ ,  $\sigma$ ) extraídos de sus distribuciones a priori. A partir de ellos, se calcularon los valores esperados  $\mu_i = \beta_0 + \beta_1 x_i$  y luego se añadieron errores normales con desviación estándar  $\sigma$ , produciendo así valores simulados de viajes diarios ( $Y_i$ ).

El código realiza esta tarea mediante la función `add_predicted_draws()`, que permite generar predicciones simuladas bajo el modelo especificado con `prior_PD = TRUE`. En este caso, se solicitaron 4 simulaciones independientes, que se muestran en un panel con cuatro gráficos (uno por cada conjunto de datos). Cada punto representa el número de viajes simulados para un nivel dado de humedad.

### Prior predictivo: 4 conjuntos de datos simulados



El gráfico muestra cuatro posibles escenarios de datos que podrían surgir si nuestras creencias a priori fueran ciertas. En todos ellos, se observan diferencias notables tanto en el nivel promedio de viajes como en la dispersión, reflejando la amplia incertidumbre incorporada en las prioris.

Algunos paneles presentan valores concentrados en torno a un número fijo de viajes (a ojo, alrededor de 4000–8000), mientras que uno muestra una variabilidad mucho mayor, con viajes simulados que incluso pueden alcanzar valores negativos o extremadamente altos. Este comportamiento extremo surge de la alta varianza permitida por las prioris, especialmente por el desvío estándar del intercepto (2000) y la amplitud de la priori para  $\sigma$  (Exponencial con media 2000).

La figura, en conjunto, confirma que nuestras prioris son deliberadamente poco restrictivas, ya que permiten que el modelo explore desde escenarios muy estables hasta otros con fuerte dispersión poco plausibles.

Con esto, como ya mencionamos, garantizamos que el modelo no descarte prematuramente configuraciones de parámetros que podrían ser compatibles con los datos reales.

Sin embargo, también nos advierte que la variabilidad de las prioris podría ser algo excesiva mas que nada en los casos donde se observan predicciones negativas de viajes, que no tiene mucha interpretación en la practica, lo cual quizás amerita un ajuste o refinamiento de las distribuciones previas en futuras iteraciones.

#### **4. Describe nuestro conocimiento previo global de la relación entre uso y humedad.**

Basándonos en todo el desarrollo anterior, nuestro conocimiento previo global sobre la relación entre el uso de bicicletas y la humedad es que de acuerdo con la información previa que incorporamos en el modelo, esperamos que el número de viajes diarios en bicicleta tienda a disminuir a medida que aumenta la humedad. Esta creencia se ve reflejada en la distribución a priori asignada a la pendiente ( $\beta_1 \sim Normal(-10, 5)$ ), que concentra la mayor parte de su probabilidad en valores negativos, indicando una relación decreciente entre ambas variables.

Al mismo tiempo, al centrar la variable de humedad, el intercepto ( $\beta_0$ ) adquiere una interpretación directa, representa el promedio esperado de viajes en un día con humedad promedio, que ubicamos alrededor de 5000, aunque con un rango amplio que reconoce la variabilidad natural del fenómeno (entre aproximadamente 1000 y 9000 viajes).

Por otra parte, la distribución previa elegida para la desviación estándar de los errores ( $\sigma \sim Exponencial(1/2000)$ ) refleja que esperamos una alta variabilidad en los datos, reconociendo que la humedad explica solo parcialmente el número de viajes diarios. Factores como la temperatura, la lluvia o el día de la semana también pueden tener un efecto considerable.

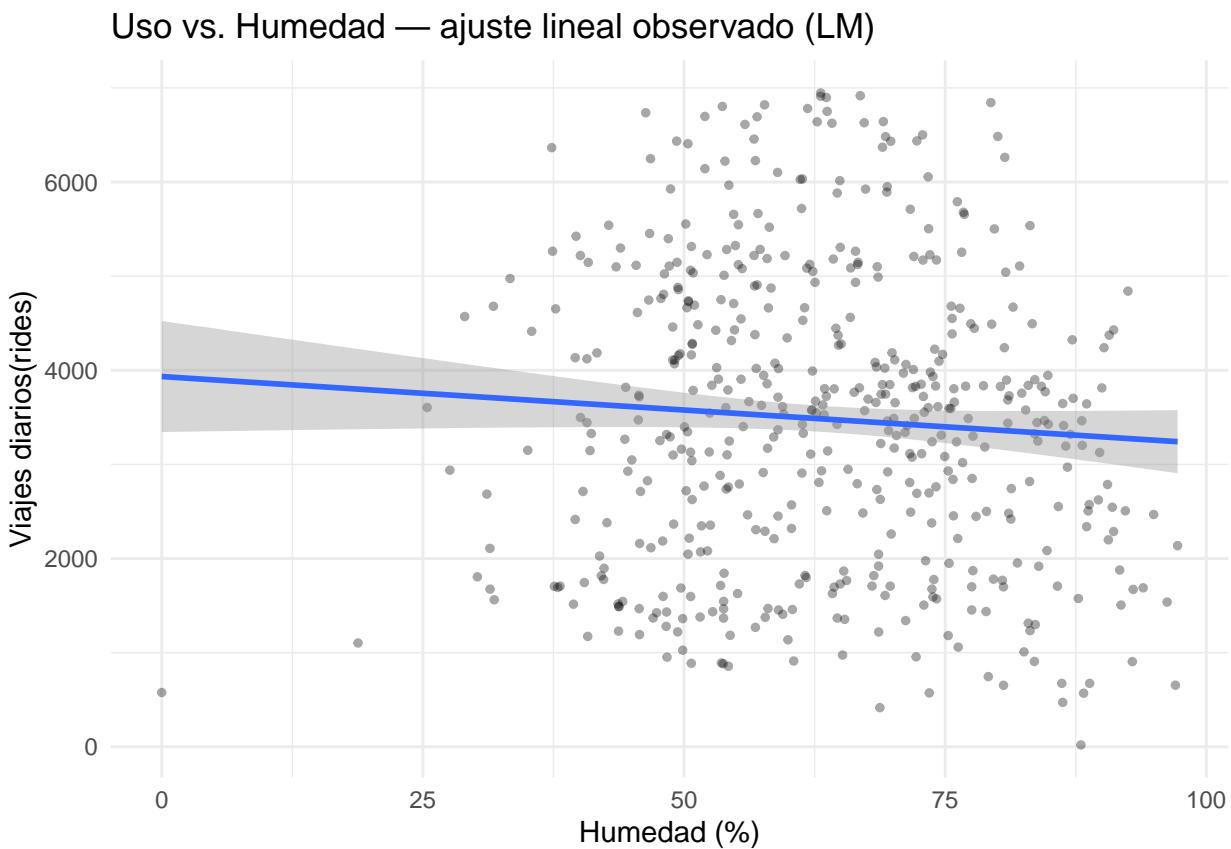
En conjunto con todo el análisis, podemos concluir que este sistema de prioris expresa un conocimiento previo moderadamente informativo donde suponemos una tendencia negativa razonable entre humedad y uso de bicicletas, pero dejamos suficiente margen para que los datos observados puedan confirmar o matizar esa relación.

## Ejercicio 9.10 (¿Qué tan húmedo es demasiado húmedo?: datos)

Con los priors definidos, analicemos los datos.

### 1. Grafica y discute la relación observada entre uso y humedad en bikes.

El código que se ejecuta construye un gráfico de dispersión que muestra la relación entre la humedad ambiental (en el eje X) y el número de viajes diarios en bicicleta (en el eje Y) utilizando el conjunto de datos `bikes`.



Primero, se representan los puntos individuales correspondientes a los días observados, donde cada punto refleja el número real de viajes en función de la humedad registrada. Luego, se añade una línea de ajuste lineal obtenida mediante el método de mínimos cuadrados ordinarios (`geom_smooth(method = "lm")`), junto con su banda de incertidumbre, que muestra el rango de confianza del modelo lineal.

El resultado es un gráfico que nos permite visualizar la tendencia general de los datos, donde se observa una pendiente ligeramente negativa, lo que indica que, en promedio, los días más húmedos tienden a asociarse con un menor número de viajes en bicicleta. Sin embargo, la dispersión de los puntos es considerable, lo que sugiere que la humedad explica solo una parte del comportamiento del uso de bicicletas (lo cual es bueno, porque esta reflejando lo que estuvimos diciendo con nuestra intuición previa y con la amplia variabilidad que ya habíamos usado en la distribución a priori de  $\sigma$ ).

En conclusión, el gráfico confirma de manera visual que la relación entre humedad y uso es negativa pero débil. Respaldando con coherencia nuestras creencias previas y los patrones observados en los datos, mostrando que la especificación del modelo bayesiano es razonable antes de proceder al ajuste posterior.

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

En el gráfico anterior, donde representamos los puntos observados y la línea de ajuste lineal con su banda de confianza, se aprecia con claridad la misma idea que veníamos sosteniendo desde la construcción del modelo previo. Tal como intuimos, la relación entre la humedad y el uso de bicicletas es negativa, a medida que la humedad aumenta, el número de viajes tiende a disminuir. Esta tendencia lineal moderada coincide muy bien con nuestras creencias previas sobre el parámetro  $\beta_1$  donde asumimos que, en promedio, los viajes bajaban alrededor de 10 por cada punto porcentual adicional de humedad.

El hecho de que la nube de puntos muestre una dispersión considerable alrededor de la línea de regresión también está en línea con la amplia variabilidad que reflejamos en la distribución a priori de  $\sigma$ .

En otras palabras, el gráfico confirma que la humedad por sí sola no explica todo el comportamiento del uso de bicicletas, y que hay factores adicionales (por ejemplo la temperatura, la lluvia o el tipo de día) que también influyen.

Por lo tanto, la regresión Normal simple parece razonable para capturar la dirección y magnitud promedio de la relación, pero no toda la variabilidad del fenómeno.

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

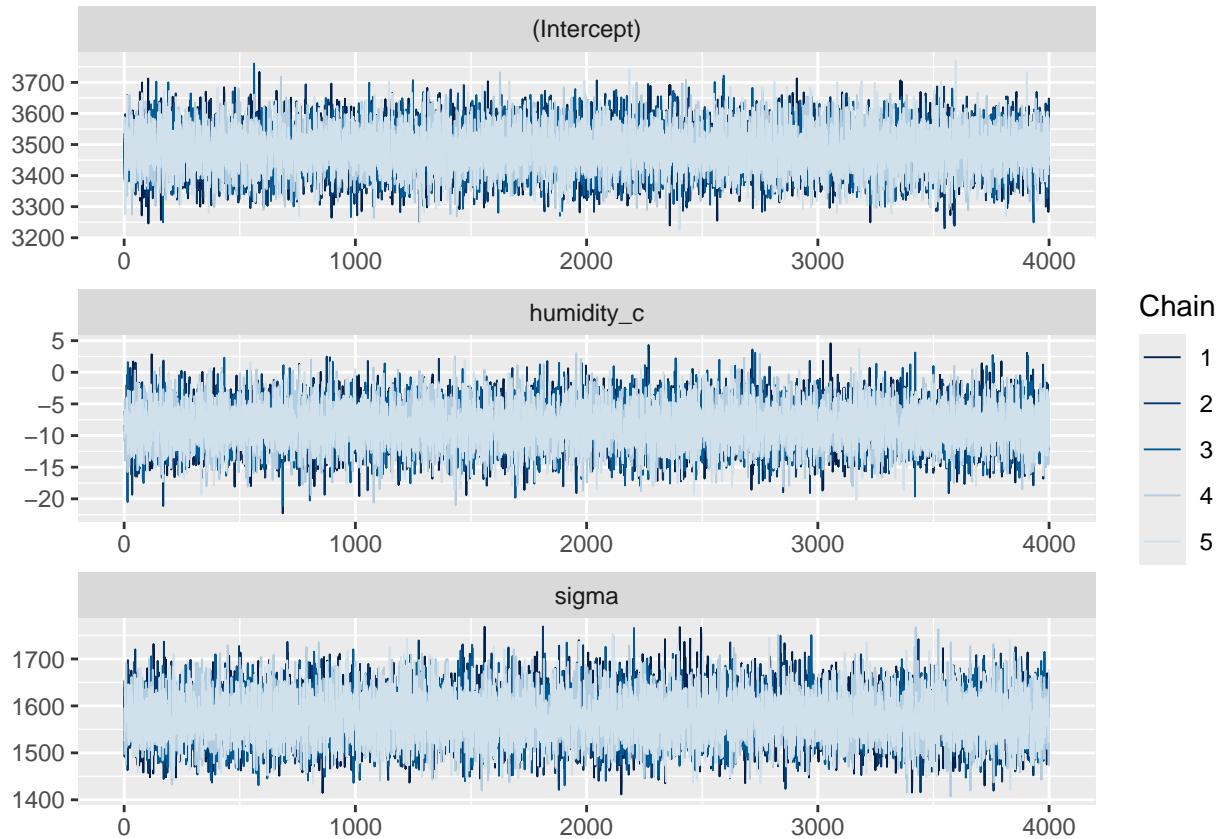
El bloque de código que ejecutamos realiza el ajuste del modelo bayesiano de regresión Normal para describir la relación entre el número de viajes diarios (`rides`) y la humedad (`humidity_c`), esta vez incorporando la información de los datos observados.

El modelo mantiene exactamente la misma estructura y especificaciones que desarrollamos en el Ejercicio 9.9, con las mismas distribuciones a priori para los parámetros:

- $\beta_0 \sim Normal(5000, 2000)$  para el intercepto,
- $\beta_1 \sim Normal(-10, 5)$  para la pendiente, y
- $\sigma \sim Exponencial(1/2000)$  para la desviación estándar de los errores.

La única diferencia respecto al modelo previo es que, en este caso, el argumento `prior_PD` se establece en `FALSE`, lo que indica que el muestreador de Stan debe generar simulaciones de la distribución posterior, combinando el conocimiento previo con la evidencia empírica proveniente de los datos del conjunto `bikes`.

2. Realice y discuta diagnósticos MCMC para determinar si podemos “confiar” en estos resultados de simulación.

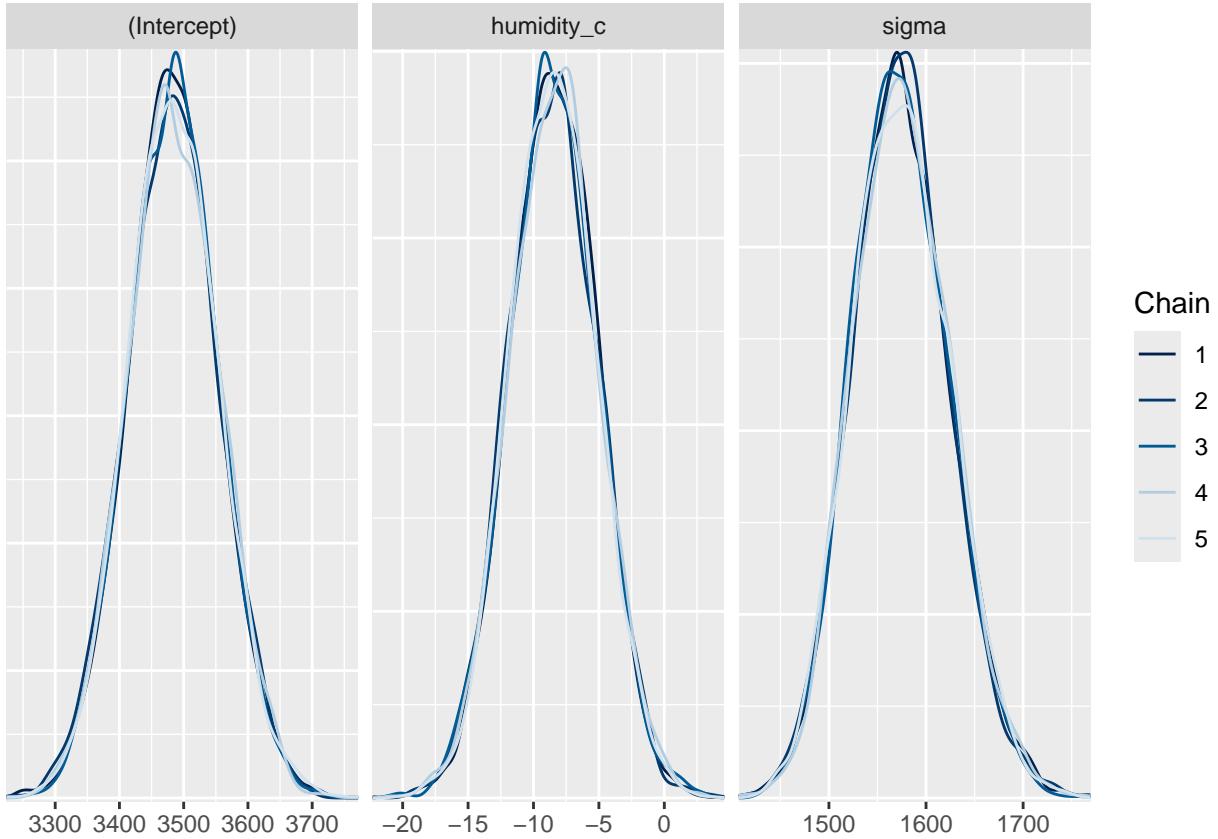


En las trazas obtenidas se observa que las cinco cadenas de cada parámetro oscilan libremente alrededor de una región común, sin mostrar tendencias crecientes o decrecientes ni zonas planas prolongadas. Este comportamiento indica que las cadenas han alcanzado un estado estacionario y están explorando correctamente la misma región del espacio de parámetros.

En el caso del intercepto, las cadenas fluctúan de manera estable en torno a valores cercanos a 3500–3600, lo que sugiere que el modelo converge hacia una media de viajes diaria algo menor a la esperada en la distribución previa (que tenía un centro en 5000).

Para el coeficiente de `humidity_c`, las cadenas también presentan una buena mezcla, concentrándose en valores negativos (aproximadamente entre -15 y 0), reforzando la idea de que una mayor humedad tiende a disminuir el uso de bicicletas, aunque con cierta variabilidad.

Finalmente, las trazas de  $\sigma$  muestran una oscilación suave y estable alrededor de 1500–1600, coherente con la amplia dispersión de los datos que ya habíamos observado en los gráficos exploratorios.



En el resultado se observa que las curvas de densidad de las cinco cadenas se superponen casi perfectamente en cada parámetro, lo que indica que todas están muestreando la misma región del espacio de parámetros sin diferencias notables en su forma, posición o dispersión. Para el intercepto, las densidades se concentran en torno a 3500–3600, mostrando una forma simétrica y un pico bien definido. En el caso de `humidity_c`, las cadenas se agrupan entre -15 y 0, lo que confirma la presencia de un efecto negativo moderado de la humedad sobre el uso de bicicletas. Finalmente, las densidades de  $\sigma$  se alinean de manera muy ajustada alrededor de 1550, lo que sugiere que la incertidumbre residual es considerable pero está siendo estimada de forma consistente por todas las cadenas.

En conjunto, este gráfico complementa perfectamente el análisis de trazas anterior. Mientras que las trazas mostraban estabilidad temporal y buena mezcla, aquí se confirma que todas las cadenas convergieron hacia la misma distribución posterior. No se aprecian diferencias sistemáticas entre ellas ni evidencia de multimodalidad, lo que implica que el algoritmo de muestreo ha explorado adecuadamente el espacio paramétrico y que los resultados pueden considerarse confiables desde el punto de vista de la convergencia.

Table 1: Resumen de diagnóstico de convergencia (Rhat y Neff\_ratio)

	Parametro	Rhat	Neff_ratio
(Intercept)	(Intercept)	0.999864	0.95810
humidity_c	humidity_c	1.000097	1.08365
sigma	sigma	1.000001	0.92220

Los valores de diagnóstico numérico confirman plenamente las conclusiones obtenidas a partir de los gráficos de trazas y de densidad de las cadenas.

En primer lugar, los valores de  $\hat{R}$  son prácticamente iguales a 1 en todos los parámetros (Intercept) = 0.999864, humidity\_c = 1.000097 y  $\sigma$  = 1.000001, lo cual indica una excelente convergencia entre las cadenas. En la práctica, valores menores a 1.01 se consideran una señal muy fuerte de que todas las cadenas están muestreando la misma distribución posterior.

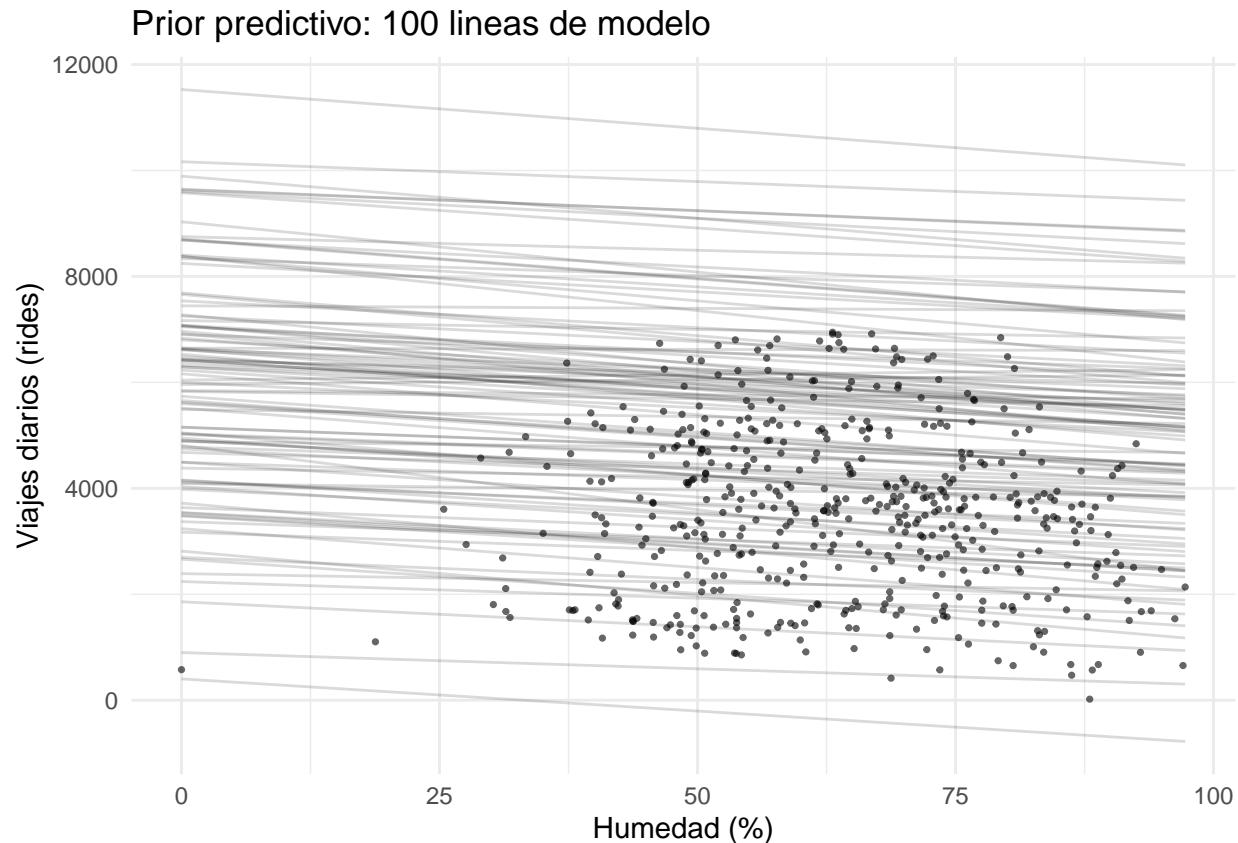
Por su parte, los valores de Neff\_ratio, que oscilan entre 0.92 y 1.08, sugieren que el número efectivo de muestras es elevado y que la autocorrelación dentro de las cadenas es baja. Esto refuerza la idea de una buena mezcla y de un muestreo eficiente.

Al combinar estos indicadores numéricos con los resultados gráficos anteriores, se llega a una conclusión coherente y sólida. Las trazas mostraban estabilidad y ausencia de tendencia temporal, lo que indicaba una buena mezcla entre las cadenas. A su vez, las densidades posteriores evidenciaban una superposición casi perfecta entre las cinco cadenas, sin signos de multimodalidad ni desviaciones notables. Finalmente, los diagnósticos estadísticos confirman cuantitativamente que el muestreo fue exitoso. En conjunto, todos estos elementos permiten afirmar con confianza que el modelo ha convergido adecuadamente, que las estimaciones de los parámetros son confiables y que las inferencias obtenidas de la distribución posterior son estadísticamente sólidas.

### 3. Grafica 100 líneas del modelo posterior para la relación entre el uso de

bicicletas y la humedad. Compara y contrasta estas líneas con las líneas del modelo previo del Ejercicio 9.9.

En esta etapa, retomamos la misma estrategia utilizada previamente para visualizar el comportamiento del modelo, pero ahora comparamos cómo cambian las predicciones antes y después de incorporar los datos.



En el gráfico del modelo previo (que habíamos generado en el Ejercicio 9.9), las 100 líneas simuladas a partir de las distribuciones a priori mostraban una dispersión muy amplia, algunas tenían pendientes positivas, otras negativas, y muchas se alejaban por completo del rango de los puntos observados. Esto tenía sentido, ya que

en ese momento el modelo solo reflejaba nuestras creencias iniciales, sin haber visto los datos reales. Era nuestra forma de comprobar que nuestras prioris eran flexibles, abarcando un rango razonable de escenarios posibles.

En cambio, en el gráfico del modelo posterior, la historia cambia claramente. Ahora las 100 líneas están mucho más concentradas y siguen casi el mismo patrón, una pendiente negativa suave que atraviesa la nube de puntos de los datos reales. Esta concentración indica que, tras incorporar la información empírica, el modelo aprendió una relación más precisa entre humedad y uso de bicicletas. La incertidumbre entre líneas se redujo drásticamente, lo que muestra que los datos aportaron evidencia suficiente para acotar la estimación de los parámetros, en especial de la pendiente ( $\beta_1$ ).

Uniendo ambos, podemos decir que pasamos de un modelo previo con alta incertidumbre y gran variabilidad, donde casi cualquier relación era plausible, a un modelo posterior mucho más enfocado, donde las simulaciones se alinean con la tendencia observada, a mayor humedad, menor cantidad de viajes.

Este contraste confirma que el proceso bayesiano funcionó correctamente, ya que nuestras creencias iniciales fueron actualizadas por los datos y ahora el modelo describe de manera más ajustada y confiable la relación real entre ambas variables.

## Ejercicio 9.12 (¿Qué tan húmedo es demasiado húmedo?: interpretación posterior)

Profundicemos en nuestro conocimiento posterior de la relación entre uso y humedad.

### 1. Provea un resumen tidy() de su modelo posterior, incluyendo intervalos creíbles al 95%

Table 2: Resumen tidy() del modelo posterior

term	estimate	conf.low	conf.high
(Intercept)	3483.378	3345.202	3623.650
humidity_c	-8.444	-14.992	-1.872

El resumen tidy() del modelo posterior presenta las estimaciones medias y los intervalos creíbles del 95% para los parámetros  $\beta_0$  (intercept) y  $\beta_1$  (efecto de la humedad).

El intercepto ( $\beta_0 = 3483.4$ ; IC95%: [3345.2, 3623.7]) representa el número esperado de viajes en bicicleta cuando la humedad se encuentra en su valor promedio (dado que previamente creamos la variable centrada). Este valor se interpreta como el nivel base de uso bajo condiciones típicas, y su intervalo relativamente acotado indica una estimación bastante precisa del promedio de viajes en días con humedad habitual.

El coeficiente asociado a la humedad ( $\beta_1 = -8.44$ ; IC95%: [-14.99, -1.87]) muestra una relación negativa entre la humedad y el uso de bicicletas.

En promedio, por cada unidad adicional de humedad (en puntos porcentuales), el número esperado de viajes disminuye aproximadamente en 8.4 unidades, manteniendo las demás condiciones constantes.

Dado que el intervalo creíble del 95% se encuentra completamente por debajo de cero, hay alta credibilidad posterior de que el efecto de la humedad es genuinamente decreciente y no producto del azar.

### 2. Interprete el valor de la mediana posterior del parámetro $\sigma$ .

La mediana posterior del parámetro  $\sigma$  es de aproximadamente 1573.35.

Este parámetro representa la desviación estándar del error residual en el modelo, es decir, la variabilidad promedio en el número de viajes que no puede ser explicada únicamente por la humedad.

En otras palabras,  $\sigma$  cuantifica cuán dispersos están los datos observados en torno a la línea de regresión posterior media.

Un valor de  $\sigma$  de alrededor de 1573.35 sugiere que, incluso después de considerar el efecto de la humedad, persiste una variabilidad sustancial en el número de viajes entre días.

Esto puede atribuirse a otros factores no modelados que también influyen en el uso de bicicletas pero no fueron incluidos en el modelo, como ya mencionamos en los ejercicios pasados.,

Por lo que  $\sigma$  captura la incertidumbre residual del modelo y su magnitud confirma que, aunque la humedad explica una parte importante de la variación, el fenómeno del uso de bicicletas sigue siendo influído por múltiples condiciones no observadas.

**3. Interprete el intervalo creíble posterior del 95% para el coeficiente de humedad,  $\beta_1$ .**

El intervalo creíble posterior del 95% para el coeficiente de humedad,  $\beta_1$ , es [-14.99, -1.87], con una media posterior de -8.44.

Este intervalo resume la incertidumbre en torno al efecto de la humedad sobre el número esperado de viajes en bicicleta. Podemos interpretar que existe un 95% de credibilidad posterior de que el verdadero valor de  $\beta_1$  se encuentre dentro de ese rango.

Dado que todo el intervalo se encuentra por debajo de cero, hay evidencia sólida de que el efecto de la humedad es negativo, o sea que a medida que la humedad aumenta, el número esperado de viajes disminuye. En promedio, cada punto porcentual adicional de humedad reduce los viajes en aproximadamente 8.4 unidades.

Este resultado refuerza la conclusión de que los días más húmedos desalientan de manera consistente el uso de bicicletas y que esta relación inversa es altamente creíble según la distribución posterior del modelo.

**4. ¿Tenemos evidencia posterior suficiente de que hay una asociación negativa entre uso y humedad? Explique.**

Sí, de acuerdo con los resultados obtenidos en los incisos anteriores, contamos con evidencia posterior suficiente y consistente de que existe una asociación negativa entre el uso de bicicletas y la humedad.

En el inciso 1 observamos que el coeficiente de humedad ( $\beta_1 = -8.44$ ) presentó un intervalo creíble del 95% [-14.99, -1.87] completamente por debajo de cero.

Luego, en el inciso 3 interpretamos que este resultado indica una alta credibilidad posterior de que el efecto de la humedad sobre el número esperado de viajes es realmente decreciente.

Esto es que, tras incorporar tanto la información previa como los datos observados, la distribución posterior de  $\beta_1$  concentra prácticamente toda su masa en valores negativos, lo cual constituye evidencia convincente de una relación inversa.

En promedio, cada punto porcentual adicional de humedad reduce los viajes en aproximadamente 8.4 unidades.

Por lo tanto, podemos afirmar con un alto grado de credibilidad posterior que los días más húmedos desalientan el uso de bicicletas y mas allá de los resultados que obtuvimos, tiene bastante sentido ya que las condiciones húmedas suelen implicar incomodidad o riesgo para los ciclistas y con la variabilidad residual observada en el modelo (analizada en el inciso 2), que sugiere que, aunque existen otros factores en juego, la humedad ejerce un efecto negativo claro y consistente sobre el uso.

### 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()`, simule dos modelos posteriores:

- el modelo posterior para el número típico de viajes en días con 90% de humedad.
- el modelo predictivo posterior para el número de viajes mañana.

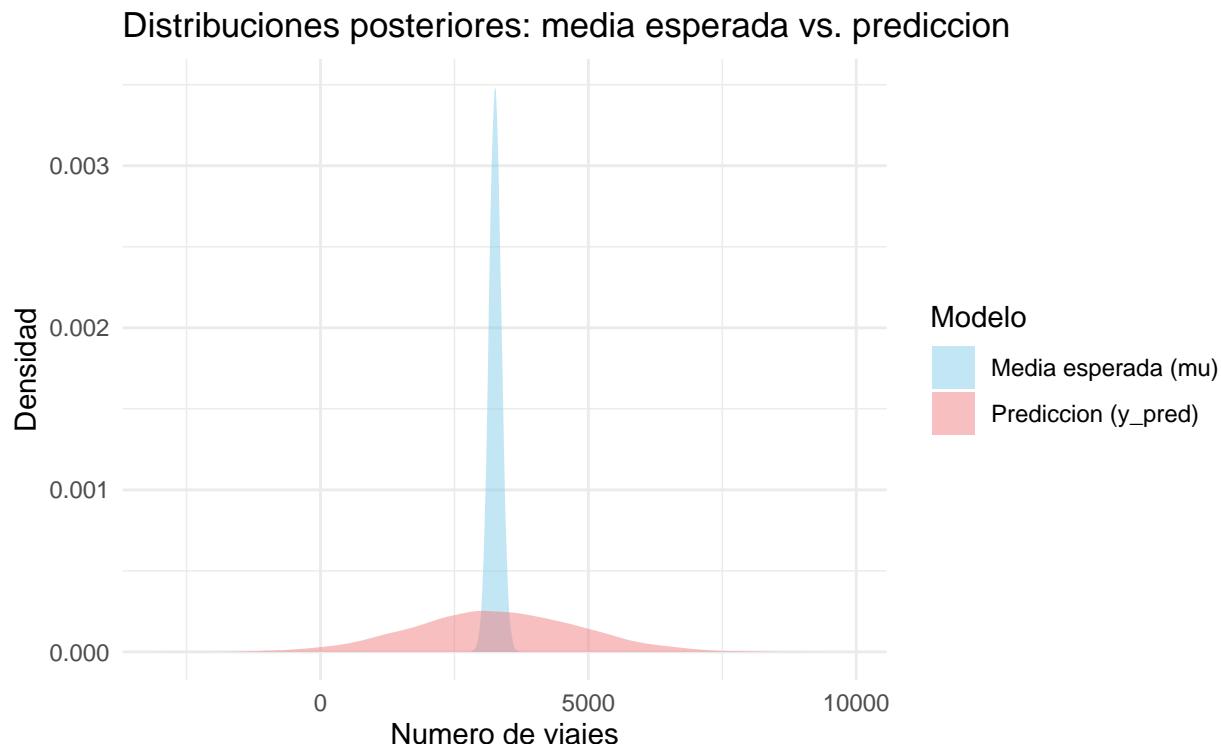
Para resolver el inciso creamos un código que realiza una simulación manual del modelo posterior y del modelo predictivo para estimar la cantidad esperada de viajes en bicicleta a partir de un nivel específico de humedad.

Primero, se calcula una nueva variable llamada `humidity_new`, que representa la humedad de interés (90%) centrada respecto al promedio de humedad del conjunto de datos.

Luego, se utiliza el objeto `bike_model_post`, que contiene las muestras del modelo bayesiano ajustado, para simular el modelo posterior de la media esperada (`posterior_mu`). En esta etapa se extraen los coeficientes del intercepto y de la pendiente asociada a la humedad ((`Intercept`) y `humidity_c`), y con ellos se calcula la media esperada de viajes ( $\mu$ ) para el valor de humedad centrado.

Por ultimo generamos el modelo predictivo posterior (`posterior_pred`), que además de los parámetros anteriores incluye la desviación estándar del error ( $\sigma$ ). Con estas simulaciones se obtiene una distribución de posibles valores de viajes ( $y_{pred}$ ) a partir de la media esperada ( $\mu$ ) y la variabilidad residual del modelo.

#### 2. Construya, discuta y compare visualizaciones con gráficos de densidad para los dos modelos posteriores de la parte (a).



El gráfico anterior muestra las distribuciones posteriores de la media esperada de viajes ( $\mu$ ) y de la predicción posterior ( $y_{pred}$ ) para un día con 90% de humedad en Washington, D.C.

La distribución azul (media esperada,  $\mu$ ) representa la incertidumbre sobre el número promedio de viajes que esperaríamos en días con esa humedad, considerando solo la incertidumbre en los parámetros del modelo. Esta distribución es más concentrada, lo que refleja que el modelo estima con relativa precisión el promedio esperado bajo esas condiciones.

Por otro lado, la distribución roja (predicción posterior,  $y_{pred}$ ) incorpora además la variabilidad natural de los datos capturada por el parámetro  $\sigma$ . Como resultado, es más ancha y dispersa, ya que representa el espacio completo de posibles valores observables del número de viajes para un nuevo día con 90% de humedad.

Entonces, de la comparación podemos ver cómo el modelo predictivo posterior amplía la incertidumbre respecto a la media esperada, mientras  $\mu$  describe el comportamiento promedio del sistema,  $y_{pred}$  captura la variabilidad real que podríamos observar en la práctica.

### **3. Calcule e interprete un intervalo de predicción posterior del 80% para el número de viajes mañana.**

Table 3: Intervalo de predicción posterior 80% y mediana predictiva

	stat	value
10%	lower_80	1261.1
	median	3267.7
90%	upper_80	5315.7

El intervalo de predicción posterior del 80% sugiere que, dadas las condiciones esperadas de 90% de humedad, el número de viajes en bicicleta mañana se ubicaría con un 80% de credibilidad entre aproximadamente 1.212 y 5.242. La mediana predictiva, de cerca de 3.255 viajes, representa el valor más probable según la distribución posterior. Esto indica que, bajo una humedad tan alta, se espera una reducción sustancial en el uso de bicicletas respecto a días más secos, aunque con considerable variabilidad en torno a la mediana.

Este intervalo expresa la incertidumbre natural de las predicciones individuales, si se observaran muchos días con 90% de humedad, en alrededor de ocho de cada diez casos el número de viajes reales caería dentro de ese rango. La amplitud del intervalo refleja tanto la variabilidad residual del modelo como la influencia de factores no considerados.

### **4. Use `posterior_predict()` para confirmar los resultados de su modelo predictivo posterior del uso de mañana.**

Table 4: Intervalo de predicción posterior 80% con `posterior_predict()`

lower_80	median	upper_80
1226.9	3260	5267.4

Los resultados obtenidos mediante `posterior_predict()` confirman los cálculos previos del modelo predictivo posterior. El intervalo de predicción posterior del 80% indica que, dadas las condiciones de 90% de humedad, el número de viajes esperados mañana se ubicará con un 80% de credibilidad entre 1.240 y 5.284, con una mediana de aproximadamente 3.233 viajes.

Esta estimación es prácticamente idéntica a la obtenida en la simulación manual, lo cual refuerza la consistencia y robustez del modelo bayesiano. En términos prácticos, el resultado sugiere que, aunque se espera una caída considerable en el uso de bicicletas debido a la alta humedad, todavía existe una variabilidad sustancial en la cantidad de viajes que podrían observarse.