Tarea 8

Santiago Robatto y Sofia Terra

Ejercicio 9.9 Bayes Rules!

Exercise 9.9 (How humid is too humid: model building) Throughout this chapter, we explored how bike ridership fluctuates with temperature. But what about humidity? In the next exercises, you will explore the Normal regression model of rides (Y) by humidity (X) using the bikes dataset. Based on past bikeshare analyses, suppose we have the following prior understanding of this relationship: Prior understanding of this relationship:

- On an *average* humidity day, there are typically around 5000 riders, though this average could be somewhere between 1000 and 9000.
- Ridership tends to decrease as humidity increases. Specifically, for every one percentage point increase in humidity level, ridership tends to decrease by 10 rides, though this average decrease could be anywhere between 0 and 20.
- Ridership is only weakly related to humidity. At any given humidity, ridership will tend to vary with a large standard deviation of 2000 rides.

En primer lugar, procederemos a entender quién es cada una de las variables del modelo de regresión normal simple para este ejercicio.

Recordemos que el modelo definido en el libro es:

$$Y_i = \beta_0 \ + \beta_1 X_i \ + \epsilon_i$$

Por ende, tenemos:

- Y: Es la variable de respuesta, es decir, la que buscamos modelar. Al igual que en todo el capitulo 9 del libro, la variable de respuesta es la cantidad de viajes realizados por bicicletas en un dia.

- X: Es la variable explicativa, la cual sera utilizada para intentar explicar Y. En este caso dejaremos de tomar la temperatura como variable explicativa y comenzaremos a usar la humedad. A simple intuicion, lo logico es pensar que esta nueva variable tendra un comportamiento decreciente, es decir que a mayor humedad menor cantidad de viajes y viceversa. Ademas, personalmente pensamos que es probable que sea menos fuerte o explicativa que la variable utilizada anteriormente,
- β_0: Se le denomina coeficiente de intercepto. Esto seria, el valor esperado de viajes cuando la humedad vale 0
- β_1: Matematicamente hablando es la pendiente de la recta. Es decir,en este contexto representa como cambia la cantidad de viajes en funcion de los cambios de la humedad.
- ϵ: Representa el error del modelo. Para ello, asumiremos que distribuye normal, con μ=0 y varianza σ^2. Seremos nosotros quienes simularemos σ con un modelo exponencial.

En resumen:

$$\epsilon_i \sim \text{Normal}(0, \sigma)\sigma \sim \text{Exponential}(\lambda)$$

De esta manera, para ajustar el modelo debemos trabajar sobre tres parametros: β_0, β_1 y σ.

Ridership tends to decrease as humidity increases. Specifically, for every one percentage point increase in humidity level, ridership tends to decrease by 10 rides, though this average decrease could be anywhere between 0 and 20 $\longrightarrow$ Nos habla de como varia la cantidad de viajes ante cambios en la humedad, por ende con dicha informacion podemos definir β_1. Confirma nuestra teoria inicial de que la relacion es decreciente, y a su vez se observa que es un valor bastente disperso entorno al centro, es decir que hay alta variabilidad, y por ende la correlacion no sera sumamente marcada.

Para definir una varianza coherente en el modelo, utilizaremos la siguiente "regla" del modelo normal:

El $\text{IC}_{95\%}$ de la normal coincide aproximademente con $\mu\pm$ 2 desviaciones estandar. De manera que:

$$\text{Rango plausible}_{95\%} \approx \mu \pm 2\sigma$$

Por ende, despejando, σ=5. Es decir que definiremos $\beta_1 \sim \text{Normal}(-10, 5)$

Podemos comprobar si nuestra manera de definir β_1 fue adecuada al texto usando los cuantiles 2.5% y 97.5% de la normmal.

Cuantil	Valor
2.5%	-19.8
97.5%	-0.2

De esta manera nos aseguramos estar modelando como nos indica el texto β_1.

On an average humidity day, there are typically around 5000 riders, though this average could be somewhere between 1000 and 9000. $\longrightarrow$ Nos da informacion sobre β_0, mas precisamente es exactamente lo que comentamos anteriormente, nos da la cantidad de viajes para un dia promeido de humedad.

La definiremos con el modelo normal, centrada entorno a 5000. Para la varianza, tomaremos el mismo criterio que en el punto anterior, de modo que σ=2000: $\beta_0 \sim \text{Normal}(5.000, 2.000)$

Verificamos el IC 95%:

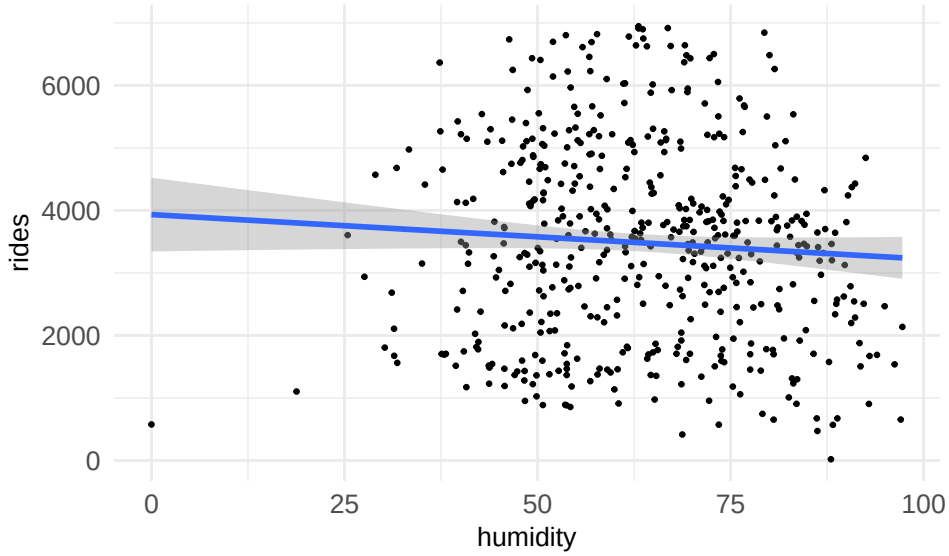
Cuantil	Valor
2.5%	1080.07
97.5%	8919.93

Ridership is only weakly related to humidity. At any given humidity, ridership will tend to vary with a large standard deviation of 2000 rides. $\longrightarrow$ nos habla sobre el error, que sera representado, tal como se realiza en el libro, primero mediante el modelo exponencial para definir σ y luego con el modelo normal.

En el modelo exponencial, el inverso del parametro es igual a la media y a la varianza, por lo que, dado que nuestra varianza es 2000 (dato letra), tomaremos $\sigma = \frac{1}{2000}$

Al graficar los datos de viajes en función de la humedad, se observa una nube de puntos bastante dispersa, sin una relación lineal fuerte (tal como esperabamo).

Aun así, la recta de tendencia ajustada muestra una leve pendiente negativa, lo que sugiere que, en promedio, a medida que aumenta la humedad, la cantidad de viajes tiende a disminuir ligeramente. Por otra parte, la gran dispersión de puntos confirma que la humedad explica solo una pequeña parte de la variabilidad en los viajes, coherente con la idea de que la relación es débil.



Juntando toda esta informacion, ya estamos en condiciones de definir el modelo. Utilizaremos para ello la funcion stan_glm, del paquete rstanarm, la cual, de manera muy resumida, lo que hace es, de forma bayesiana, combina los priors con los datos para simular la distribución posterior de los parámetros mediante MCMC.

Una manera estandar de definir el modelo es la siguiente:

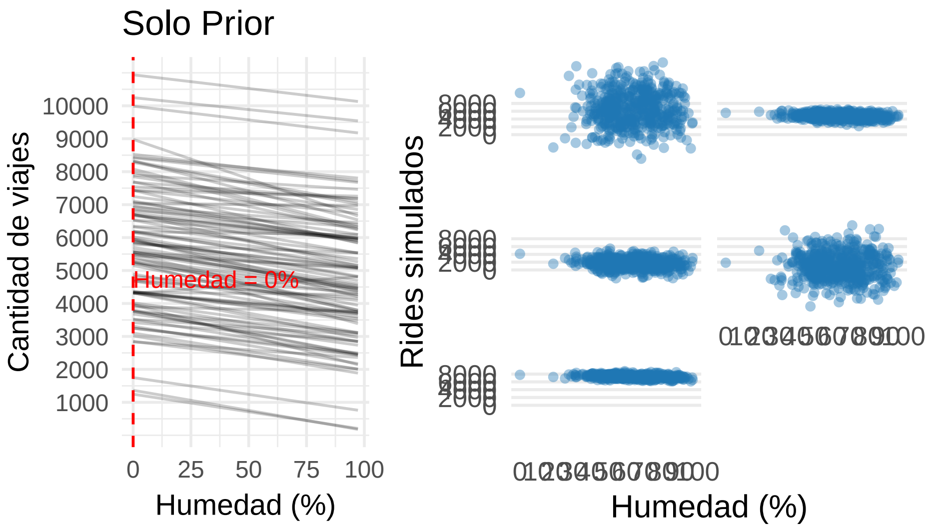
```
bike_model_hum<- stan_glm(rides ~ humidity,  #Rides en funcion de la humedad
                          data = bikes, #Dataset
                          family = gaussian, #Asumir errores bajo modelo normal
                          prior_intercept = normal(5000, 2000), #Definicion del intercepto
                          prior = normal(-10, 5),  #Definicion de B1
                          prior_aux = exponential(1/2000), #Defnicion de sigma (para el error)
                          chains = 5, # Cant. cadenas
                          iter = 8000*2, #Iteraciones por cadena
                          seed = 84735) #Semilla
```

Definida la base del modelo, procederemos a correrlo pero solo a partir de los priors, es decir, sin observar la data observada. Para ello, utilizaremos la funcion update combinada con el argumento prior_pd=TRUE, de manera de asegurarnos de que Stan ignore los datos observados y que genere valores simulados usando solo los priors.

Ajustamos y corremos el modelo:

```
bike_model_prior <- update(bike_model_hum, prior_PD = TRUE)
```

DIOS QUE CHOTO EL P2. ARREGLARLO!!!



El grafico de la izquierda muestra las 100 combinaciones aleatorias del prior (β_0 y β_1) y genera 100 lineas a partir de esos valores tomados de manera aleatoria, donde cada línea es una versión posible de la relación entre humedad y viajes (según nuestros priors).

Claramente se observa que la relacion es negativa. Las pendientes de las lineas tienden a ser parecidas entre si, lo que nos hace pensar que tenemos cierta informacion consistente sobre como es la relacion entre humedad y viajes. Esto confirma que, a priori, esperamos que la humedad tenga un efecto levemente negativo sobre la cantidad de viajes.

Por otro lado, la ordenada en el origen es sumamente variable, oscilando entre 3000 y 9000 la mayoria de puntos de corte. Esto es, en un día promedio de humedad, suele haber entre unas 3000 y 9000 personas, pero podría variar bastante, con valores extremos menores a 1.000 y mayores a 10.000.

Al mirar el gráfi|co de la derecha, los conjuntos simulados bajo los priors muestran una amplia dispersión vertical, lo que refleja nuestra incertidumbre sobre la fuerza de la relación entre humedad y viajes. Los priors, son débilmente informativos y mantienen la incertidumbre en un rango realista: se centran en niveles de viajes del orden de miles, no en valores absurdos o imposibles para el contexto delos viajes en bici. Esto garantiza que nuestro modelo exprese creencias plausibles, sin desviarse hacia valores imposibles.

A su vez, se observa a simple vista que los distintos paneles presentan niveles de dispersión variables. Esto ocurre porque en cada simulación los valores del parámetro σ (la desviación

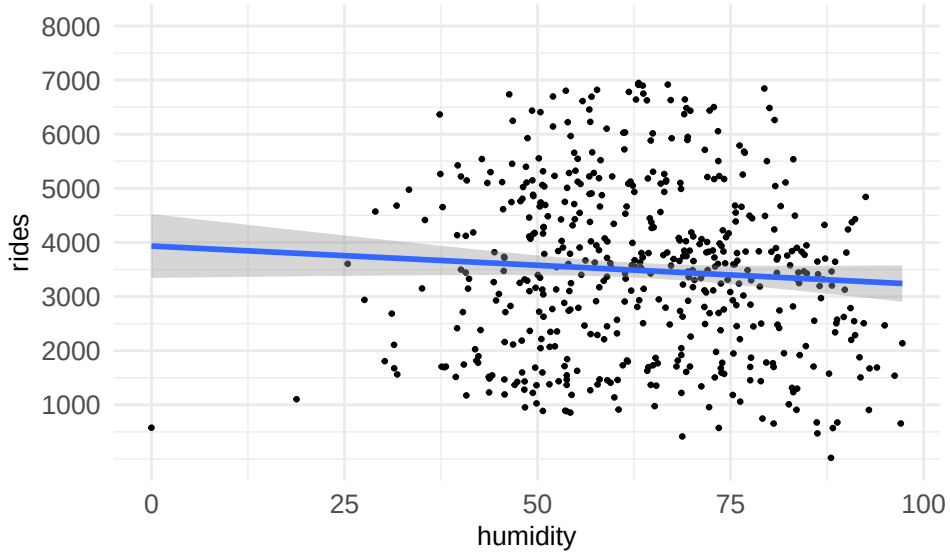
del error) son distintos. Cuando σ es pequeño, los puntos se concentran más cerca de la recta; cuando es grande, la nube de puntos se vuelve más dispersa.

Nuestro entendimiento previo sugiere que el número diario de viajes en bicicleta suele ser variable, siendo siempre menor a 10.000 viajes por dia, y oscilando tre aproximadamente 2.000 y 9.000. Esperamos una relación apenas negativa entre la humedad y la cantidad de viajes: a medida que aumenta la humedad, el número de viajes tiende a disminuir, pero muy levemente. No obstante, dicha relacion no se ve del todo clara en todos los datasets simulados. De esta manera, seria mas "seguro" afirmnar que lo que nos dice nuestro entendimiento previo es que la relacion no es positiva.

Es decir que nuestros priors nos otorgan una incertidumbre considerable al respecto de que tan fuerte es la correlacion.

Exercise 9.10: Data

Si bien ya graficamos previamente la relacion entre viajes y hunedad, reiteraremos el grafico y profundizaremos en su analisis:



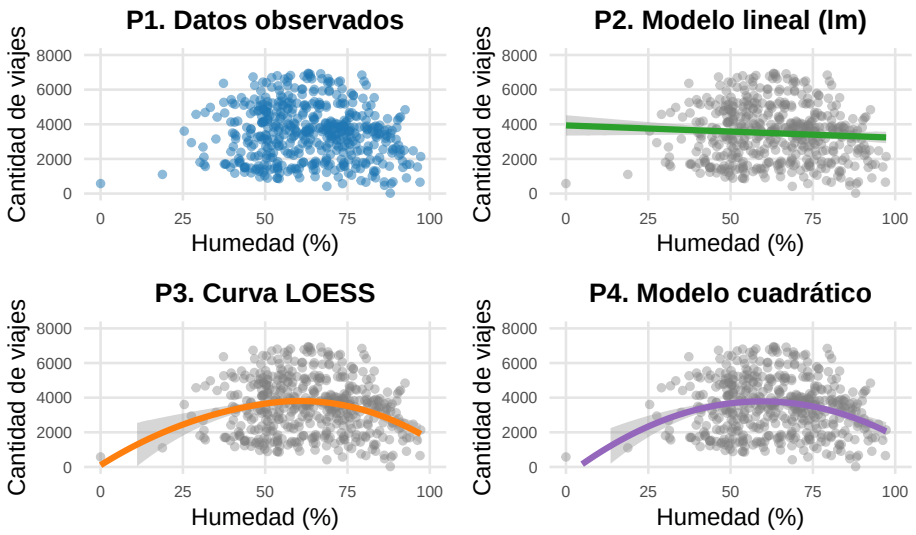
Se observa en los datos una muy leve tendencia negativa o decreciemte entre la humedad y la cantidad de viajes, es decir que los dias mas humedos tienden levemente a registrar apenas menos viajes que los dias mas secos. Esta pendiente casi plana nos indica que no hay una correlacion estricta entre las dos variables.

Si observamos los valoers extremos (cuando hay humedad cerca de cien), vemos que tiende a bajar la cantidad de viajes, Pasa al reves para los valores de menor humedad, tiende a subir levemente la cantidad de viajes.

Si miramos los valores promedio de humedad, esta relacion se vuelve sumamente debil y dispersa. Por ejemplo , para una humedad de aprox 60%, la cantidad de viajes oscila entre 500 y 7000, es decir, que soloo saber la humedad del dia, no nos da informacion real sobre la cantidad de viajes. De esta manera, podemos interpretar que la humedad capta muy poco de la variabilidad.

Esto que mencionamos rompe un poco con uno de las suposiciones del modelo normal, llamado heterocedasticidad, o en palabras mas sencillas, el criterio de varianza constante. Distintos valores de humedad nos devuelven varianzas levemente distintas. Esta diferencia, si bien debemos marcarla, no rompe con la posibilidad de utilizar el modelo, tal como se menciono en clase.

Si graficamos la dispersion con otro geom_method, observamos que mejoran levemente las aproximaciones obtenidas. No obstante, a efectos del ejercicio y dado que la mejora no es tan importante, decidimos avanzar con el modelo de regresion normal simple, a pesar de no ser perfecto. Para esta parte, nos apoyamos en ChatGpt dado que desconociamos los otros metodos y que es lo que realiza cada uno.



El modelo *LOESS* lo que hace es ajustar una curva flexible localmente, sin asumir una forma específica; se adapta a la tendencia real de los datos punto por punto. LOWESS (o LOESS) significa Locally Weighted Scatterplot Smoothing, es decir que ajusta una curva a los datos utilizando observaciones cercanas a cada punto de interés, no obstante es un metodo no parametrico, que entendemos escapa a los contenidos del curso. En cambio, el modelo cuadratico ajusta una parábola (relación polinómica de grado 2), permitiendo una curvatura global suave en la relación entre humedad y viajes. En este caso nos aporta valores practicamente iguales al metodo definido anteriormente.

Por otro lado, si cambiamos el foco del analisis y ahora comparamos contra el dataset con datos simulados en base a nuestros priors, tenemos ciertas diferencias considerables:

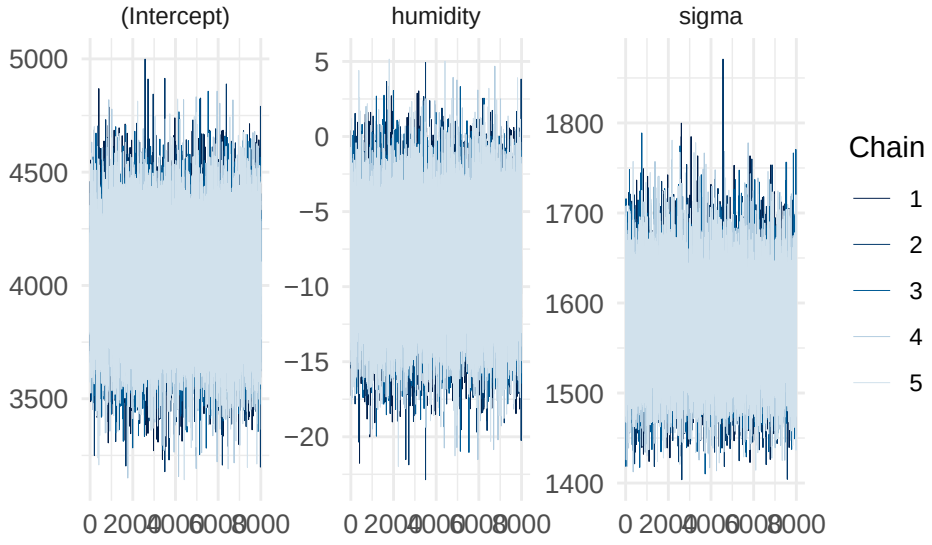
- Diferencias en la media de viajes: En el prior las centramos entornoa 5000, mientras que los datos se encuentran entorno a 4000
- Dispersion: En los priors oscilaban entre 1000 y 9000, mientras que en los datos van de 500 a 7000

Exercise 9.11 Posterior simulation

Simularemos los posterior a partir de 5 cadenas, utilizando la funcion update sobre la simulacion anterior (sin el parametro prior_pd, para ahora si incluir los datos)

```
bike_model_pos <- update(bike_model_hum, prior_PD = FALSE)
```

Para determinar si las cadenas se mezclaron correctamente, utilizaremos el diagnostico que usamos habitualmente para MCMC: rhat, neff_ratio y los graficos tipicos de diagnostico:



En los gráficos de traza se puede ver cómo se comportaron las cinco cadenas del muestreo para los tres parámetros del modelo. En general, las cadenas se mezclan bien y no muestran ninguna tendencia rara ni demasiados saltos abruptos, lo que sugiere que el proceso de simulación fue estable.

En el caso del intercepto, los valores oscilan entre más o menos 3500 y 5000, lo cual tiene sentido porque representa el promedio de viajes cuando la humedad es promedio. Se alinea con lo que vimos en los datos anteriormente.

El parámetro de la pendiente (humidity) también se mueve dentro de un rango lógico (alrededor de -15 a 0), mostrando que la estimación del efecto de la humedad es negativa, como esperábamos, y que las cadenas exploraron bien esa zona. La gran mayoria de los valores fue negativa.

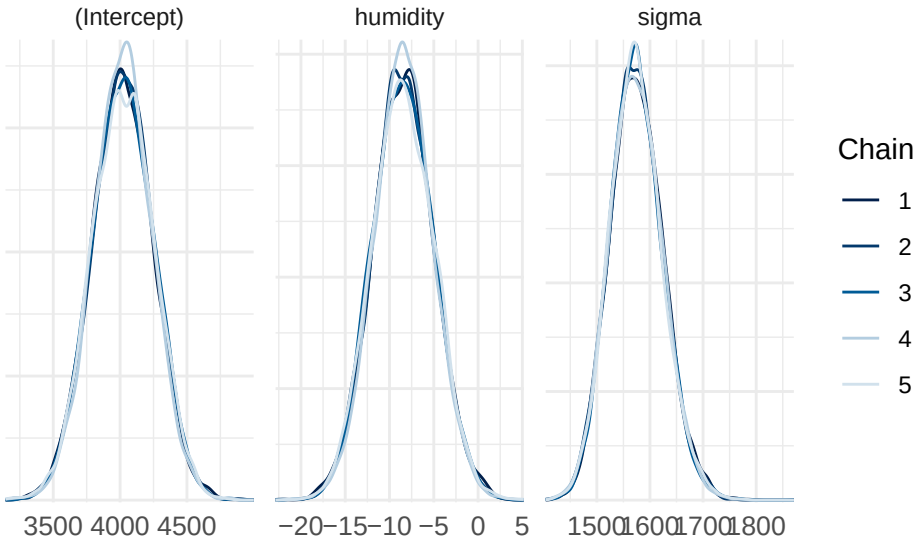
Por último, el parámetro sigma, que representa la variabilidad de los datos, también se mantiene estable entre aproximadamente 1400 y 1800.

En conjunto, las trazas no presentan señales de falta de convergencia, así que parece que las cadenas se mezclaron correctamente.

En el caso de las densidades posteriores se observa que para cada parámetro del muestreo las curvas de las distintas cadenas prácticamente se superponen, lo que indica que todas convergieron hacia la misma distribución y que el muestreo fue estable.

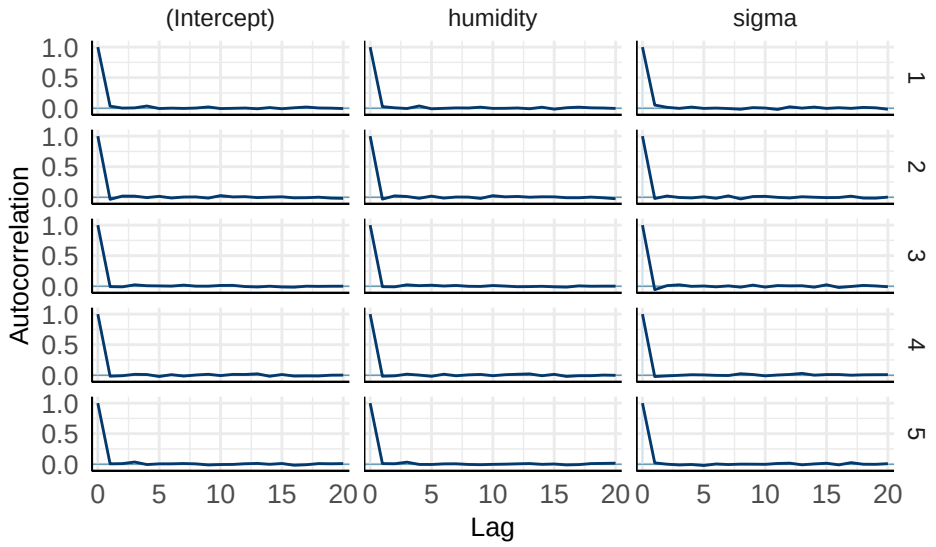
Las distribuciones de los tres parámetros tienen una forma aproximadamente normal, sin irregularidades ni multimodalidad. El intercepto está centrado alrededor de 4000, lo que representa el promedio de viajes cuando la humedad es baja.

No queremos dejar de mencionar que la cadena celeste (cadena nro 4), tanto para el intercepto como la pendiente, esta levemente mas concentrada que el resto de cadenas, lo que indica que tuvo levemente menor variabilidad.



Al mirar las autocorrelaciones para cada cadena y parámetro, vemos que la mismas decaen muy rápido hacia cero, lo cual indica que las simulaciones son bastante independientes. A diferencia

de en cadenas anteriores, vemos quela línea de autocorrelación no se mantiene pegada a cero perfectamente, sino que en algunos casos hay una leve correlación entre valores consecutivos. Este patrón se observa en mayor o menor medida en todas las cadenas.



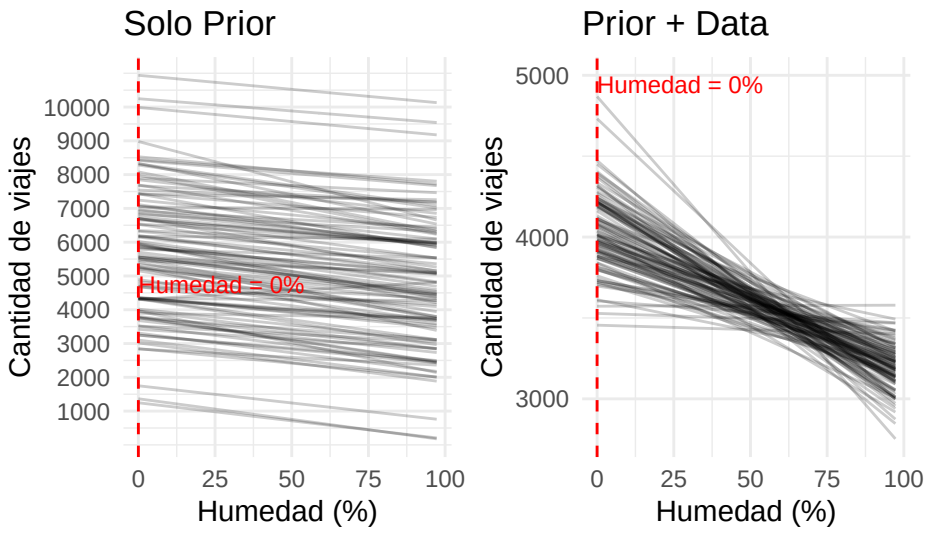

```
(Intercept)    humidity       sigma
   0.914300    0.940150    0.992825
```

Los valores de neff_ratio son cercanos a 1 para los tres parámetros, lo que indica que las cadenas presentan muy baja autocorrelación y que la mayoría de las simulaciones son efectivamente independientes.

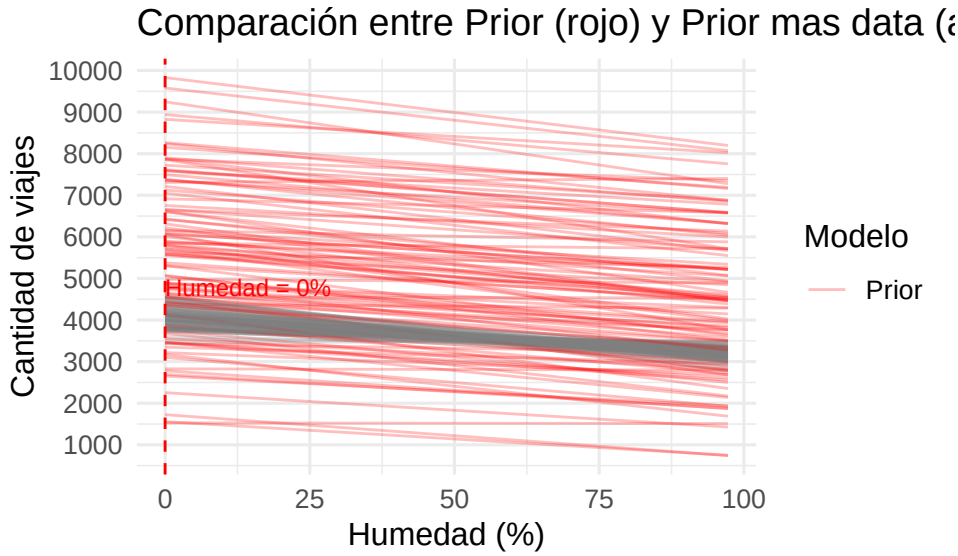
Se observa que sigma presenta un neff_ratio más alto, lo cual indica que sus simulaciones son aún más independientes que las del intercepto o la pendiente.

Ver bien por qué chota no funciona

Comparacion entre 100 priors y posteriors



Sinceramente, es muy dificil extraer info de estos dos graficos presentados de esta manera, se ven practicamente iguales, por lo tanto decidimos superponerlos en uno solo:



“‘

Fuentes utilizadas:

https://www.scribbr.com/methodology/explanatory-and-response-variables/ https://rpubs.com/cristina_gil/re
gresion_lineal_simple

https://ggplot2.tidyverse.org/reference/geom_smooth.html

https://www.rdocumentation.org/packages/ggplot2/versions/2.0.0/topics/geom_smooth

https://www.maximaformacion.es/blog-dat/que-es-la-regresion-local-loess-o-lowess/