Tarea 8

Santiago Robatto y Sofía Terra

Ejercicio 9.9 Bayes Rules!

La letra del ejercicio 9.9 nos pide estudiar cómo la humedad afecta la cantidad de viajes en bicicleta. Para eso,nos da ciertos supuestos, que tomaremos para definir nuestros priors:

- 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 en el contexto de este ejercicio.

Recordemos que el modelo definido en el libro es:

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

Por ende, tendremos:

- Y: 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 en la ciudad de Washington DC.
- X: Es la variable explicativa, la cual sera utilizada para intentar entender o 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 nuestra intuición nos hace pensar que es probable que la relación con la cantidad de viajes sea menos fuerte que en el caso de la temperatura.
- 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, representa la "razón de cambio". Esto es, en este caso, cómo 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 deberemos 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, a priori, no sera sumamente fuerte.

Para definir una varianza coherente en el modelo, utilizaremos la "regla" del modelo normal, que dice que 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, sabemos que σ=5.

De esta manera 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 y por ende $\beta_0 \sim \text{Normal}(5.000, 2.000)$

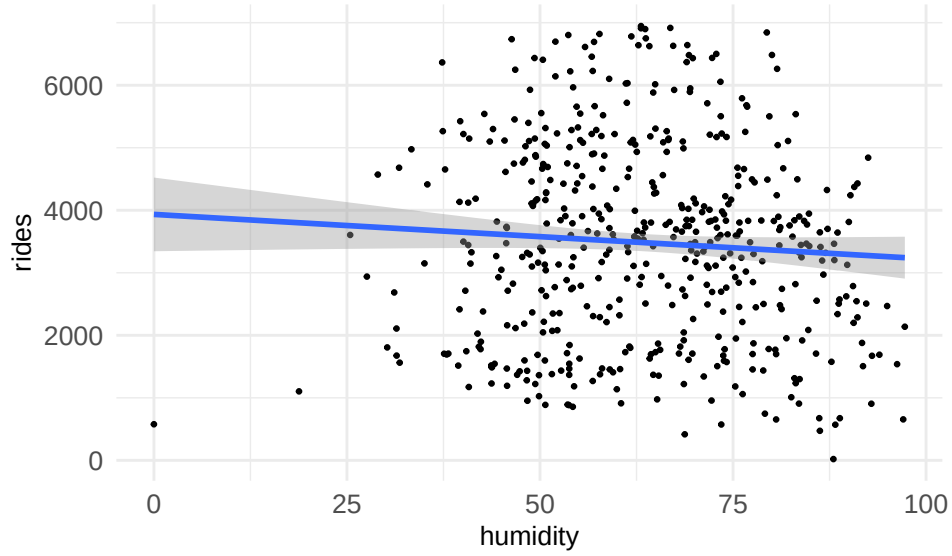
Verificamos el IC 95%:

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 que combina los priors con los datos para simular la distribución posterior de los parámetros utilizando MCMC.

Una manera estandar de definir el modelo es la siguiente:

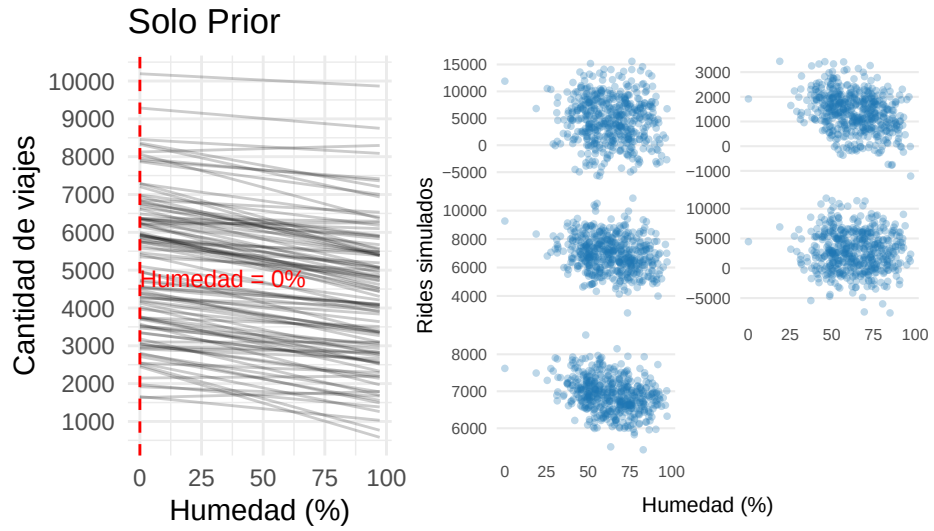
```
bike_model_hum<- stan_glm(rides ~ humidity,  #Rides en funcion de la humedad
                          data = bikes, #Dataset
                          family = gaussian, #Asumir errores bajo modelo normal
                          prior_intercept = normal(5000, 2000), #Definicion del intercepto
                          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 nuestra 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)
```



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 gráfico de la derecha, los conjuntos simulados bajo los priors muestran una amplia dispersión vertical y horizontal, lo que refleja nuestra incertidumbre sobre la fuerza de la relación entre humedad y viajes y el punto de corte con el origen. De esta manera, los priors son débilmente informativos y mantienen la incertidumbre en un rango alto pero 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.

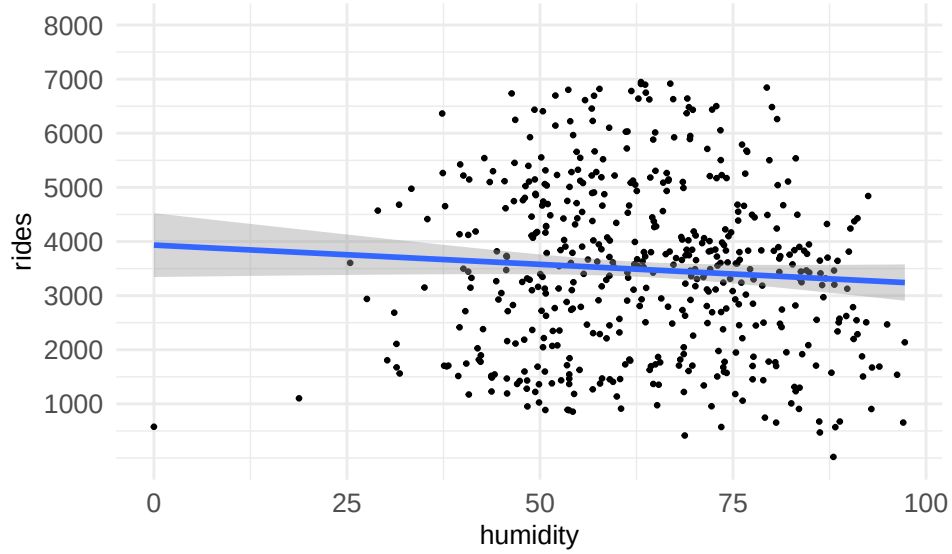
A su vez, se observa a simple vista que los distintos paneles presentan niveles de dispersión variables. Esto ocurre porque en cada simulación los valores del parámetro σ (la desviación del error) son distintos. Cuando σ es pequeño, los puntos se concentran más cerca de la recta; cuando es grande, la 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:



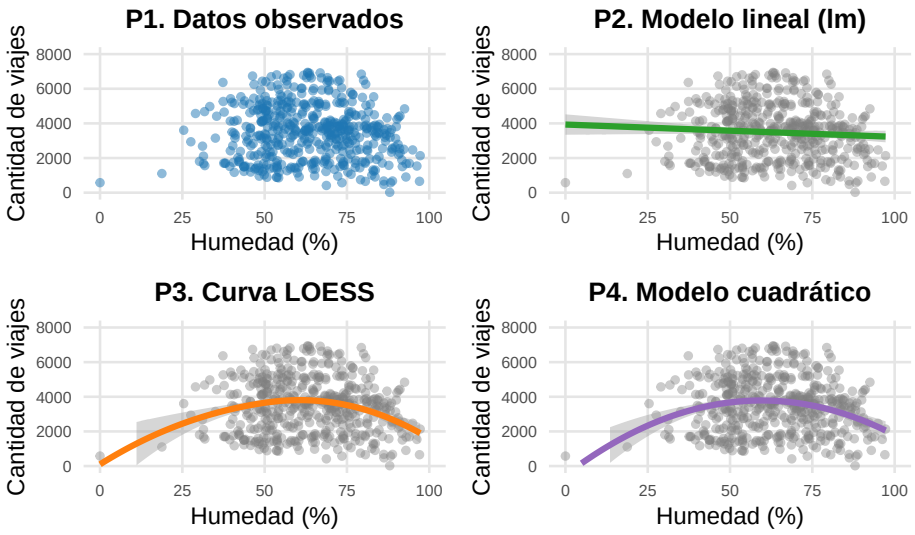
Se observa en los datos una muy leve tendencia negativa o decreciemte entre la humedad y la cantidad de viajes, es decir que los dias mas humedos tienden levemente a registrar apenas menos viajes que los dias mas secos. Esta pendiente casi plana nos indica que no hay una correlacion estricta entre las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

Si observamos los valoers extremos cuando hay humedad cercana a 100% vemos que tiende a bajar la cantidad de viajes. Ocurre 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 solo 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 poca información al respecto de la cantidad de viajes.

Esto que mencionamos anteriormente 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 no queríamos dejar de mencionarla, no nos impide de utilizar el modelo dado que es leve.

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 LOESS, radicando la principal diferencia en el ancho de los intervalos de confianza.

Por otro lado, volviendo al ejerciicio y cambiando el foco del analisis para ahora comparar contra el dataset con datos simulados en base a nuestros priors, tenemos ciertas diferencias considerables:

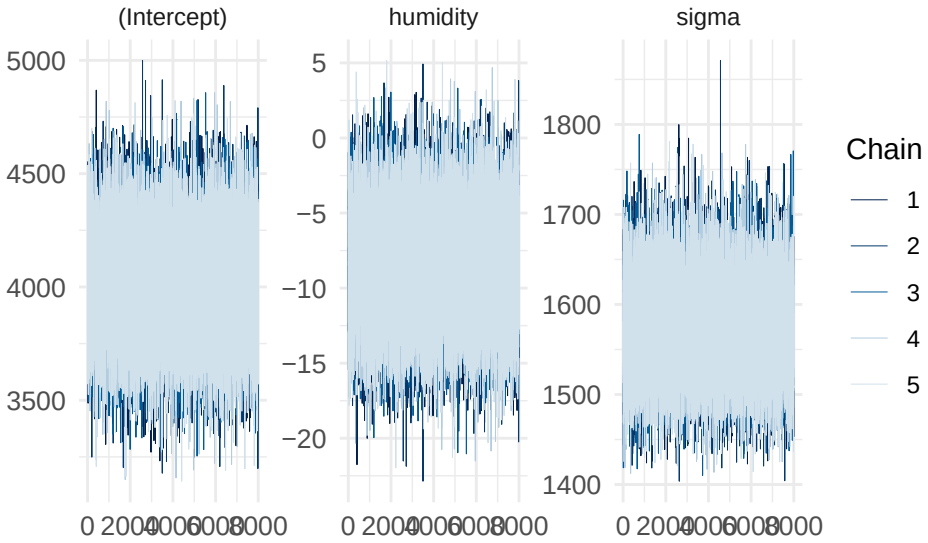
- Diferencias en la media de viajes: En el prior las centramos entornoa 5000, mientras que los datos se encuentran entorno a 4000
- Dispersion: En los priors oscilaban entre 1000 y 9000, mientras que en los datos van de 500 a 7000

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 con los metodos MCMC: rhat, neff_ratio y los graficos tipicos de diagnostico:



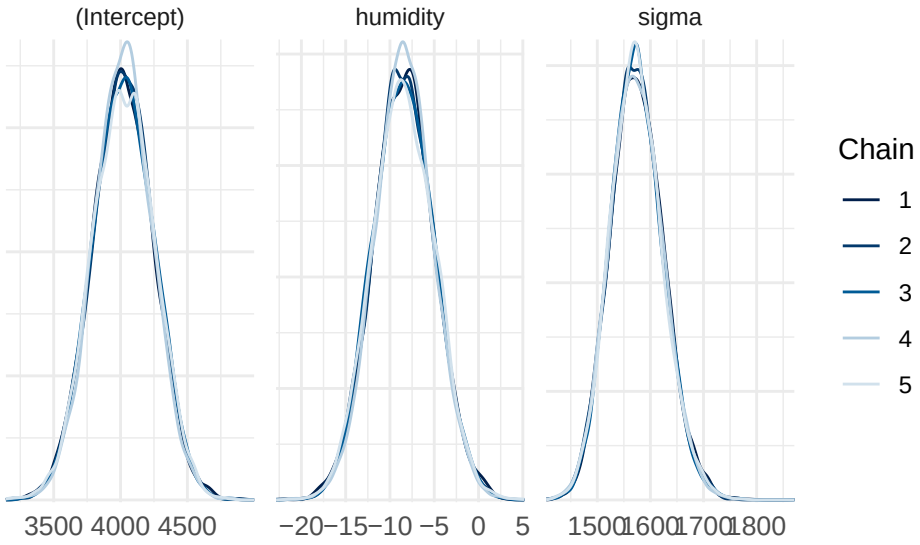
En los gráficos de traza se puede ver cómo se comportaron las cinco cadenas del muestreo para los tres parámetros del modelo. En general, las cadenas se mezclan bien y no muestran ninguna tendencia rara ni demasiados saltos abruptos, lo que sugiere que el proceso de simulación fue estable.

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.

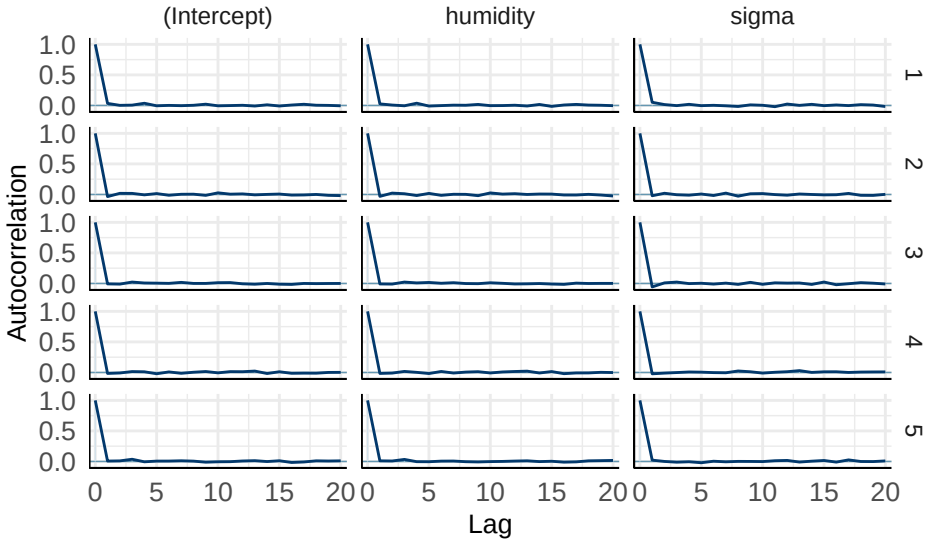
En conjunto, las trazas no presentan señales de falta de convergencia, así que parece que las cadenas se mezclaron correctamente.



En el caso de las densidades a posteriori 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
```

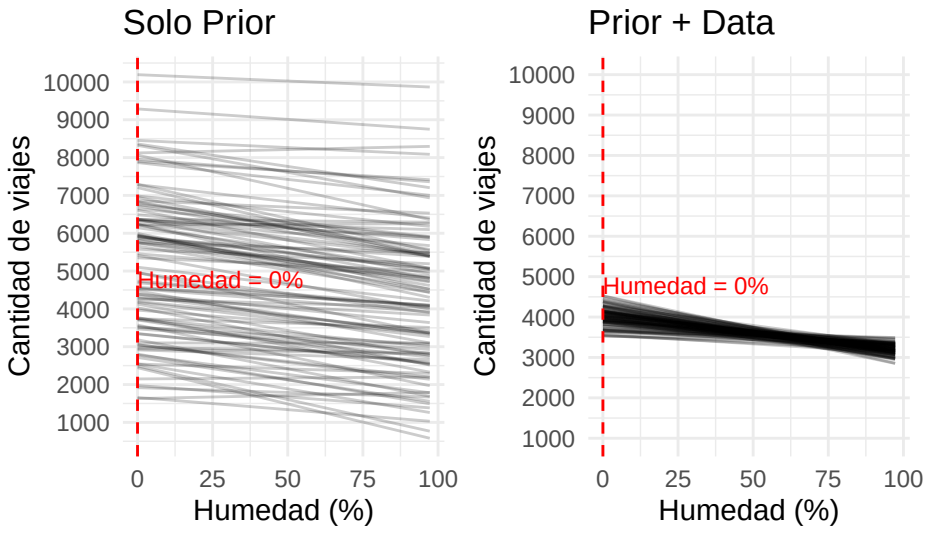
Los valores de neff_ratio son cercanos a 1 para los tres parámetros, lo que indica que las cadenas presentan muy baja autocorrelación y que la mayoría de las simulaciones son efectivamente independientes.

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.

En el caso del rhat estoy teniendo un problema puntualmente en R para correrlo, tal como les comenté en clase el día lunes. Intenté solucionarlo probando distintas alternativas pero no lo logré. El código extraído del libro tampoco me funcionó.

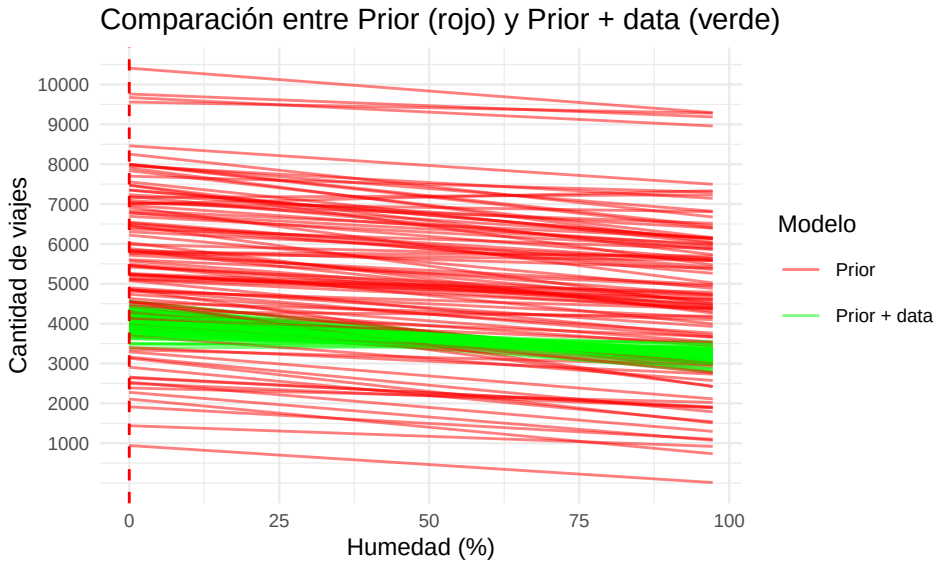
Comparacion entre 100 priors y posteriors

Procederemos a comparar las 100 rectas obtenidas a partir de los prioris únicamente contra las rectas que combinas prior + data.



A simplemente se ve claramente que al combinar prior + data las rectas son mucho más consistentes entre sí, reduciendo principalmente la dispersión vertical entre rectas.

Para poder comparar las pendientes decidimos además superponerlos en un solo gráfico:



Ahora, claramente se observa que las líneas rojas tambien están mucho más dispersas. Algunas incluso muestran una relación casi nula o incluso levemente positiva. Esto refleja la alta incertidumbre previa sobre la magnitud y dirección del efecto de la humedad que definimos a priori.

De esta manera, el gráfico muestra claramente cómo los datos actualizan nuestras creencias previas: el modelo pasa de considerar una gran variedad de posibles relaciones (prior amplio) a una relación más específica.

Excersie 9.12: How humid is too humid: posterior interpretation

El objetivo ahora es entender cualitativamente la posteriori, en cuanto a la relación entre los viajes de las personas que andan en bicicleta y la humedad.

Comenzamos utilizando la función tidy() (del paquete broom.mixed) la cual nos proporcionara el resumen de estadísticas sobre nuestro modelo. Debido a que nuestra Regresión Normal tiene dos parámetros (viajes y humedad), debemos especificar los "effects", es decir, nuestros parámetros de interés, que en este caso son β_0 y β_1 (ambos "fixed") y (σ) (auxiliar). Asimismo, en este caso analizaremos los intervalos de confianza al 95%.

```
# A tibble: 4 x 5
  term        estimate std.error conf.low conf.high
  <chr>          <dbl>     <dbl>    <dbl>     <dbl>
1 (Intercept)  4021.      230.     3574.    4460.
2 humidity       -8.47      3.42    -15.1     -1.72
3 sigma        1573.       49.6    1482.    1677.
4 mean_PPD     3484.       99.8    3291.    3680.
```

Comenzamos analizando la mediana de σ. Ya sabemos que σ mide la variabilidad de los datos que el modelo no puede explicar, podriamos decir que es el error que asumimos vamos a tener en la predicción. Entonces, el valor estimado de σ es 1573. Es decir que en promedio, esperamos desviarnos 1573 de la regresion. Es una dispersion considerable teniendo en cuenta que estamos trabajando con valores de 4000 y 5000 viajes promedio, es decir que aceptamos una desviacion estandar de mas del 30%. Su intervalo de confianza es sumamente ajustado, lo que nos hace pensar que la estimacion de σ es precisa.

El valor de Mean_ppd nos indica el promedio de viajes predicha por el modelo, es decir que esperamos 3484 viajes en promedio por dia. Esto refleja el "centro" de las predicciones: el modelo espera alrededor de 3484 viajes por día en promedio, con un rango típico entre 3291 y 3680.

Por otro lado, nos interesa interpretar el intervalo de confianza al 95% para el parámetro de la humedad, siendo este β_1. Sabemos que β_1 representa la pendiente de nuestro modelo, nos dice cómo cambia ridership por cada unidad que se aumenta en humidity. Entonces bajo un intervalo de confianza del 95% podemos afirmar que con ese nivel de confianza un aumento en una unidad de humidity en cuanto al número de viajes se encuentra entre [-15.093 ; -1.717]. Este intervalo es considerablemente amplio comparado con la precision que veniamos teniendo antes.

En el caso del intercepto centrado, su intervalo de confianza es [3574, 4460]. Esto muestra una incertidumbre moderada, y es coherente con lo esperado inicialmente en el prior. Como comentamos cuando miramos la data, se noto que los viajes eran menos a los esperados, por lo tanto era esperable que el posterior ajuste hacia abajo el intervalo (actualiza los datos hacia 4000).

Finalmente, nos preguntamos si existe evidencia del posteriori sobre una asociación negativa entre el número de viajes en bicicleta respecto al nivel de humedad. Gracias al intervalo planteado anteriormente, podemos concluir que sí existe una relación negativa. El IC del 95% no incluye al 0, lo que significa que todos los valores plausibles para coeficientes de β_1 son negativos. Con mas de un 95% de confianza concluimos que si aumenta la humedad, habrá menos viajes en bicicleta.

Exercise 9.13: How humid is too humid: prediction

En este ejercicio queremos analizar el número de viajes en bicicleta para un día donde la humedad es del 90%.

Primero empezamos simulando dos modelos posteriori: uno para el número típico de viajes en bicicleta en días con 90% de humedad y otro que sea un modelo predictivo para el número de viajes del día siguiente. Para hacer esto debemos comenzar extrayendo los datos del modelo a un data frame.

Ahora podemos hacer el modelo sobre la media (μ) de viajes en un día con 90% de humedad. Debemos simular la regresión lineal con la ecuación $\mu = \beta_0 + \beta_1 \cdot 90$, que tiene en cuenta el dato conocido pero también la incertidumbre de los otros parámetros. Nos definimos el objeto viajes1_promedio_H90 de la siguiente manera:

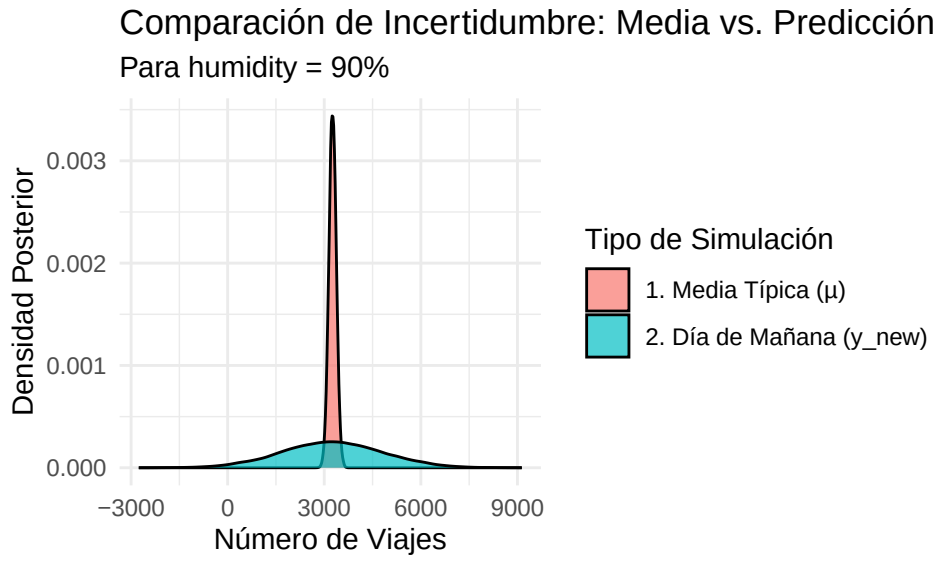
```
#aplicamos la formula de la regresion lineal teniendo en cuenta que
#la humedad es 90%
viajes1_promedio_H90 <- dataframe_model$`(Intercept)` + dataframe_model$humidity * 90

#viajes1_promedio_H90
```

Por otro lado debemos realizar el modelo sobre la incertidumbre de el siguiente día. Se realiza esto para observar la variabilidad que presenta nuestro muestreo sobre nuestra media. No todos los días tienen el mismo promedio. Entonces, podemos simular nuestra predicción y_{new} tomando un día aleatorio del modelo Normal. Hacemos esto con la función rnorm().

Tenemos entonces que el objeto viajes2_promedio_H90 representa un número plausible de viajes para un solo día con 90% de humedad.

Si graficamos ambos modelos obtenemos:



Se observa que ambas distribuciones están centradas en prácticamente el mismo valor, pero la dispersión del modelo que representa el día siguiente es mucho mayor que la del modelo de la

media. La gráfica de la media muestra la incertidumbre del modelo, mientras que la gráfica del día siguiente muestra la incertidumbre sobre la predicción. Tiene en cuenta la incertidumbre del modelo y la de los datos también, por lo tanto, es más acertado que el modelo que predice un evento específco. Presenta una realidad más balanceada.

Luego, nos interesa calcular el IC al 80% para el día siguiente, donde usamos el segundo modelo simulado.

```
     10%      90%
1247.188 5284.125
```

Por lo tanto, con un 80% de confianza sabemos que el número de viajes en bicicleta el día siguiente, sabiendo que hay un 90% de humedad, se encontrará en el intervalo [1247 ; 5284].

Finalmente, nuestro objetivo es comprobar que lo anterior es correcto con la función posterior_predict():

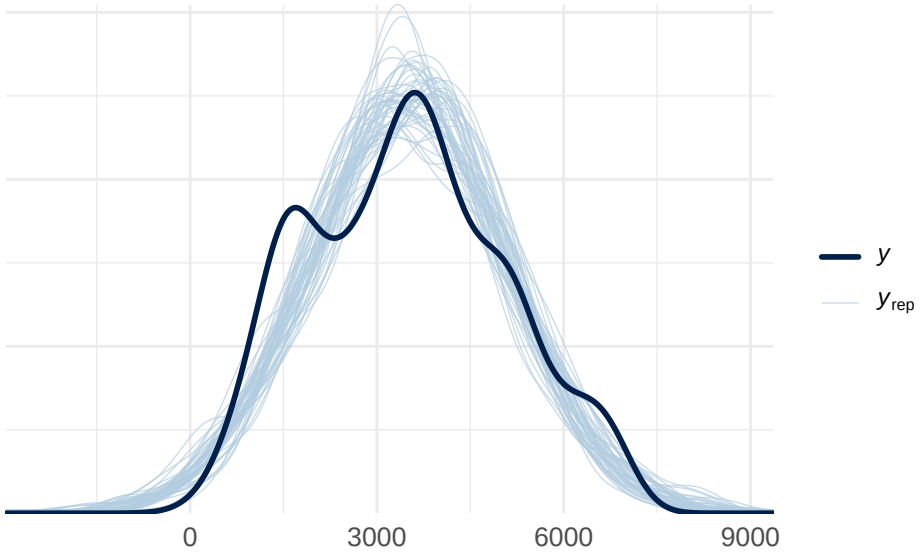
```
     10%      90%
1246.966 5302.915
```

Se observa que el intervalo obtenido es prácticamente igual al calculado de forma manual anteriormente. Por lo tanto, la simulación se realizó correctamente.

Ahora, queremos evaluar nuestro modelo de regresión lineal, queremos saber qué tan "justo" o adecuado es. Para eso se utilizan diversas funciones de diagnóstico.

Comenzamos con la función pp_check() (posterior predictive check) de rstanarm. Lo que pretende estudiar es si los datos simulados son coherentes con los datos reales, los muestreados. La curva y representa los datos reale, mientras que las curvas y_rep fueron las simuladas.

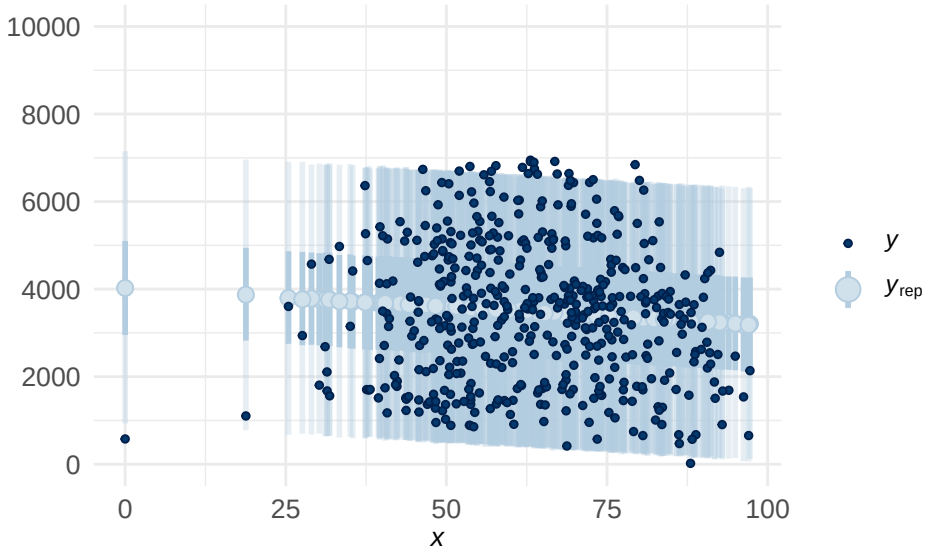
Observamos que y es bimodal, mientras que las y_rep son unimodales. No logramos que las curvas simuladas contemplen esta bimodalidad. Esto no significa necesariamente que el modelo de viajes en bicicleta sea malo, ya que el mismo nos aportó información que no teníamos antes, pero claramente puede mejorar. Quizá necesitaríamos implementar un modelo que contemple de forma explícita la bimodalidad, quiza debamos incorporar mas parametros o quizas deberiamos agregar una variable, por ejemplo location.



Por otro lado, analiazamos la función ppc_intervals(), de bayesplot, la cual es un complemento de pp_check(). Esta nos proporciona un resumen visual sobre aproximadamente 500 modelos posteriori predictivos. Nos enfocamos en los valores individuales. Queremos ver qué tan bien se ajustan los intervalos de confianza que se generaron para los valores con los reales correspondientes. Tenemos dos tipos de IC diferentes: por un lado los indicados por prob = 0.5 (IC 50%), que visualmente son las barras más marcadas, y por otro lado los indicados por prob_outer = 0.95 (IC 95%), que son las líneas más delgadas. El eje x representa humidity, mientras que los puntos y son los datos muestreados para cada nivel de humedad. Las rectas y_rep indican las predicciones para cada punto.

Observamos que la mayoría de los datos muestreados se encuenteran contenidos en los respectivos intervalos de confianza. Solamente algunos pocos caen fuera de los IC del 95%. Esto implica que el modelo se ajusta bien a la realidad, predice de forma correcta.

Esto puede parecer contrario a los que vimos en pp_check(). Sin embargo, como fue mencionado anteriormemnte, no es que el modelo estuviera mal, simplemente no contemplaba todos sus aspectos. Con este nuevo diagnóstico confirmamos que el modelo simula correctamente la dispersión de los datos, pero no logra aproximar correctamente qué "forma" tienen los mismos.



Por último, utilizamos prediction_summary del paquete bayesrules. Este nos proporciona el análisis cuantitativo que nos faltó en las dos funciones de diagnóstico que vimos. Utiliza tres medidas de resumen diferentes. La primera es mae (Median Absoulte Error). Este valor es la mediana de todos los errores de predicción absolutos (mide la diferencia típica entre lo observado y las medias del modelo predictivo posteriori). La segunda es mae_scaled, que nos dice el número típico de desviaciones típicas de los datos observados que "caen" de sus medias posteriori predictas. La tercera es within_50 y within_95, que nos da la proporcion de datos que se encuentran contenidos en sus correspondientes IC del 50% y 95%.

En nuestro caso, mae es aproximadamente 1165, lo que quiere decir que el 50% de las predicciones tuvieron un error absoluto igual o menor a 1165 viajes, mientras que la otra mitad tuvo un error igual o mayor a esa cantidad de viajes. El error parece ser grande, lo que indica que el modelo podría no llegar a ser preciso, lo cual es coherente con lo que vimos en pp_check().

En cuanto a mae_scaled, tenemos que esta vale aproximadamente 0.74. Esto significa que el Median Absolute Error es aproximadamente 74% de la desviación típica total de los datos sobre viajes en bicicleta. Por lo tanto, se confirma que el modelo es poco preciso. Se confirma la sospecha que se tenía de mae.

Finalmente, within_50 nos muestra que aproximadamente el 46% de los viajes en bicicleta "cayeron" dentro de sus correspondientes IC del 50%. Por otra parte, within_95 nos dice que el 96% de los viajes "cayeron" dentro de sus respectivos IC del 95%, lo que también es una buena señal. Entonces, a pesar de que el modelo no es preciso, representa de forma casi perfecta la incertidumbre. Entonces, ¿nuestro modelo es "justo"? La respuesta es que depende de nuestros objetivos y qué análisis hagamos de todo esto, en conclusion, es relativo al objetivo que persigamos.

```
       mae mae_scaled within_50 within_95
1 1165.385  0.7400281     0.462     0.964
```

Fuentes utilizadas:

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