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%:

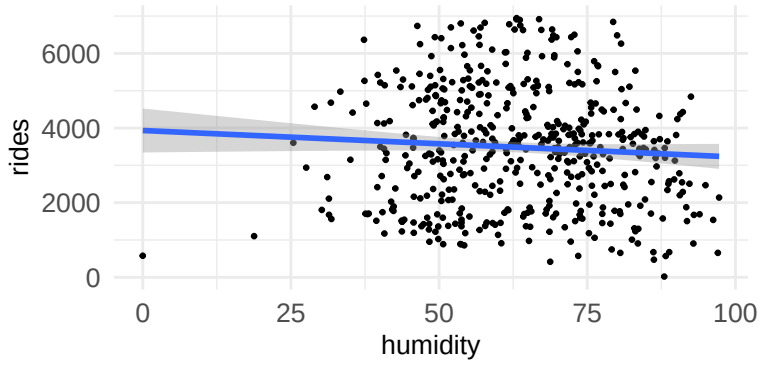
Cuantil	Valor
2.5%	1080.07
97.5%	8919.93

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.



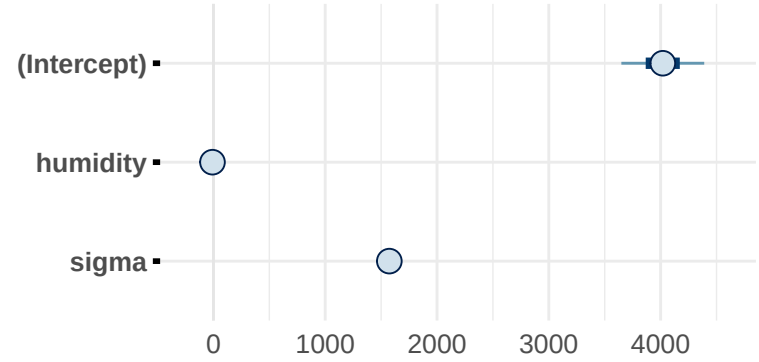
Juntando toda esta informacion, ya estamos en condiciones de definir el modelo. Utilizaremos para ello la funcion stan_glm, del paquete rstanarm, la cual, de manera muy resumida, lo que hace es, de forma bayesiana, combina los priors con los datos para simular la distribución posterior de los parámetros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

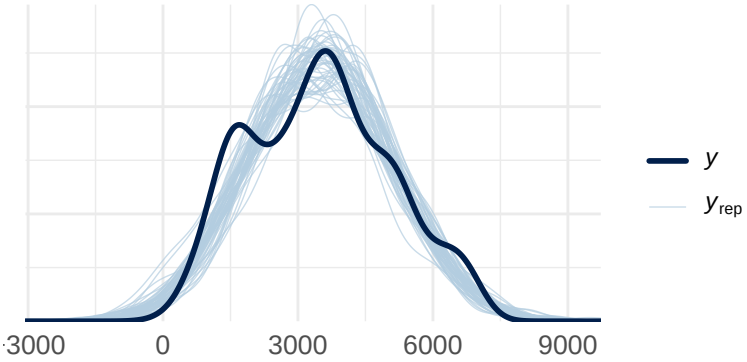
```
bike_model_hum<- stan_glm(rides ~ humidity,  #Rides en funcion de la humedad
                          data = bikes, #Dataset
                          family = gaussian, #Asumir errores bajo modelo normal
                          prior_intercept = normal(5000, 2000), #Definicion del intercepto
                          prior = normal(-10, 5),  #Definicion de B1
                          prior_aux = exponential(5e-04), #Defnicion de sigma (para el error)
                          chains = 4, # Cant. cadenas
                          iter = 5000*2, #Iteraciones por cadena
                          seed = 84735) #Semilla
```

Antes de avanzar a la parte b, checkearemos los resultados de la simulacion

```
Model Info:
 function:     stan_glm
 family:       gaussian [identity]
 formula:      rides ~ humidity
 algorithm:    sampling
 sample:       20000 (posterior sample size)
 priors:       see help('prior_summary')
 observations: 500
 predictors:   2

Estimates:
              mean   sd     10%    50%    90%
(Intercept) 4019.5  226.8 3729.5 4019.7 4310.4
humidity      -8.4    3.4  -12.8   -8.4   -4.1
sigma       1574.8   49.6 1512.6 1573.4 1639.1

Fit Diagnostics:
           mean   sd     10%    50%    90%
mean_PPD 3484.5   99.5 3356.5 3485.2 3611.6

The mean_ppd is the sample average posterior predictive distribution of the outcome variable

MCMC diagnostics
              mcse Rhat n_eff
```

```
(Intercept)   1.6  1.0  18989
humidity      0.0  1.0  18587
sigma         0.4  1.0  19215
mean_PPD      0.7  1.0  19148
log-posterior 0.0  1.0   8957

For each parameter, mcse is Monte Carlo standard error, n_eff is a crude measure of effective
```





Fuentes utilizadas:

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