Proyecto:

Regresión Logística Bayesiana Para La Predicción De Lluvia

Autor:

Martínez Hernández Luis Antonio

20 de Junio, 2022

Contents

RESÚMEN

Los modelos de regresión logística se utilizan comúnmente para modelar variables de respuesta en forma de variables categóricas con varias variables predictoras. La contribución de la variable predictora a la variable respuesta se expresa a través de un coeficiente de regresión (β). Por lo tanto, es necesario estimar β. Este estudio discute la estimación de β usando el método bayesiano.

El enfoque bayesiano utiliza una combinación de información de datos de muestra e información previa (a priori) sobre las características de los parámetros de interés, lo que da como resultado la información actualizada, es decir, la posterior (a posteriori). Por lo tanto, el método bayesiano puede superar el problema si la calidad de los datos de la muestra no respalda la observación. El método de regresión logística bayesiana se utilizará para analizar los datos de la ciudad de Perth, Australia. Perth experimenta veranos secos e inviernos húmedos. Nuestro objetivo será predecir si mañana lloverá o no. Es decir, queremos modelar Y, una variable de respuesta categórica binaria.

SUMMARY

Logistic regression models are commonly used to model response variables in the form of categorical variables with several predictor variables. The contribution of the predictor variable to the response variable is expressed through a regression coefficient (β). Therefore, it is necessary to estimate β. This study discusses the estimation of β using the Bayesian method.

Bayesian approach utilizes a combination of information from sample data and prior information about the characteristics of the parameters of interest, resulting in the updated information, namely the posterior. Bayesian method thus can overcome the problem if the quality of the sample data does not support observation. The Bayesian logistic regression method will be used to analyze the data from the city of Perth, Australia. Perth experiences dry summers and wet winters. Our objective will be to predict if it will rain tomorrow or not. That is, we want to model Y, a binary categorical response variable.

INTRODUCCIÓN

El análisis de regresión es un proceso de búsqueda de modelos matemáticos que sean más adecuados a los datos que tiene como objetivo estudiar la forma de la relación entre una o más variables explicativas con una variable comúnmente llamada variable de respuesta. Un tipo de modelo de análisis de regresión es el modelo de regresión logística, el cual cuenta con una variable de respuesta que tiene solo dos valores posibles (es decir, una variable de respuesta binaria). El modelo de regresión logística se puede utilizar si la variable de respuesta es una variable categórica. Por ejemplo, la variable de respuesta en el modelo de regresión logística donde un valor de 1 indica éxito y 0 indica fracaso.

En el modelo de regresión, los coeficientes de regresión juegan un papel importante, ya que el cálculo del riesgo relativo y la interpretación del modelo se basan en estos coeficientes. Por lo tanto, la estimación de estos parámetros debe realizarse de manera que los valores estimados resultantes puedan producir una predicción precisa y una interpretación perspicaz. Existen dos métodos de estimación de parámetros, el método frecuentista y el método bayesiano. El método frecuentista es un método de estimación de parámetros que solo utiliza información de datos de muestra, por lo que se depende en gran medida de la calidad de los datos.

JUSTIFICACIÓN

Cuando los datos utilizados no respaldan la observación, la precisión del método frecuentista es cuestionable. Por lo tanto, es necesario utilizar otra información que pueda respaldar los datos, por ejemplo, información previa sobre los parámetros a estimar. El enfoque bayesiano utiliza ambas fuentes de información: los datos de la muestra y la información previa (a priori). Con este enfoque, si hay una falta de confianza en los datos de la muestra, entonces, la incorporación de la distribución previa todavía produce un resultado óptimo, a diferencia de simplemente confiar en la muestra.

En el método bayesiano, el coeficiente de regresión se trata como una variable aleatoria. La información de los datos de observación se resume en la función de verosimilitud. Luego, la distribución a priori y la función de verosimilitud se combinarán y producirán una distribución posterior (a posteriori). Además, la extracción de conclusiones sobre los parámetros a estimar se basará en la distribución a posteriori. La distribución posterior se puede obtener en forma cerrada y no cerrada. Si la distribución a posteriori es una función de densidad de probabilidad de una distribución particular, se dice que la distribución a posteriori es de forma cerrada. La ventaja de obtener la distribución a posteriori de forma cerrada es que el estimador de Bayes se puede obtener sin utilizar técnicas computacionales complejas. Si se obtiene una distribución a posteriori de forma no cerrada, se necesitan técnicas computacionales para obtener estimadores bayesianos como el método Cadenas de Markov Monte Carlo (MCMC).

OBJETIVOS

General:

- Crear un modelo de regresión logística bayesiana para la predicción de si lloverá o no en la ciudad de Perth, Australia.

Específicos:

- Utilizar este modelo para clasificar o predecir el resultado de Y para un conjunto dado de valores predictores X.
- Evaluar la calidad de esta técnica de clasificación.

MÉTODO

Esta sección contiene la base teórica utilizada en este estudio, incluido el modelo de regresión logística y el método bayesiano, así como los datos y las variables utilizadas. Las discusiones sobre el método bayesiano incluyen la distribución a priori y la construcción de la distribución a posteriori.

Modelo De Regresión Logística

El modelo de regresión logística es un modelo de regresión con una variable de respuesta binaria. Por lo tanto, se puede suponer que la variable de respuesta es una variable aleatoria de Bernoulli y, si suponemos que hay observaciones, la variable de respuesta para la i-ésima observación es y_i (y_i sigue una distribución de Bernoulli). El modelo de regresión lineal es como sigue:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_{i2} + \cdots + \beta_k x_{ik} + \varepsilon_i.$$

con $\varepsilon_i \sim NIID(0, \sigma^2)$. Como $E(\varepsilon_i) = 0$, entonces,

$$E(y_i|x_{i1}, x_{i2}, \ldots, x_{ik}) = \beta_0 + \beta_1 x_{i1} + \beta_{i2} + \cdots + \beta_k x_{ik}$$

De la distribución de probabilidad de Bernoulli, $P(y_i = 1) = \pi_i$ y $P(y_i = 0) = 1 - \pi_i$, se obtiene

$$E(y_i|x_{i1}, x_{i2}, \ldots, x_{ik}) = \pi_i,$$

Entonces, tenemos

$$\pi_i = \beta_0 + \beta_1 x_{i1} + \beta_{i2} + \cdots + \beta_k x_{ik}$$

Hay un problema en la ecuación anterior. Es decir, la probabilidad de π_i en el lado izquierdo debe estar entre 0 y 1, pero $\beta_0 + \beta_1 x_{i1} + \beta_{i2} + \cdots + \beta_k x_{ik}$ en el lado derecho puede tener cualquier valor numérico real. No hay garantía de que el valor previsto esté en el rango correcto a menos que se impongan restricciones complejas al coeficiente. Una solución a este problema es transformar la probabilidad para eliminar el rango de restricciones y modelar la transformación como una función lineal de $x_{i1}, x_{i2}, \ldots, x_{ik}$. La transformación se puede realizar en los siguientes dos pasos:

- Calcular las odds de π_i. Odds es la razón de una probabilidad y su complemento.

$$odds_{\pi_i} = \frac{\pi_i}{1 - \pi_i}$$

- Calcular el logaritmo de $odds_{\pi_i}$ comúnmente llamado transformación logit o log-odds de probabilidad π_i, es decir

$$\eta_i = logit(\pi_i) = \log\left(\frac{\pi_i}{1 - \pi_i}\right)$$

Se puede ver que si la probabilidad de π_i cae hacia 0 entonces $odds_{\pi_i}$ irá a 0 y $logit(\pi_i)$ irá a $-\infty$. Mientras que si la probabilidad de π_i es 1, entonces $odds_{\pi_i}$ será ∞ y $logit(\pi_i)$ también será ∞. Por lo tanto, logit asigna la probabilidad del rango (0, 1) a toda la línea real. Después de eso, se debe realizar la transformación inversa, para obtener lo siguiente,

$$\pi = logit^{-1}(\eta_i) = \frac{e^{\eta_i}}{1 + e^{\eta_i}}$$

La ecuación anterior es una función logística que tiene un valor de curva no lineal en forma de S (en forma de S inversa) de entre 0 y 1.

Supongamos que el logit de π_i es un predictor lineal, es decir

$$\eta_i = \beta_0 + \beta_1 x_{i1} + \beta_{i2} + \cdots + \beta_k x_{ik}.$$

Puesto que $E(y_i|x_{i1}, x_{i2}, \ldots, x_{ik}) = \pi_i$, entonces

$$E(y_i|x_{i1}, x_{i2}, \ldots, x_{ik}) = \pi_i = logit^{-1}(\eta_i) = \frac{e^{\eta_i}}{1 + e^{\eta_i}} = \frac{e^{\beta_0 + \beta_1 x_{i1} + \beta_{i2} + \cdots + \beta_k x_{ik}}}{1 + e^{\beta_0 + \beta_1 x_{i1} + \beta_{i2} + \cdots + \beta_k x_{ik}}}$$

La ecuación anterior se llama función de respuesta logística.

Función de Verosimilitud

Sea $Y_1, Y_2, \ldots, Y_n$ una muestra aleatoria con la función de densidad de probabilidad $f(y|\beta) : \beta \in \Omega$ donde β es un parámetro cuyo valor se desconoce y Ω es un espacio de parámetros. Suponga que la observación da el resultado $Y_1 = y_1, Y_2 = y_2, \ldots, Y_n = y_n$. La información de esa observación se resumirá en la función de verosimilitud. La función de verosimilitud se puede expresar como una función de densidad de probabilidad conjunta de $Y_1, Y_2, \ldots, Y_n$, es decir

$$L(\beta) = f(y_1, y_2, \ldots, y_n|\beta).$$

Dado que $Y_1, Y_2, \ldots, Y_n$ son muestras aleatorias, la verosimilitud se puede reescribir de la siguiente manera:

$$L(\beta) = f(y_1, y_2, \ldots, y_n|\beta) = \prod_{i=1}^{n} f(y_i|\beta).$$

Distribución A Priori

En el método bayesiano, el parámetro a estimar (β) se trata como una variable aleatoria y tiene un valor en el dominio de B. La distribución anterior se denota por $f_B(\beta)$, indicando información sobre el parámetro antes de hacer una observación. Las distribuciones a priori se pueden obtener a partir de información previa sobre los parámetros estimados, por ejemplo, de investigaciones previas o de teorías disponibles. La elección de la distribución a priori influye en gran medida en los resultados de la evaluación utilizando el método bayesiano.

La distribución a priori se divide en dos, a saber, a priori conjugada y a priori no conjugada. Los a priori que no cumplen con la definición de a priori conjugados son a prioris no conjugados. Se dice que la distribución a priori es la distribución a priori conjugada para un modelo en particular si la distribución a posteriori resultante proviene de la misma familia que la distribución a priori. La selección de distribuciones conjugadas a priori se basa en la similitud de las formas funcionales con el modelo de verosimilitud.

Con a prioris conjugados, la información a posteriori se puede obtener en forma cerrada. Se dice que la información a posteriori es de forma cerrada, si la información posterior es una función de la densidad de probabilidad de una distribución particular, lo que facilita el análisis o la inferencia. Sin embargo, no siempre es posible obtener formas a priori que sean similares a las formas de verosimilitud. Por lo tanto, existe otra alternativa en la determinación de las a priori, denominadas a prioris no conjugados. La selección de distribuciones a prioris no conjugadas se basa en la información de las características de los parámetros. En este caso, la similitud de la forma funcional con el modelo de verosimilitud ya no es una consideración importante.

Distribución A Posteriori

La distribución posterior (a posteriori) es una función de densidad de probabilidad condicional de β dado $Y_1 = y_1, Y_2 = y_2, \ldots, Y_n = y_n$, se denota por $f(\beta|y_1, y_2, \ldots, y_n)$, por lo que en base al teorema bayesiano se puede escribir,

$$f(\beta|y_1, y_2, \ldots, y_n) = \frac{f(y_1, y_2, \ldots, y_n|\beta) f_B(\beta)}{g(y_1, y_2, \ldots, y_n)} = \frac{f(y_1, y_2, \ldots, y_n|\beta) f_B(\beta)}{\int_{-\infty}^{\infty} f(y_1, y_2, \ldots, y_n|\beta) f_B(\beta) d\beta}$$

donde $f_B(\beta)$ es una distribución a priori y $f(y_1, y_2, \ldots, y_n|\beta)$ es una función de verosimilitud. Dado que el denominador es solo una constante de normalización que no afecta la forma de la distribución, podemos simplificar el problema considerando solo los componentes que son proporcionales a la posterior, es decir

$$aposteriori \propto verosimilitud \times apriori.$$

Datos De La Ciudad De Perth, Australia.

Perth es una ciudad del oeste de Australia, capital del estado de Australia Occidental. Tiene 2,125,114 habitantes, lo que la convierte en la cuarta ciudad más poblada de Australia. Se encuentra en el estuario del río Swan. Su denominación procede de la ciudad de Perth, en Escocia. El área metropolitana se encuentra entre el océano Índico y una baja escarpadura costera conocida como Montes Darling. La ciudad más cercana a Perth con una población de más de un millón de personas es Adelaida, a 2104 kilómetros de distancia, lo que convierte a Perth en la ciudad con más de un millón de habitantes que se encuentra más aislada del mundo.

Perth, ubicada en la costa suroeste, experimenta veranos secos e inviernos húmedos. Contamos con un conjunto de datos de 1000 días con registros de humedad y si el día siguiente llovió o no. Nuestro objetivo será predecir si mañana lloverá o no. Es decir, queremos modelar Y, una variable de respuesta categórica binaria , convertida a un indicador 0-1 por conveniencia:

$$Y = \begin{cases} 1 & \text{si llueve mañana,} \\ 0 & \text{de lo contrario.} \end{cases}$$

Aunque hay varios predictores potenciales de lluvia, consideraremos solo uno:

$$X_1 = \text{Humedad de hoy.}$$

Y, también tenemos como información previa que:

- En un día promedio, hay aproximadamente un 20% de probabilidad de lluvia.
- La probabilidad de lluvia puede aumentar hasta en un 52% cuando está precedida por un día con mucha humedad o lluvia.

Los datos se presentan a continuación (solo las primeras 20 filas):

	raintomorrow	humidity
1	No	55
2	No	43
3	Yes	62
4	No	53
5	No	65
6	No	84
7	No	48
8	No	51
9	Yes	47
10	Yes	90
11	No	53
12	No	87
13	No	83
14	No	96
15	No	81
16	No	53
17	No	79
18	No	77
19	No	65
20	Yes	96

Especificación De Los A Priori

Sea n el tamaño de muestra con Y_i distribuida como una Bernoulli entonces, el modelo de regresión logística para estos datos es:

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 X_{i1}$$
$$\pi_i = \frac{e^{\beta_0+\beta_1 X_{i1}}}{1+e^{\beta_0+\beta_1 X_{i1}}}$$

De este modo, la función de verosimilitud es:

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n}\left(\frac{e^{\beta_0+\beta_1 X_{i1}}}{1+e^{\beta_0+\beta_1 X_{i1}}}\right)^{y_i}\left(1-\frac{e^{\beta_0+\beta_1 X_{i1}}}{1+e^{\beta_0+\beta_1 X_{i1}}}\right)^{1-y_i}$$

Tomando como variable X_1 la humedad de hoy. Para completar el modelo de regresión logística bayesiana de Y, debemos poner modelos a priori en nuestros dos parámetros de regresión, β_0 y β_1. Dado que estos parámetros pueden tomar cualquier valor en los reales, los valores a priori normales con media μ_0 y varianza σ_0^2 son apropiados para ambos, también asumiremos independencia, por lo que la distribución a priori conjunta para todos los coeficientes de regresión se puede escribir como:

$$f(\beta_0,\beta_1) = \prod_{j=0}^{1}\frac{1}{\sigma_{j0}\sqrt{2\pi}}\exp\left[\frac{(\beta_j-\mu_{j0})^2}{2\sigma_{j0}^2}\right]$$

Entonces, la distribuición a posteriori para los datos es:

$$f(\boldsymbol{\beta}|datos) \propto \prod_{i=1}^{n}\left(\frac{e^{\beta_0+\beta_1 X_{i1}}}{1+e^{\beta_0+\beta_1 X_{i1}}}\right)^{y_i}\left(1-\frac{e^{\beta_0+\beta_1 X_{i1}}}{1+e^{\beta_0+\beta_1 X_{i1}}}\right)^{1-y_i}\prod_{j=0}^{1}\frac{1}{\sigma_{j0}\sqrt{2\pi}}\exp\left[\frac{(\beta_j-\mu_{j0})^2}{2\sigma_{j0}^2}\right]$$

La distribución a posteriori es una forma no cerrada ya que no forma una distribución particular. Por tanto, se necesitan técnicas computacionales para obtener el estimador de Bayes (en este caso $\boldsymbol{\beta}$). Se utilizará simulación MCMC (Cadena de Markov Monte Carlo). Sin embargo, contamos con información previa, la cual nos puede ayudar a elegir las medias y varianzas iniciales para β_j.
Comenzando con la intersección β_0, recordemos la información a priori de que en un día promedio, hay aproximadamente un 20% de probabilidad de lluvia, es decir, $\pi \approx 0.2$. Por lo tanto, establecemos la media a priori para β_0 en la escala $\log(odds)$ en -1.4:

$$\log\left(\frac{\pi}{1-\pi}\right) = \log\left(\frac{0.2}{1-0.2}\right) \approx -1.4.$$

El rango de esta normal anterior indica nuestra vaga comprensión de que el $\log(odds)$ también podría variar entre aproximadamente -2.8 y 0 ($-1.4 \pm 2 \times 0.7$). Más significativamente, creemos que las *odds* de lluvia en un día promedio podrían estar entre 0.06 y 1:

$$\left(e^{-2.8},\ e^{0}\right) \approx (0.06, 1)$$

y, por lo tanto, que la probabilidad de lluvia en un día promedio podría estar entre 0.057 y 0.50 (un rango bastante amplio en el contexto de la lluvia):

$$\left(\frac{0.06}{1+0.06}, \frac{1}{1+1}\right) \approx (0.057,\ 0.50).$$

Ahora, para el coeficiente de humedad β_1, sabemos que la probabilidad de lluvia aumenta hasta un 52% cuando está precedida por un día con mucha humedad o lluvia. Por lo tanto, establecemos la media a priori para β_1 en la escala $\log(odds)$ en 0.08:

$$\log\left(\frac{0.52}{1-0.52}\right) \approx 0.08.$$

Específicamente, en la escala log(*odds*), asumimos que la pendiente β_1 oscila entre 0 y 0.16 ($0.08 \pm 2 \times 0.04$). O, en la escala de *odds*, las odds de lluvia pueden aumentar entre un 0% y un 17% por cada punto porcentual adicional en el nivel de humedad:

$$\left(e^0,\ e^{0.16}\right) \approx (1,\ 1.17).$$

Entonces:

$$\begin{aligned} \text{Datos:}\quad & Y_i|\beta_0,\beta_1 \overset{ind}{\sim} Bern(\pi_i) \quad \text{con} \quad \log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 X_{i1} \\ \text{Priors:}\quad & \beta_0 \sim N(-1.4, 0.7^2) \\ & \beta_1 \sim N(0.08, 0.04^2). \end{aligned}$$

HMC Y Stan

Stan es un programa para generar muestras de una distribución posterior de los parámetros de un modelo, el nombre del programa hace referencia a Stanislaw Ulam (1904-1984) que fue pionero en los métodos de Monte Carlo. A diferencia de JAGS y BUGS, los pasos de la cadena de Markov se generan con un método llamado Monte Carlo Hamiltoniano (HMC). Aunque los detalles difieren, ambos algoritmos son variaciones del algoritmo fundamental de Metropolis-Hastings. HMC es computacionalmente más costoso que Metropolis o Gibbs, sin embargo, sus propuestas suelen ser más eficientes, y por consiguiente no necesita muestras tan grandes. En particular cuando se ajustan modelos grandes y complejos (por ejemplo, con variables con correlación alta) HMC supera a otros.

Simulación De La A Priori

Primero, con la ayuda del paquete `rstanarm`, usamos la función `stan_glm()` con `prior_PD=True` para simular 20,000 conjuntos de parámetros (β_0, β_1) de los modelos a priori. Al hacerlo, especificamos `family=binomial` para indicar que el nuestro es un modelo de regresión logística con una estructura de datos especificada por un modelo de Bernoulli/binomial. Primero debemos especificar la información del modelo mediante:

- `formula`: en nuestro caso sería `raintomorrow`∼`humidity`.
- `data`: nuestro conjunto de datos.
- `family`: como se trata de un modelo de regresión logística, indicamos `binomial`.

En segundo lugar, debemos especificar la información de la Cadena de Markov Monte Carlo (MCMC) deseada utilizando tres argumentos adicionales:

- `chains`: número de cadenas. En nuestro caso, elegimos 4 cadenas, ya que producen aproximaciones estables y ejecutar más cadenas no producen aproximaciones drásticamente mejores.
- `iter`: número de iteraciones o la longitud de cada cadena. De forma predeterminada, la primera mitad de estas iteraciones se eliminan como muestras de "calentamiento". La segunda mitad se conserva como muestra final de la cadena de Markov. Se especifican 10,000 iteraciones.
- `seed`: semilla de generación de números aleatorios.

Este objeto incluye cuatro cadenas de Markov paralelas que se ejecutan durante 10,000 iteraciones cada una. Después de descartar las primeras 5000 iteraciones de las cuatro cadenas, terminamos con cuatro muestras separadas de cadenas de Markov de tamaño 5000, o una muestra combinada de cadenas de Markov de 20,000.

```
# Librerías
library(rstanarm)
library(bayesplot)
library(tidybayes)
library(tidyverse)
```

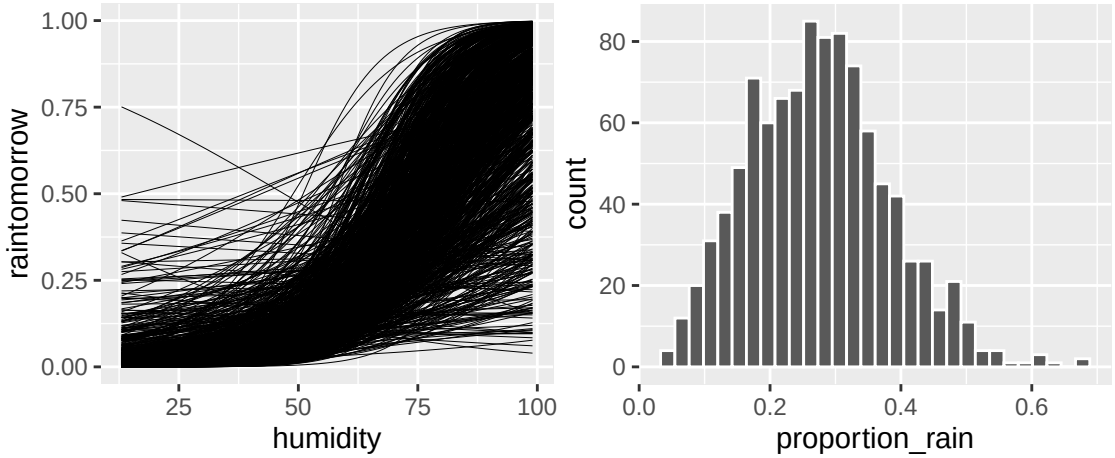
```
# Simulación de la apriori
rain_model_prior <- stan_glm(raintomorrow ~ humidity,
                             data = weather, family = binomial,
                             prior_intercept = normal(-1.4, 0.7),
                             prior = normal(0.08, 0.04),
                             chains = 4, iter = 5000*2, seed = 123,
                             prior_PD = TRUE)
```

Cada uno de los 20,000 pares anteriores plausibles resultantes de β_0 y β_1 describe una relación plausible a priori entre la probabilidad de lluvia mañana y la humedad de hoy:

$$\pi = \frac{e^{\beta_0+\beta_1 X_1}}{1+e^{\beta_0+\beta_1 X_1}}.$$

Graficamos solo 100 de estas relaciones plausibles a priori a continuación, para cada uno, trazamos la relación entre la probabilidad de lluvia y el nivel de humedad del día anterior (izquierda) y la proporción observada de días en los que llovió (derecha), estos reflejan adecuadamente nuestra comprensión previa de que la probabilidad de lluvia aumenta con la humedad.
El porcentaje de días en los que llovió varió desde aproximadamente el 5% en un conjunto de datos hasta aproximadamente el 50% en otro. Esto coincide adecuadamente con nuestra comprensión apriori y la incertidumbre sobre la lluvia en Perth. Por el contrario, si nuestras predicciones anteriores tendieran a centrarse en valores altos, cuestionaríamos nuestro ajuste anterior ya que no creemos que Perth sea un lugar lluvioso.

```
set.seed(123)
# Graficar 100 modelos apriori
plot1 <- weather %>%
  add_fitted_draws(rain_model_prior, n = 1000) %>%
  ggplot(aes(x = humidity, y = raintomorrow)) +
    geom_line(aes(y = .value, group = .draw), size = 0.1)
# Proporción observada de lluvia para los 100 conjuntos a priori
plot2 <- weather %>%
  add_predicted_draws(rain_model_prior, n = 1000) %>%
  group_by(.draw) %>%
  summarize(proportion_rain = mean(.prediction == 1)) %>%
  ggplot(aes(x = proportion_rain)) +
    geom_histogram(color = "white")
plot1; plot2
```



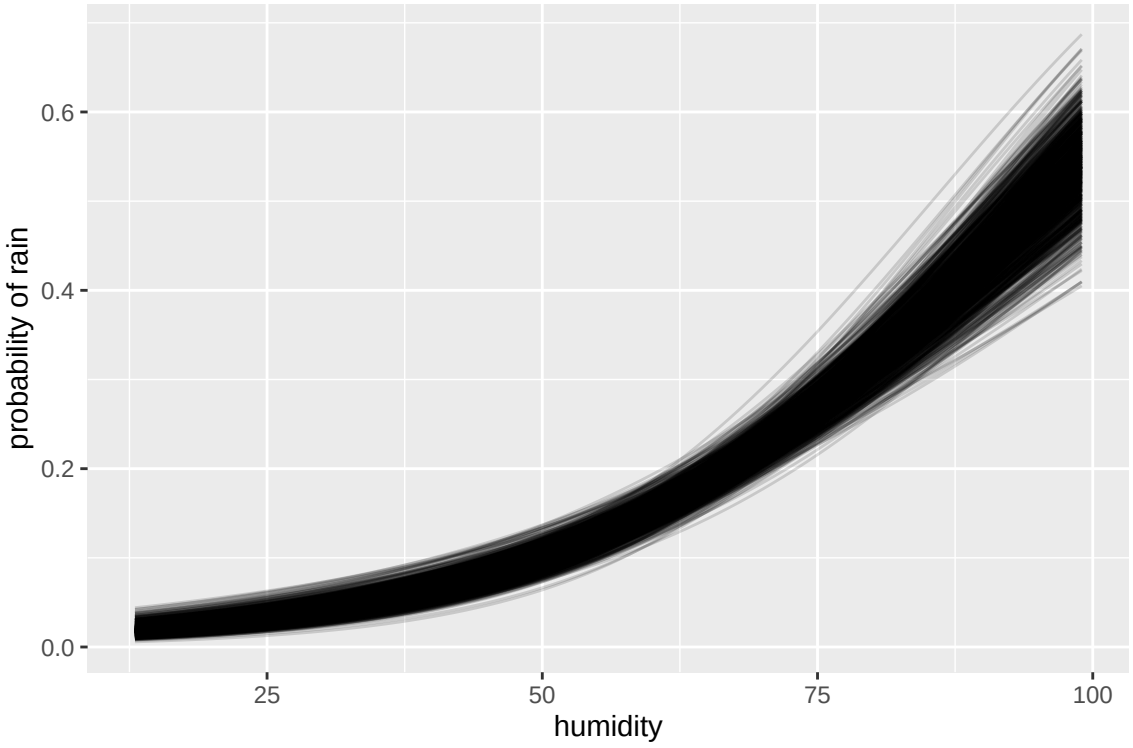
Simulación De La A Posteriori

Para simular el modelo a posteriori de los parámetros β_0 y β_1 de nuestro modelo de regresión logística, podemos actualizar (`update()`) la simulación a priori `rain_model_prior`.

```
# modelo a posteriori
rain_model_1 <- update(rain_model_prior, prior_PD = FALSE)
```

Trazamos 100 modelos plausibles a posteriori en la gráfica siguiente. Naturalmente, estos modelos son mucho menos variables que las contrapartes anteriores, es decir, estamos mucho más seguros acerca de la relación entre la lluvia y la humedad. Ahora entendemos que la probabilidad de lluvia aumenta constantemente con la humedad, sin embargo, no es hasta una humedad de aproximadamente el 95% que alcanzamos el punto de inflexión cuando la probabilidad de lluvia es mayor. Por el contrario, cuando la humedad de hoy está por debajo del 25%, es muy poco probable que llueva mañana.

```
weather %>%
  add_fitted_draws(rain_model_1, n = 1000) %>%
  ggplot(aes(x = humidity, y = raintomorrow)) +
    geom_line(aes(y = .value, group = .draw), alpha = 0.15) +
    labs(y = "probability of rain")
```



Información más precisa sobre la relación entre la humedad y la lluvia se encuentra en el parámetro β_1, cuyo modelo a posteriori aproximado se resume a continuación:

```
# resúmenes a posteriori en la escala log(odds)
posterior_interval(rain_model_1, prob = 0.80)
```

```
                     10%         90%
(Intercept) -5.08202038 -4.14574058
humidity     0.04159425  0.05493655
```

```
# resúmenes a posteriori en la escala de odds
exp(posterior_interval(rain_model_1, prob = 0.80))
```

```
                     10%        90%
(Intercept) 0.006207355 0.01583171
humidity    1.042471414 1.05647358
```

Por cada aumento de un punto porcentual en la humedad de hoy, existe un 80% de probabilidad posterior de que el $log(odds)$ aumente entre 0.0416 y 0.0549. Esta tasa de aumento es menor que nuestra media anterior de 0.08 para β_1 (la probabilidad de lluvia aumenta significativamente con la humedad, pero no en el grado que habíamos anticipado). Más significativamente, por cada aumento de un punto porcentual en la humedad de hoy, las odds de lluvia aumentan entre un 4.2% y un 5.6%:

$$(e^{0.0416},\ e^{0.0549}) = (1.042,\ 1.056).$$

Predicción Y Clasificación

Más allá de utilizar nuestro modelo de regresión logística bayesiana para comprender mejor la relación entre la humedad de hoy y la lluvia de mañana, también queremos predecir si lloverá o no mañana. Por ejemplo, suponga que está en Perth hoy y experimentó una humedad del 99%. Para predecir si lloverá mañana, podemos aproximar el modelo predictivo a posteriori para los resultados binarios de Y, llueva o no, dónde:

$$Y|\beta_0, \beta_1 \sim Bern(\pi) \quad \text{con} \quad \log\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 * 99.$$

Para ello, la función `posterior_predict()` simula 20,000 resultados de lluvia Y, uno por cada uno de los 20,000 conjuntos de parámetros de nuestra cadena de Markov:

```
# Predicciones a posteriori
set.seed(123)
binary_prediction <- posterior_predict(
  rain_model_1, newdata = data.frame(humidity = 99))
```

Como sugiere el nombre, un objetivo común en un análisis de clasificación es convertir nuestras observaciones de la probabilidad pronosticada de lluvia (π) o el resultado pronosticado de lluvia (Y) en una clasificación binaria, sí o no de Y. De este modo:

```
# Resultados de las predicciones a posteriori
table(binary_prediction)
```

```
 binary_prediction
    0     1
 9189 10811
```

```
colMeans(binary_prediction)
```

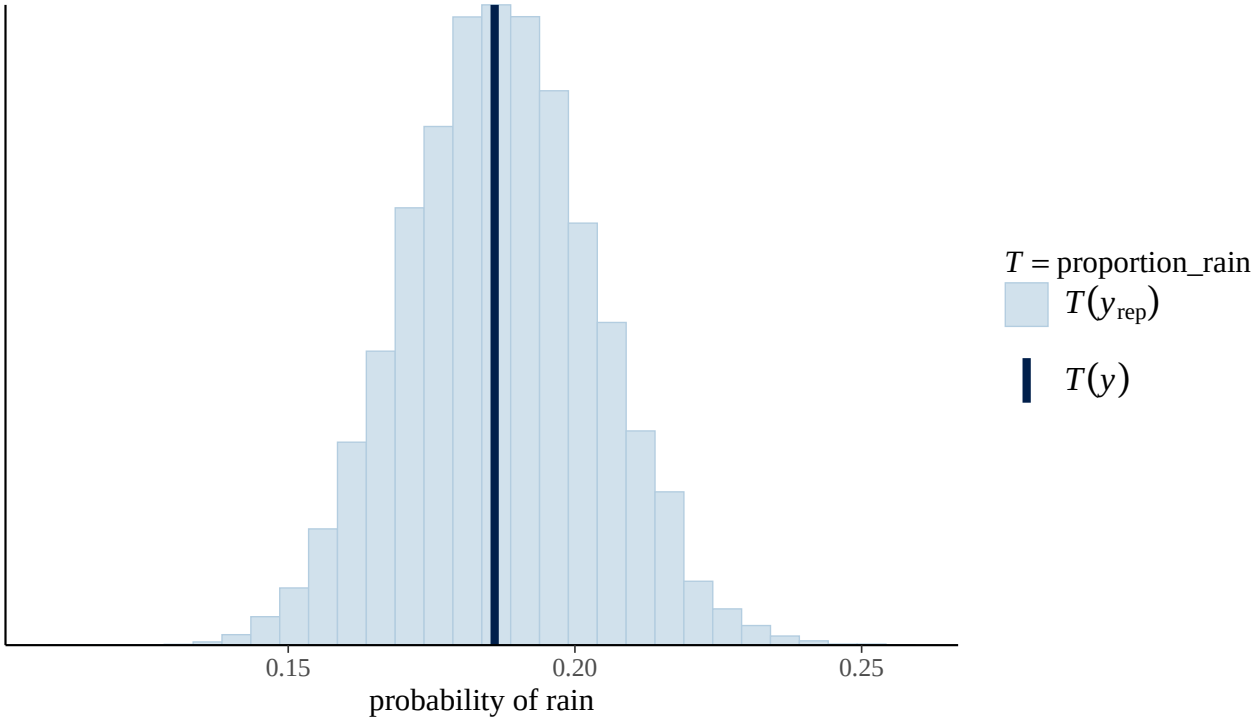
```
       1
0.54055
```

Entre nuestras 20,000 predicciones posteriores de Y, 10811 (o 54.05%) pidieron lluvia. Por lo tanto, dado que la lluvia era más probable que la ausencia de lluvia en el modelo predictivo posterior de Y, es razonable clasificar Y como "1" (lluvia).

Evaluación Del Modelo

Es fundamental evaluar la calidad de un modelo de clasificación, entonces, nos haríamos la siguiente pregunta: ¿Qué tan equivocado está el modelo? Entonces, primero grafiquemos 100 conjuntos de los datos simulados a posteriori, registramos la proporción de resultados Y que son 1, es decir, la proporción de días en los que llovió, utilizando la función `ratio_rain()`. Un histograma de estas tasas de lluvia simuladas confirma que son consistentes con los datos originales. La mayoría de nuestros conjuntos de datos simulados a posteriori vieron lluvia en aproximadamente el 18% de los días, cerca de la incidencia de lluvia observada en los datos meteorológicos, sin embargo, algunos vieron lluvia en tan solo el 12% de los días o hasta el 24% de los días.

```
proportion_rain <- function(x){mean(x == 1)}
pp_check(rain_model_1, nreps = 100,
         plotfun = "stat", stat = "proportion_rain") +
  xlab("probability of rain")
```



Otra pregunta frecuente a la hora de evaluar un modelo es: ¿Qué tan precisas son las clasificaciones a posteriori del modelo?
Comencemos evaluando las clasificaciones de `rain_model_1` de los mismos datos meteorológicos que usamos para construir este modelo. Para comenzar, construímos modelos predictivos a posteriori de Y para cada uno de los 1000 días en el conjunto de datos meteorológicos:

```
# Modelos predictivos a posteriori para cada día en el conjunto de datos
set.seed(123)
rain_pred_1 <- posterior_predict(rain_model_1, newdata = weather)
dim(rain_pred_1)
```

```
[1] 20000  1000
```

Cada una de las 1000 columnas en `rain_pred_1` contiene 20,000 predicciones 1 o 0 de si lloverá o no en el día correspondiente en los datos. La media de cada columna indica la proporción de estas predicciones que son 1; por lo tanto, las medias de 1000 columnas estiman la probabilidad de lluvia para los 1000 días correspondientes en nuestros datos. Luego podemos convertir estas proporciones en clasificaciones binarias

de lluvia comparándolas con un límite de clasificación elegido. Comenzaremos con un límite de 0.5: si la probabilidad de lluvia excede 0.5, entonces se predice lluvia.

```
weather_classifications <- weather %>%
  mutate(rain_prob = colMeans(rain_pred_1),
         rain_class_1 = as.numeric(rain_prob >= 0.5)) %>%
  select(humidity, rain_prob, rain_class_1, raintomorrow)
```

Finalmente, para estimar la precisión a posteriori general de nuestro modelo, podemos comparar las clasificaciones del modelo (`rain_class_1`) con los resultados observados (`raintomorrow`) para cada uno de nuestros 1000 días de muestra. Esta información se resume en la siguiente tabla o "matriz de confusión":

```
# Matriz de confusión
weather_classifications %>%
  tabyl(raintomorrow, rain_class_1) %>%
  adorn_totals(c("row", "col"))
```

```
   raintomorrow   0  1 Total
             No 805  9   814
            Yes 170 16   186
          Total 975 25  1000
```

Observe que nuestra regla de clasificación, junto con nuestro modelo bayesiano, clasificó correctamente 821 de los 1000 casos de prueba totales (805 + 16). Por lo tanto, la tasa de precisión de la clasificación general es del 82.1% (817/1000). A primera vista, ¡esto parece bastante bueno!, sin embargo, nuestro modelo es mucho mejor para anticipar cuándo no lloverá que cuándo lloverá. Entre los 814 días en los que no llueve, clasificamos correctamente 805, o el 98.89%.
Y esto lo podemos comprobar con la función `classification_summary()`:

```
set.seed(123)
classification_summary(model = rain_model_1, data = weather, cutoff = 0.5)
```

```
 $confusion_matrix
   y   0  1
  No 805  9
 Yes 170 16

 $accuracy_rates

 sensitivity      0.08602151
 specificity      0.98894349
 overall_accuracy 0.82100000
```

CONCLUSIONES

Como hemos visto, nuestro modelo bayesiano tiene un nivel de precisión aceptable, además de que la regresión logística bayesiana tiene la ventaja de que nos brinda una distribución a posteriori en lugar de una estimación de un solo punto como en el enfoque clásico, también llamado frecuentista.
Aunque en este tipo de ejemplos relativamente simples, no hay mucha diferencia entre las inferencias bayesianas y las frecuentistas, en modelos más complejos es cuando la regresión bayesiana puede llegar a tener mejor rendimiento, más aún si se cuenta con información previa.

REFERENCIAS BIBLIOGRÁFICAS

- Jaakkola, T.S. & amp; Jordan, M.I.. (1997). A Variational Approach to Bayesian Logistic Regression Models and their Extensions. Proceedings of the Sixth International Workshop on Artificial Intelligence and Statistics in Proceedings of Machine Learning Research. Disponible en https://proceedings.mlr.press/r1/jaakkola97a.html. Reeditado por PMLR el 30 de marzo de 2021.
- William DuMouchel. "Multivariate Bayesian Logistic Regression for Analysis of Clinical Study Safety Issues." Statist. Sci. 27 (3) 319 - 339, August 2012. Disponible en https://doi.org/10.1214/11-STS381
- Joyee Ghosh. Yingbo Li. Robin Mitra. "On the Use of Cauchy Prior Distributions for Bayesian Logistic Regression." Bayesian Anal. 13 (2) 359 - 383, June 2018. Disponible en https://doi.org/10.1214/17-BA1051