Modelo Linear bayesiano

joseferson da silva barreto

.

Professor(a): Fenanda Clotilde

Projeto final Inferência bayesiana apresentado na última aula da disciplina .

dezembro,07,2022

Sumário

.

.

Objetivo

O objetivo é apresentar de forma clara oque são os modelos lineares bayesianos , e verificar sua importância para análises estatísticas.

Método

Foi feito os tratamentos dos dados para buscar relações entre as variáveis , a limpeza dos dados foi feita no software Rstudio ,utilizando a linguagem R para as demais análises. Para essa análise foi utilizado um banco de dados sobre os prejuízos causados as seguradoras pelos furacões nos Estados Únidos de 1954 a 1984.

Introdução

Os modelos Bayesianos são utilizados quando estamos buscando explicar o desconhecido atrávez de distribuições de probabilidades de modo subjetivo ,ou seja, a inferência Bayesiana é uma metodologia estatística baseada na definição de probabilidade como um grau de informação. Neste artigo veremos um modelo de poisson bayesiano generalizado.

Matematicamente o modelo Linear Generralizado de poisson é aplicado quando temos a ocorrência de eventos discretos ao longo de intervalos especificos, entrentanto, para utilizar os modelos de poisson é necessário verificar se não tem excesso de zeros e se nossos dados não estão tão dispersos, além disso nossos dados tem que ser discretos .

- sem exceso de zeros e com dispersão: poisson com dispersão e binomial negativa sem/com dispersão

Os dados utilizados nesse artigo é de uma pesquisa feita em 2014 sobre os prejuízos causados as seguradoras pelos furacões que ocorre regularmente nos estados únidos, esse dataset é de 1954 a 1984.

Carregando o Banco de Dados

Para carregar nosso conjunto de dados vamos utilizar os comandos a baixo, nosso dataset utilizado nesse arquivo está disponível no guithub

```
library (readr)
urlfile="https://raw.githubusercontent.com/metodosexatos/mlgbayes/main/DatasetsES15/hurricanes.csv"
mydata<-read_csv2(url(urlfile)) # para csv no formato brasileiro use read_csv2#head(mydata)
```

Verificando os Histograma da Priori

A distribuição a priori de Poisson é definida de forma caso uma variável aleatória discreta Y tem uma distribuição de Poisson com parâmetro $\theta > 0$ se sua função de massa de probabilidade satisfaz:

$$f(y|\theta) : \frac{e^{-\theta}\theta^{y}}{y!}, \quad y = 0, 1, 2...$$

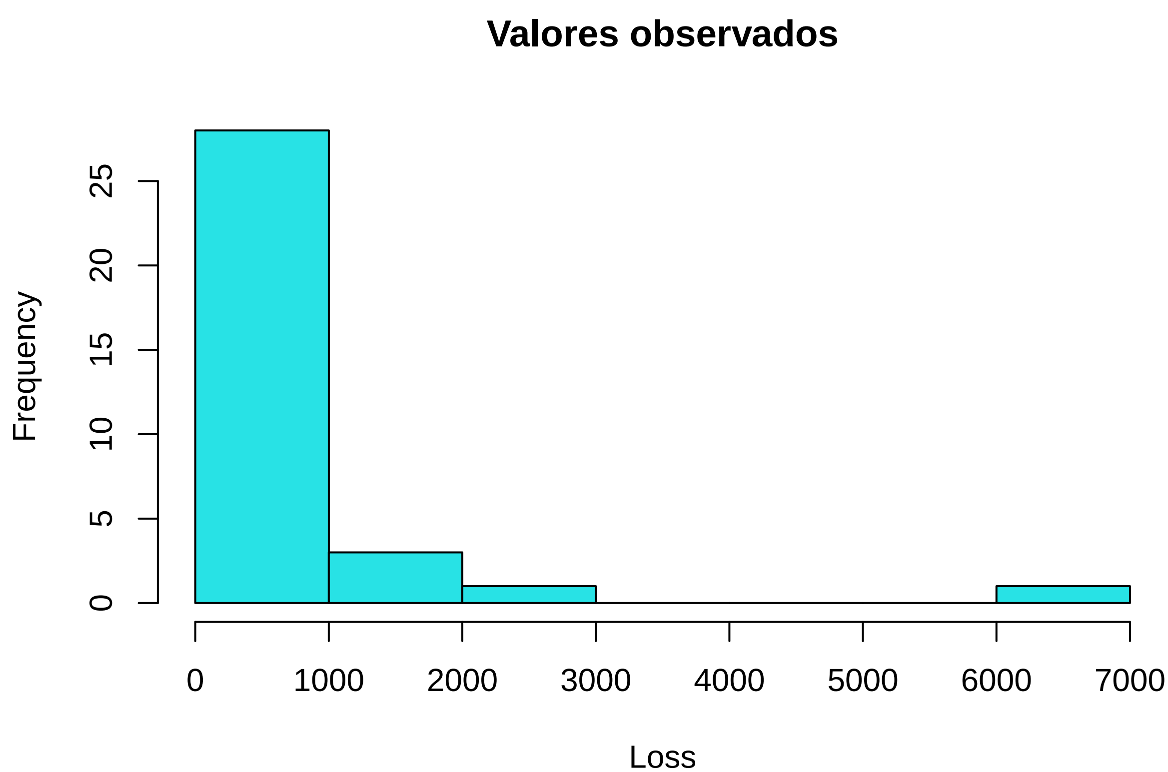
Se Y tem um Poisson(θ) distribuição então

$E(y) = \theta$
$Var(y) = \theta$

Vamos observar a nossa distribuição apriori atravéz do histograma gerado com o código abaixo:

```
#------------
# Histograma da vari?vel dependente:

k <- round(1+3.3*log10(nrow(mydata)),0) # N?mero de classes: Regra de Sturges
hist(mydata$loss, main = "Valores observados", xlab = "Loss", nclass = k, col = 5)
```




```
mean(mydata$loss)
```

```
## [1] 563.4545
```

Como temos poucas observações nós não conseguimos observar muito bem se ele segue a distribuição de poisson, sabemos que só temos informação sobre o conjuto dos naturais

Criação do Modelo

Para criação do modelo vamos utilizar o pacote **rstanarm** com a função stan_glm que criará nosso modelo

```
library(rstanarm)


model_poisson <- stan_glm(loss ~ hurr, data = mydata, family = poisson())
```

Para ver as informações sobre o modelo gerado vamos utilizar a função **summary**

```
summary(model_poisson)
```

```
##
## Model Info:
##  function:     stan_glm
##  family:       poisson [log]
##  formula:      loss ~ hurr
##  algorithm:    sampling
##  sample:       4000 (posterior sample size)
##  priors:       see help('prior_summary')
##  observations: 33
##  predictors:   2
##
## Estimates:
##              mean   sd   10%   50%   90%
## (Intercept) 6.0    0.0  6.0   6.0   6.0
## hurr        0.3    0.0  0.3   0.3   0.3
##
## Fit Diagnostics:
##            mean   sd    10%   50%   90%
## mean_PPD 563.4    5.9 556.0 563.4 571.1
##
## The mean_ppd is the sample average posterior predictive distribution of the outcome variable (for de
##
## MCMC diagnostics
##              mcse Rhat n_eff
## (Intercept)   0.0  1.0  1826
## hurr          0.0  1.0  2451
## mean_PPD      0.1  1.0  2447
## log-posterior 0.0  1.0  1504
##
## For each parameter, mcse is Monte Carlo standard error, n_eff is a crude measure of effective sample
```

Aqui temos algumas informações importantes sobre o modelo:
function: mostra a função utilizada para criação do modelo(stan_glm)
family: mostra a família de distribuição utilizada para geração do modelo e o respectivo lik de ligação, no nosso caso foi a poisson com o link(log)
formula: mostra a formula com as variáveis usandas no modelo
algorithm: mostra o método utilizado para criação do modelo, no nossoo caso foi amostragem
sample: mostra o número de observações geradas na nossa posteriori, temos um total de 4000 observações geradas

observations: mostra o número de observações utilizadas para a geração do modelo

predictors: mostra o número de variaveis utilizadas para a criação do modelo

Ainda temos outras informações como média e desvio padrão dos estimadores ee dos valores preditos ,mas vamos focar na análise dos coeficientes

Análisando os Coeficientes

```
coeff <- round(exp(model_poisson$coefficients),2)
coeff
```

```
## (Intercept)        hurr
##      404.94        1.29
```

Podemos ver que o nosso intercpto é 404.91 e o $\beta_1 = 1.29$

considerando que o intercepto foi 404.91 isso representa que a perda esperada de ínicio é de apróximadamente 400 dolares na média e o $\beta_1 = 1.29$ significa que a cada furacão que acontece por ano nós esperamos um acréssimo, uma perda de uma seguradora de 1.29,ou seja, essa é uma taxa associada a a cada furacão a mais que vai ocorrer, por exemplo, se ocorrem 3 furacões pegasse 1.29 e multiplica-se por 3 furacões

Histograma da Distribuição Posterior

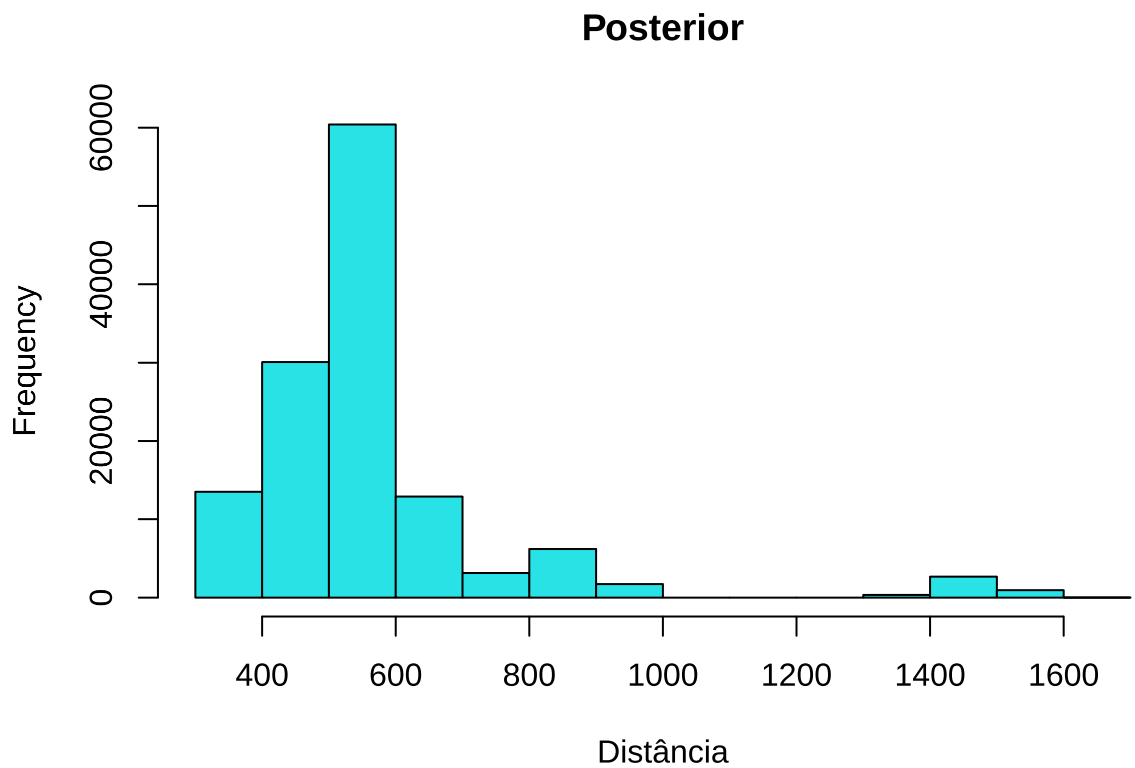
Por definição se $Y \sim Poisson(\theta)$ a priori , logo Y a posteriori será de forma $Y \sim Poisson(\theta)$,ou seja,

$$f(y|\theta) : \frac{e^{-\theta}\theta^y}{y!}, \quad y = 0, 1, 2...$$

Se Y tem um Poisson(θ) distribuição então

$E(y) = \theta$
$Var(y) = \theta$

```
k <- round(1+3.3*log10(nrow(posterior_predict(model_poisson))),0) # N?mero de classes: Regra de Sturges
hist(posterior_predict(model_poisson), main = "Posterior", xlab = "Distância", nclass = k, col = 5)
```



Agora podemos observar um comportamento mais parecido com a poisson ,ela começa com um leve crescimento, ela continua até atingir o ponto máximo e depois começa a cair ,Vamos análisar o intervalor de credibilidade.

Intervalo de Credibilidade

```
exp(posterior_interval(model_poisson))
```

```
##                     5%        95%
## (Intercept) 397.277445 412.252406
## hurr          1.281583   1.305576
```

segundo o intervalo de credibilidade é esperado que tenhamos por exemplo um acontecimento por ano por furacão uma perda por seguradora de 397 a 412 milhões de dólares.

Conclusão

Como podemos ver os modelos bayesianos vem sendo amplamente utilizados devido às suas vantagens e desenvolvimento de computadores, principalmente em cenários onde há Falta de métodos claros para incluir dados(conhecimentos) existentes e para lidar com incerteza nos métodos frequentistas. Demos uma breve introdução ao modelo de poisson bayesiano generalizado, que utilizamos para prever o impacto dos furacões a seguradoras, ou seja, o seu prejuízo econômico,registrando por ano um prejuízo entre 397 a 412 milhoes de dólares.

Referências

EFL Amaral, INÁCIO Magna - Modelos Bayesianos

Dogucu, Mine, Alicia Johnson e Miles Ott. 2022. Regras de Bayes! Uma Introdução à Modelagem Bayesiana Aplicada . 1ª ed. Boca Raton, Flórida: Chapman; Salão/CRC. https://www.bayesrulesbook.com/ .

Kruschke, John. 2015. Fazendo análise de dados bayesiana: um tutorial com r, JAGS e Stan. 2ª ed. Imprensa Acadêmica. https://sites.google.com/site/doingbayesiandataanalysis/ .

GELMAN, Andrew. Objections to Bayesian statistics. Bayesian Analysis, v. 3, n. 3, pp. 445–449, 2008.

Gelman, Andrew; Carlin, John B.; Stern, Hal S.; Dunson, David B.; Vehtari, Aki; Rubin, Donald B. Bayesian data analysis. 3rd ed. Chapman & Hall/CRC, Boca Raton, 2014.