Universidade Federal de Minas Gerais (UFMG)

Departamento de Estatística - ICEx

Isabelle Fernandes de Oliveira Sannier (2021432208)

Implementação Bayesiana via Stan do modelo de Regressão Poisson

Belo Horizonte, 24 de junho de 2025

Exercício

Os gerentes de uma empresa resolveram desenvolver uma pesquisa coletando dados sobre o número de defeitos observados na superfície de um tipo de peça produzida pelo setor de fabricação. Suponha que a variável Y_i representa o número de defeitos registrados na peça i (Y_i é uma contagem e seus valores possíveis são 0, 1, 2, . . .). Dados referentes a uma amostra de tamanho n=300 foram coletados, ou seja, i=1,2, . . . , 300. Além de Y_i, a base de dados também contém duas covariáveis: X_{1i} = covariável binária (1 = usou maquinário novo na fabricação, 0 = usou maquinário antigo) e X_{2i} = anos de experiência do funcionário que operou a máquina de fabricação (unidade de medida: anos/10).

Será admitido a distribuição:

$Y_i \sim \text{Poisson}(\theta_i)$

Sendo $\theta_i > 0$ a média de defeitos esperados na superfície de uma peça i. Também, será utilizado regressão Poisson para estabelecer uma relação entre as covariáveis (X_{1i}, X_{2i}) e a resposta Y_i.

$\ln(\theta_i) = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i}$

Carregando dados

```
dados <- read.table("DadosDefeitos.txt")
n = 300 # tamanho amostral
q <- 3 # 3 betas a estimar
y <- dados$V1
x <- cbind(rep(1, n), dados$V2, dados$V3)
```

Especificação a priori

A informação inicial sobre $\beta = (\beta_0, \beta_1, ..., \beta_k)^T$ será expressa através da distribuição Normal Multivariada $\beta \sim N_q(m_\beta, S_\beta)$.

```
set.seed(123)

m_beta <- rep(0, q)  # Vetor de medias
S_beta <- 10 * diag(q) # Matriz de covariancia
```

Priori especificada:

$\boldsymbol{\beta} \sim N_3(\mathbf{0}_3, 10\ \mathbf{I}_3)$

Transmitindo informações para o Stan

```
data <- list(n = n, q = q, y = y, x = x, m_beta = m_beta, S_beta = S_beta)
```

```
# Lista requisitando que beta e theta sejam salvos.
pars = c("beta", "theta")
```

```
# Lista de sementes de inicialização
init = list()
init[[1]] = list(beta = rep(0, q))
init[[2]] = list(beta = runif(q,-1, 1))
```

```
iter = 2000 # Total de iterações (incluindo burn-in).
warmup = 1000 # Numero de iterações do burn-in.
chains = 2 # Numero de cadeias do MCMC.
```

```
// Bloco de declaração de dados

data{
  int<lower=1> n;
  int<lower=1> q;
  int<lower=0> y[n];
  matrix[n,q] x;
  vector[q] m_beta;
  matrix[q,q] S_beta;
}

// Bloco de declaração de parâmetros

parameters{
  vector[q] beta;
}

// Bloco de parâmetros transformados

transformed parameters{
  vector[n] theta;
  for(i in 1:n){
    theta[i] = exp(x[i,] * beta);
  }
}

// Bloco do modelo
```

```
model{
  // Verossimilhança
  for(i in 1:n){
    y[i] ~ poisson(theta[i]);
  }

  // Priori:  Normal Multivariada com vetor de medias e matriz de covariâncias
  beta ~ multi_normal(m_beta, S_beta);

}
```

```r
aux = stan_model(file = "RegPoissonStan.stan", verbose = FALSE)
output <- sampling(aux, data = data, iter = iter, warmup = warmup,
                   chains = chains, pars = pars, init = init, verbose = FALSE)
```

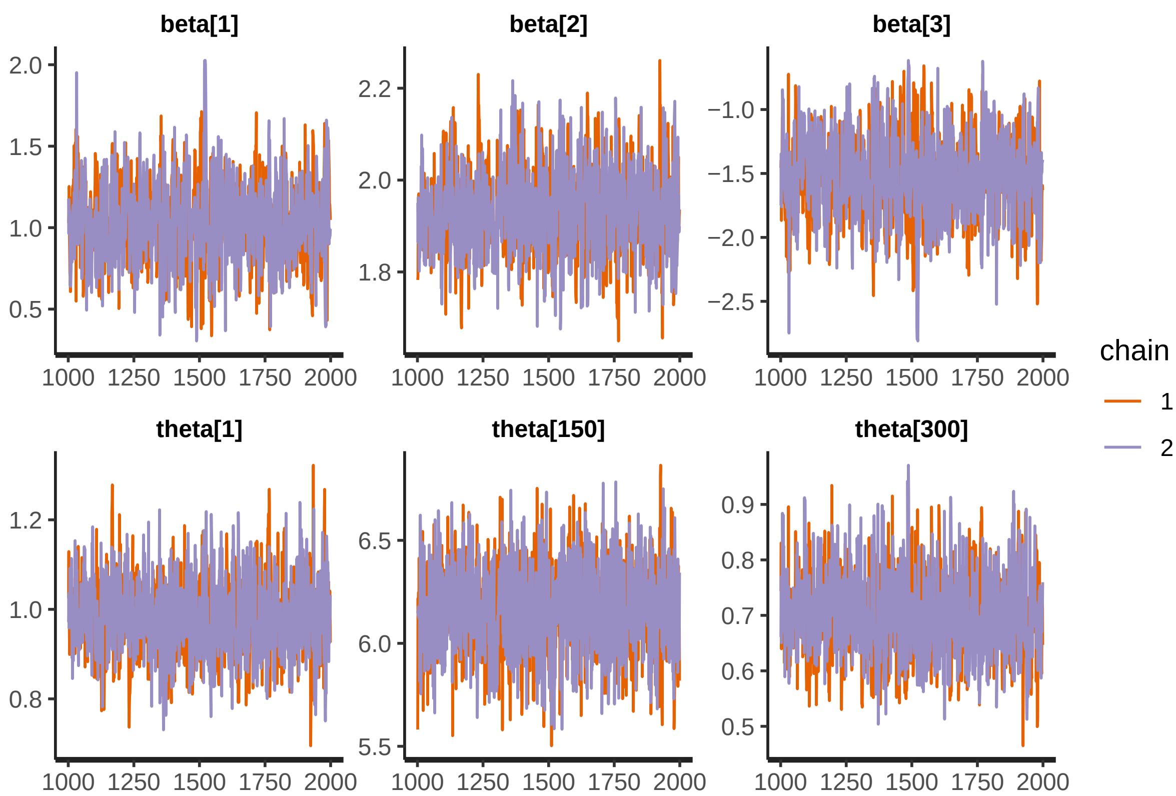
Explorando os resultados

```r
# Sumario global do objeto stan fit.
# print(output, pars = c("beta","theta"))
print(output, pars = c(paste0("beta[",c(1,2,3),"]"),
                       paste0("theta[",c(1,150,300),"]")))
```

```
## Inference for Stan model: anon_model.
## 2 chains, each with iter=2000; warmup=1000; thin=1;
## post-warmup draws per chain=1000, total post-warmup draws=2000.
##
##             mean se_mean   sd  2.5%   25%   50%   75% 97.5% n_eff Rhat
## beta[1]     1.03    0.01 0.25  0.55  0.87  1.03  1.19  1.54   644 1.00
## beta[2]     1.93    0.00 0.09  1.76  1.87  1.93  1.99  2.13   712 1.00
## beta[3]    -1.53    0.01 0.33 -2.18 -1.75 -1.54 -1.32 -0.90   641 1.00
## theta[1]    0.98    0.00 0.08  0.82  0.92  0.98  1.03  1.15   777 1.00
## theta[150]  6.16    0.01 0.21  5.75  6.02  6.17  6.31  6.57  1585 1.00
## theta[300]  0.70    0.00 0.07  0.57  0.65  0.70  0.75  0.85   779 1.01
##
## Samples were drawn using NUTS(diag_e) at Tue Jun 24 18:49:17 2025.
## For each parameter, n_eff is a crude measure of effective sample size,
## and Rhat is the potential scale reduction factor on split chains (at
## convergence, Rhat=1).
```

Visualmente, e pela tabela acima (estatística Rhat) indica que as cadeias convergiram. Também, na tabela acima, são obtidas as principais estatísticas das estimativas dos parâmetros obtidas pelo algoritmo NUTS.

```r
rstan::traceplot(output, pars = c("beta","theta[1]",
                                  "theta[150]", "theta[300]"))
```




```r
samp = extract(output)
```

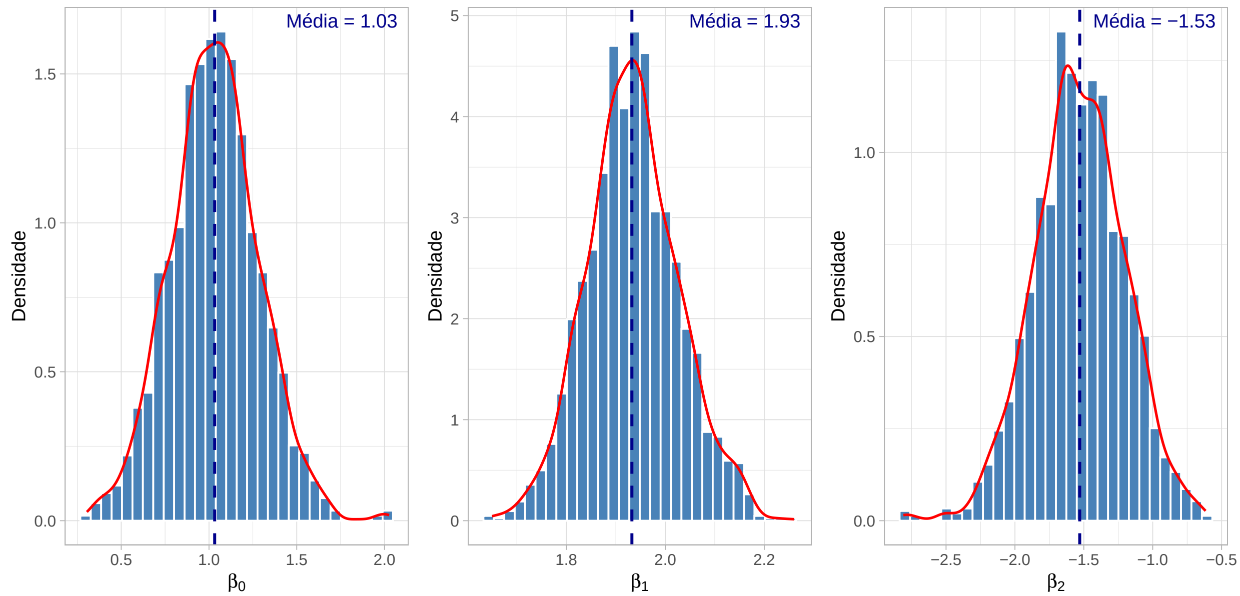


Figura 1: Histograma a posteriori dos parâmetros coeficientes betas da regressão poisson estimados

O histograma acima foi construído para cada parâmetro β. Os valores estimados foram obtidos realizando o cálculo da média do resultado obtido pelo algorítmo NUTS, retirando os primeiros 1000 valores (warm-up). Em linha vertical tracejada está localizada a estimativa no parâmetro em sua distribuição a posteriori.

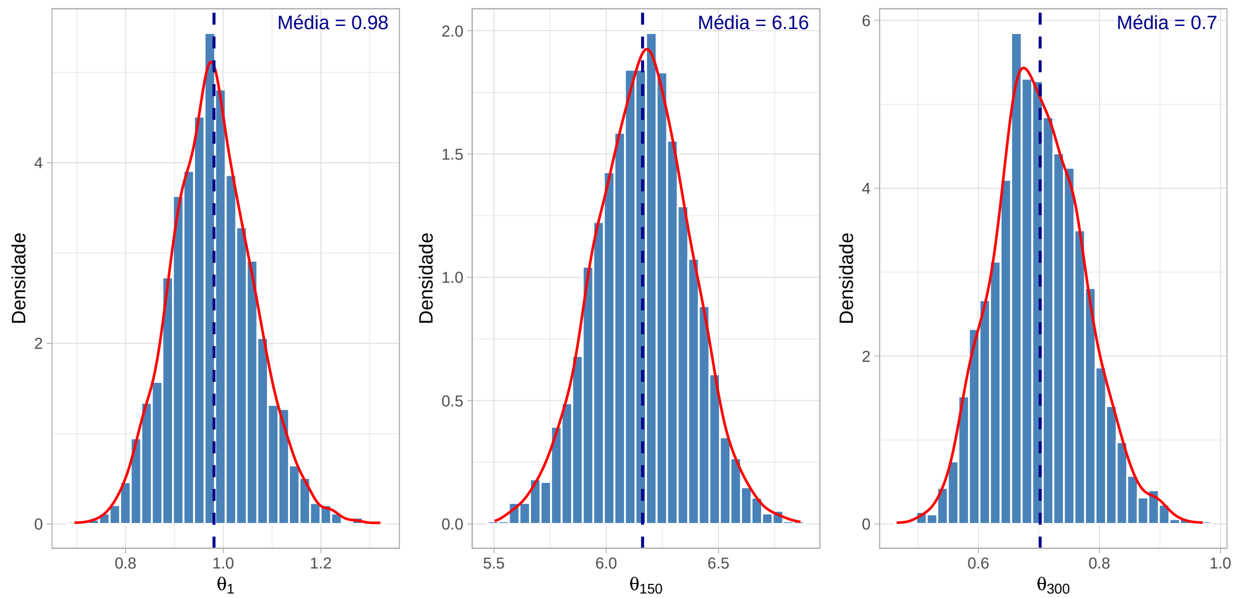


Figura 2: Histograma a posteriori das médias thetas de defeitos estimadas para as peças 1, 150 e 300

O histograma acima foi construído para as médias de defeitos estimadas θ_i para as peças i = 1, 150 e 300. Os valores estimados foram obtidos realizando o cálculo da média do resultado obtido pelo algoritmo NUTS, retirando os primeiros 1000 valores (warm-up).

```
calc_moda <- function(x) {
  d <- density(x)
  d$x[which.max(d$y)]
}

beta0 <- data_beta$beta0[1001:2000]
beta1 <- data_beta$beta1[1001:2000]
beta2 <- data_beta$beta2[1001:2000]
theta1 <- data_theta$theta1[1001:2000]
theta150 <- data_theta$theta150[1001:2000]
theta300 <- data_theta$theta300[1001:2000]


tabela <- tibble::tibble(
  Parâmetro = c("beta0", "beta1", "beta2", "theta1", "theta150", "theta300"),
  Média = c(mean(beta0), mean(beta1), mean(beta2),
            mean(theta1), mean(theta150), mean(theta300)),
  Moda = c(calc_moda(beta0), calc_moda(beta1), calc_moda(beta2),
           calc_moda(theta1), calc_moda(theta150), calc_moda(theta300)),
  Mediana = c(median(beta0), median(beta1), median(beta2),
              median(theta1), median(theta150), median(theta300)),
  `Desvio padrão` = c(sd(beta0), sd(beta1), sd(beta2),
                      sd(theta1), sd(theta150), sd(theta300)),
  `HPD 2.5%` = c(
    HPDinterval(as.mcmc(beta0), prob = 0.95)[1],
    HPDinterval(as.mcmc(beta1), prob = 0.95)[1],
    HPDinterval(as.mcmc(beta2), prob = 0.95)[1],
    HPDinterval(as.mcmc(theta1), prob = 0.95)[1],
    HPDinterval(as.mcmc(theta150), prob = 0.95)[1],
    HPDinterval(as.mcmc(theta300), prob = 0.95)[1]
  ),
  `HPD 97.5%` = c(
    HPDinterval(as.mcmc(beta0), prob = 0.95)[2],
    HPDinterval(as.mcmc(beta1), prob = 0.95)[2],
    HPDinterval(as.mcmc(beta2), prob = 0.95)[2],
    HPDinterval(as.mcmc(theta1), prob = 0.95)[2],
    HPDinterval(as.mcmc(theta150), prob = 0.95)[2],
    HPDinterval(as.mcmc(theta300), prob = 0.95)[2]
  )
)
```

```r
kable(tabela, digits = 3,
      caption = "Tabela de Resumo da Inferência Bayesiana para os parâmetros
      beta e theta estimados")
```

Tabela 1: Tabela de Resumo da Inferência Bayesiana para os parâmetros beta e theta estimados

Parâmetro	Média	Moda	Mediana	Desvio padrão	HPD 2.5%	HPD 97.5%
beta0	1.033	1.092	1.040	0.255	0.494	1.464
beta1	1.932	1.939	1.932	0.092	1.748	2.116
beta2	-1.530	-1.618	-1.538	0.331	-2.200	-0.946
theta1	0.981	0.974	0.977	0.085	0.826	1.157
theta150	6.162	6.162	6.161	0.210	5.775	6.591
theta300	0.702	0.675	0.694	0.075	0.566	0.848

A tabela apresenta as estatísticas de resumo das distribuições a posteriori para os parâmetros do modelo. Observa-se uma alta consistência nas estimativas de tendência central, com valores muito próximos para a média, moda e mediana em todos os parâmetros, sugerindo distribuições posteriores simétricas.

A hipótese nula $\beta_i = 0$ é rejeitada quando analisados os intervalos de credibilidade, isto é, com 95% de probabilidade o valor verdadeiro de β_i está contido no intervalo proposto e, como ele não inclui zero, isso significa dizer que as covariáveis maquinário novo/não novo e anos de experiência são significativas para explicar o número de falhas.

Especificamente, β_1 tem relação positiva sobre o número de falhas, isto é, se a máquina for nova, espera-se 6,9 (exp^{β_1}) vezes mais defeitos se comparado com a máquina velha. Já para β_2, o seu impacta negativamente no número de falhas, isto é, quanto mais anos de experiência tem o operador da máquina, espera-se menor o número de falhas nas peças. A cada dez anos de experiência, o número de falhas esperados é multiplicado por 0.214 (exp^{β_2}).

As médias de falhas esperadas para as peças 1, 150 e 300 estão expressas na tabela em theta1, theta150 e theta300 respectivamente. Também, observa-se que há mais certeza nas estimativas de beta1, theta1 e theta300 por apresentarem menor desvio padrão.