Universidade de Brasília

DEPARTAMENTO DE CIÊNCIA DA COMPUTAÇÃO

2 de dezembro de 2022

Lista 2: Inferência estatística via simulação.

Prof. Guilherme Rodrigues

Métodos Computacionais Intensivos para Mineração de Dados

Programa de pós-graduação em Computação Aplicada (PPCA)

(A) **As questões deverão ser respondidas em um único relatório *PDF* ou *html*, produzido usando as funcionalidades do *Rmarkdown* ou outra ferramenta equivalente.**

(B) **O aluno poderá consultar materiais relevantes disponíveis na internet, tais como livros, *blogs* e artigos.**

(C) **O trabalho é individual. Suspeitas de plágio e compartilhamento de soluções serão tratadas com rigor.**

(D) **Os códigos *R* utilizados devem ser disponibilizados na integra, seja no corpo do texto ou como anexo.**

(E) **O aluno deverá enviar o trabalho até a data especificada na plataforma Microsoft Teams.**

(F) **O trabalho será avaliado considerando o nível de qualidade do relatório, o que inclui a precisão das respostas, a pertinência das soluções encontradas, a formatação adotada, dentre outros aspectos correlatos.**

(G) **Escreva seu código com esmero, evitando operações redundantes, comentando os resultados e usando as melhores práticas em programação.**

"O Monte Rainier é um estratovulcão, e a montanha mais alta do estado norte-americano de Washington. (...) Sua altitude é de 4392m e, em dias de tempo claro, seu pico permanentemente nevado pode ser facilmente avistado de Seattle e outras cidades da região." (*wikipédia*)

Um conjunto de dados sobre tentativas de se escalar o Monte Rainier está disponível no site *Kaggle*, e pode ser obtido pelo link https://www.kaggle.com/codersree/mount-rainier-weather-and-climbing-data/version/3.

Usaremos Modelos Lineares Generalizados para descrever como o número de montanhistas que alcançam o cume do monte em um dado dia (sucessos) varia em função da temperatura média do ar (em graus Celsius).

A seguir apresentamos a estrutura do banco de dados.

```
str(dados, width = 60, strict.width = "cut")
```

```
## tibble [1,889 x 10] (S3: tbl_df/tbl/data.frame)
##  $ Data            : Date[1:1889], format: "2015-11-27" ...
##  $ Sucessos        : num [1:1889] 0 0 0 0 0 0 0 0 0 0 ...
##  $ Route           : chr [1:1889] "Disappointment Cleaver"..
##  $ Tentativas      : num [1:1889] 2 3 2 8 2 10 2 2 2 2 ...
##  $ Temperatura     : num [1:1889] -3.155 -0.389 8.027 4.98..
##  $ Umidade_relativa: num [1:1889] 19.7 21.7 27.2 28.3 74.3..
##  $ Velocidade_vento: num [1:1889] 27.84 2.25 17.16 19.59 6..
##  $ Direc_vento     : num [1:1889] 68 118 259 280 265 ...
##  $ Radiacao_solar  : num [1:1889] 88.5 93.7 138.4 176.4 27..
##  $ Cleaver         : logi [1:1889] TRUE TRUE TRUE FALSE TR..
```

Considere o modelo

$$\begin{aligned} Y_{ij} &\sim \text{Poisson}(\lambda_{ij}) \\ \lambda_{ij} &= \exp(\alpha + \beta t_i), \end{aligned}$$

onde Y_{ij} representa o número de montanhistas que atingiram o cume pela rota j no dia i, α e β são parâmetros desconhecidos do modelo e t_i indica a temperatura média no dia i. Para uma determinada temperatura, o modelo prevê o mesmo valor para todas as rotas. Desconsidere os dados da rota *glacier only - no summit attempt*. Por fim, note que para obter o valor de y_{ij} é preciso somar todos os sucessos registrados no dia i para a rota j.

Questão 1)

a) Conduza um teste de hipóteses por simulação para avaliar a hipótese nula de que a média do número de sucessos obtidos pela rota "Disappointment Cleaver'' é igual a média das demais rotas (conjuntamente).

b) Obtenha o estimador de máxima verossimilhança de α e β considerando o modelo proposto. Dica: Use a função `optim` do `R` para achar o ponto que maximiza a log-verossimilhança.

c) Estime a distribuição de probabilidade do número de sucessos previstos para um dia em que a temperatura seja de 15 graus.

d) Construa um intervalo de confiança de 95% para $\exp(\beta)$ a partir do método de bootstrap paramétrico. Interprete o resultado considerando o contexto dos dados. Dica: calcule o aumento percentual da média esperada quando a temperatura aumenta em 1 grau Celsio.

e) Faça um diagnóstico do modelo via simulação. Para tanto, gere dados sintéticos usando o modelo obtido no item b), ajuste um novo modelo sobre os dados sintéticos e calcule o Erro quadrático médio (MSE). Repita esse procedimento 10000 vezes e compare os MSEs gerados com aquele do modelo obtido em b). Comente os resultados.

Questão bônus)

Use o método de integração por Monte Carlo para estimar o volume de uma elipsoide definida por

$$\frac{x^2}{2} + \frac{y^2}{3} + \frac{z^2}{4} = 1.$$

Anexo

Código usado para organizar o banco de dados.

```
library(readr)
require(tidyverse)
require(broom)
require(lubridate)
library(corrplot)

climbing <- read_csv("climbing_statistics.csv")
weather <- read_csv("Rainier_Weather.csv")
convert <- function(x) (x-32) * 5/9
shift <- function(x) x - mean(x)
dados <- inner_join(climbing, weather) %>%
  select(-matches("Percent|Battery")) %>%
  filter(Attempted >= Succeeded) %>%
  mutate(`Temperature AVG`= convert(`Temperature AVG`),
         Cleaver = Route=="Disappointment Cleaver",
         Date = mdy(Date)) %>%
  select(Date, Succeeded, everything()) %>%
  rename(Data = Date,
         Sucessos = Succeeded,
         Tentativas = Attempted,
         Temperatura = `Temperature AVG`,
         Umidade_relativa = `Relative Humidity AVG`,
         Velocidade_vento = `Wind Speed Daily AVG`,
         Direc_vento = `Wind Direction AVG`,
         Radiacao_solar = `Solare Radiation AVG`)
```