# CEFET/RJ
# Bacharelado em Ciência da Computação
# GCC1625 - Inferência Estatística
# Trabalho 04

Prof. Eduardo Bezerra (ebezerra@cefet-rj.br)

Novembro/2023

# Sumário

# 1 Teste de Permutação

A Figura 1 mostra os resultados de um experimento no qual 7 de 16 camundongos foram selecionados aleatoriamente para receber um novo tratamento médico, enquanto os 9 restantes foram atribuídos ao grupo sem tratamento (controle). O tratamento tinha como objetivo prolongar a sobrevivência após uma cirurgia de teste. Em particular, a coluna "Data" mostra o tempo de sobrevivência após a cirurgia, em dias, para todos os 16 camundongos. Essa mesma figura também apresenta, para cada amostra: tamanho, média, desvio padrão.

Utilize o teste de permutação para responder à seguinte pergunta de pesquisa (use nível de significância igual a 5%): *O tratamento prolongou a sobrevivência?*. Você deve apresentar a declaração das hipóteses, descreva como calculou a estatística de deste e o $p$-valor, a finalmente apresente sua conclusão.

| Group | Data | | | Sample Size | Mean | Estimated Standard Error |
|---|---|---|---|---|---|---|
| Treatment | 94 | 197 | 16 | | | |
| | 38 | 99 | 141 | | | |
| | 23 | | | (7) | 86.86 | 25.24 |
| Control | 52 | 104 | 146 | | | |
| | 10 | 50 | 31 | | | |
| | 40 | 27 | 46 | (9) | 56.22 | 14.14 |
| | | | | Difference: | 30.63 | 28.93 |

Figura 1: Mouse data - retirada de Efron and Tibshirani [1994]

# 2 Bootstrap

Essa parte do trabalho é uma adaptação do Problema 9 na seção 5.4 de *An Introduction to Statistical Learning*[1]. O conjunto de dados usado aqui é o denominado Boston dataset. Uma descrição desse conjunto de dados pode ser encontrada em `http://lib.stat.cmu.edu/datasets/boston`.

(i) Com base neste conjunto de dados, forneça uma estimativa pontual para a média populacional da variável medv. Chame essa estimativa $\hat{\mu}$.

[1] `https://www.statlearning.com`

(ii) Forneça uma estimativa do erro padrão de $\hat{\mu}$. Interprete o resultado.

(iii) Agora estime o erro padrão de $\hat{\mu}$ usando o método Bootstrap. Como essa estimativa se compara com sua resposta de (ii)?

(iv) Com base em sua estimativa de bootstrap de (iii), forneça um intervalo de confiança de 95% para a média de medv. Compare-o com os resultados obtidos usando t-test sobre o atributo `medv`.

(v) Com base neste conjunto de dados, forneça uma estimativa, $\hat{\mu}_{med}$, para a mediana populacional de `medv`.

(vi) Agora você deve estimar o erro padrão de $\hat{\mu}_{med}$. Infelizmente, não há uma fórmula simples para calcular o erro padrão da mediana. Em vez disso, estime o erro padrão da mediana usando o método bootstrap. Comente suas descobertas.

(vii) Forneça uma estimativa para o décimo percentil do atributo `medv`. Chame essa quantidade de $\hat{\mu}_{0.1}$.

(viii) Use o método bootstrap para estimar o erro padrão de $\hat{\mu}_{0.1}$. Comente suas descobertas.

# 3 Teste de Permutação *vs* Bootstrap

Uma empresa quer saber se é eficiente ensinar novas ferramentas aos seus funcionários usando cursos pela internet. A empresa seleciona aleatoriamente 7 trabalhadores e os atribui a dois grupos de tamanhos 4 e 3. O primeiro grupo frequentou aulas tradicionais, e o segundo frequentou cursos pela internet. Após a realização dos cursos, foi aplicado um teste aos trabalhadores, cujos resultados foram:

- Cursos na Internet: 37, 49, 55, 57
- Cursos tradicionais: 23, 31, 46

Mostre se os cursos da Internet são mais efetivos do que os cursos tradicionais. Para isso, aplique um teste de permutação e um teste de bootstrap. Os dois testes levam à mesma conclusão?

# 4 Regressão Linear Múltipla

Considere o conjunto de dados `Auto`[2]. O arquivo `Auto.csv` contém os dados para essa parte do trabalho. Esse arquivo está na plataforma MS Teams.

[2] `https://islp.readthedocs.io/en/latest/datasets/Auto.html`

(i) Produza uma matriz de gráfico de dispersão[3] que inclua todas as variáveis no conjunto de dados.

(ii) Calcule a matriz de correlações entre as variáveis usando a função corr() do `pandas.DataFrame`. Você precisará excluir a variável name, que é qualitativa.

(iii) Use a função ols() da biblioteca *statsmodels* para realizar uma regressão linear múltipla com `mpg` como resposta e todas as outras variáveis (exceto nome) como os preditores. Use a função summary() para imprimir os resultados. Comente sobre a saída. Por exemplo:

  (a) Existe uma relação entre os preditores e a resposta?

  (b) Quais preditores parecem ter um valor estatisticamente significativo com relação à resposta?

  (c) O que sugere o coeficiente correspondente à variável ano?

(iv) Use a função `regplot`[4] da biblioteca seaborn para produzir gráficos de diagnóstico do ajuste de regressão linear. Comente sobre quaisquer problemas que você encontrar com o ajuste.

# 5 DiD: Salário mínimo x taxa de empregos

Nesta parte, você irá replicar um estudo realizado originalmente por Card and Krueger [1994] sobre o efeito do aumento do **salário mínimo** sobre a **taxa de empregos**[5]. A teoria econômica convencional sugere que num mercado de trabalho com concorrência perfeita, um aumento no salário mínimo leva a um aumento no desemprego. Em abril de 1992, o estado americano de Nova Jersey (NJ) aumentou o salário mínimo (por hora) de US$ 4,25 para US$ 5,05. Card e Krueger (1994) utilizaram a técnica *Difference-in-Difference* (DiD) e mostraram que este aumento nos salários mínimos levou a um aumento no emprego no setor dos restaurantes de *fast food*. O grupo de controle utilizado nesse estudo foi o estado vizinho da Pensilvânia (PA), que não foi sujeito a essa mudança de política. Os autores realizaram uma pesquisa antes e depois do aumento do salário mínimo com uma amostra representativa de restaurantes de fast food em NJ e PA. Esta configuração pode ser considerada quase experimental, uma vez que ambos os estados não são idênticos em muitos aspectos e o processo legislativo, para aumentar o salário mínimo, não foi iniciado ao acaso.

O arquivo `card_krueger_1994_mod.csv` contém os dados para essa parte do trabalho. Esse arquivo está na plataforma MS Teams.

[3] `https://seaborn.pydata.org/examples/scatterplot_matrix.html`

[4] `https://seaborn.pydata.org/tutorial/regression.html`

[5] `https://davidcard.berkeley.edu/papers/njmin-aer.pdf`

(i) Reproduza o gráfico apresentado na Figura 1 do estudo supra-mencionado.

(ii) Calcule a estimativa DiD usando a abordagem de computar a diferença das médias. Ou seja, neste item você não deve usar a statsmodels para realizar a regressão. Use apenas Python (ou R) para computar as médias da variável de interesse para os dois grupos, antes e depois da intervenção.

(iii) Agora compute novamente a estimativa DiD, dessa vez utilizando a regressão linear. Para isso, você deve inicialmente criar duas variáveis *dummy*. Um indica o início do tratamento (tempo) e é igual a zero antes do tratamento e igual a um após o tratamento. A outra variável separa as observações em grupo de tratamento e grupo controle (tratado). Essa segunda variável *dummy* é igual a um para restaurantes *fast food* localizados em NJ e igual a zero para restaurantes *fast food* localizados no PA. Em seguida, crie a variável de interação multiplicativa. Finalmente, use a biblioteca statsmodels para gerar o modelo de regressão linear. Apresente sua análise e interpretação do resultado obtido.

# 6 DiD: validação

A validade da abordagem diferença-em-diferenças baseia-se na suposição de que há **tendências iguais** (*equal trends*) nos grupos de controle e de tratamento. De acordo com essa suposição, na ausência da intervenção (programa, tratamento), não existiriam diferenças variáveis no tempo entre os grupos de tratamento e de controle. Embora esta suposição não possa ser provada, sua validade pode ser avaliada de quatro maneiras:

1. Comparar repetidamente as mudanças nos resultados dos grupos de tratamento e controle antes de o programa ser implementado (ou seja, em t-3, t-2, t-1). Se a tendência dos resultados se mover em paralelo antes do início do programa, teria provavelmente continuado a mover-se em conjunto na ausência do programa.

2. Fazer um teste de placebo usando um grupo de tratamento falso. O grupo de tratamento falso deveria ser um grupo que não foi afetado pelo programa. Um teste placebo que revela impacto zero apoia a suposição de tendência igual.

3. Fazer um teste de placebo usando um resultado falso. Um teste placebo que revela impacto zero apoia a suposição de tendência igual.

4. Executar a estimativa de diferenças em diferenças usando diferentes grupos de comparação. Estimativas semelhantes do impacto do programa confirmam a suposição de tendência igual.

Sua tarefa nesta parte do trabalho é revisitar o conjunto de dados denominado `Panel101.dta`, que foi usado no exercício realizado em aula. Ao realizar esse exercício, dividimos os países em dois grupos (controle e tratamento) usando o ano 1994 como ponto no tempo em que houve a intervenção. Use os dados relativos aos anos anteriores a 1994 para realizar o teste de validação descrito no item 1 acima.

# 7 Consumo de álcool x taxa de mortalidade

As estatísticas relacionadas com o efeito do consumo de álcool são preocupantes, desde as elevadas taxas de mortalidade por acidentes de trânsito até problemas de saúde, especialmente entre os jovens adultos. Nesta parte, você irá replicar um estudo realizado por Carpenter and Dobkin [2009] sobre o efeito do consumo de álcool nas taxas de mortalidade[6]. Os dados a serem usados podem ser obtidos em `http://masteringmetrics.com/wp-content/uploads/2015/01/AEJfigs.dta`. A Tabela 1 descreve as colunas desse conjunto de dados.

| **Variável** | **Descrição** |
|---|---|
| `agecell` | Idade do indivíduo (o estudo se concentra em adultos entre 19 e 22 anos) |
| `all` | Taxa de mortalidade geral |
| `alcohol` | Taxa de mortalidade por causas relacionadas ao álcool |
| `homicide` | Taxa de mortalidade por homicídios |
| `suicide` | Taxa de mortalidade por suicídio |
| `mva` | Taxa de mortalidade por acidentes de carro |
| `drugs` | Taxa de mortalidade por causas relacionadas a drogas (excluindo álcool) |
| `externalother` | Taxa de mortalidade por outras causas externas |

Tabela 1: Descrição dos dados usados no artigo Carpenter e Dobkin (2009).

A descontinuidade de regressão é um desenho apropriado para estudar estas questões, uma vez que os jovens adultos são "naturalmente" selecionados em dois grupos com base na sua idade: os jovens adultos com menos de 21[7] anos não estão legalmente autorizados a beber, enquanto os jovens adultos com mais de 21 anos são legalmente proibidos de beber. Podemos comparar a taxa de mortalidade entre esses dois grupos.

(i) Reproduza o gráfico apresentado na Figura 3 do estudo acima mencionado.

[6] `https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2846371/pdf/nihms68174.pdf`

[7] Na maior parte dos EUA, a maioridade é obtida aos 21 anos, diferente do Brasil, no qual esse valor é de 18 anos.

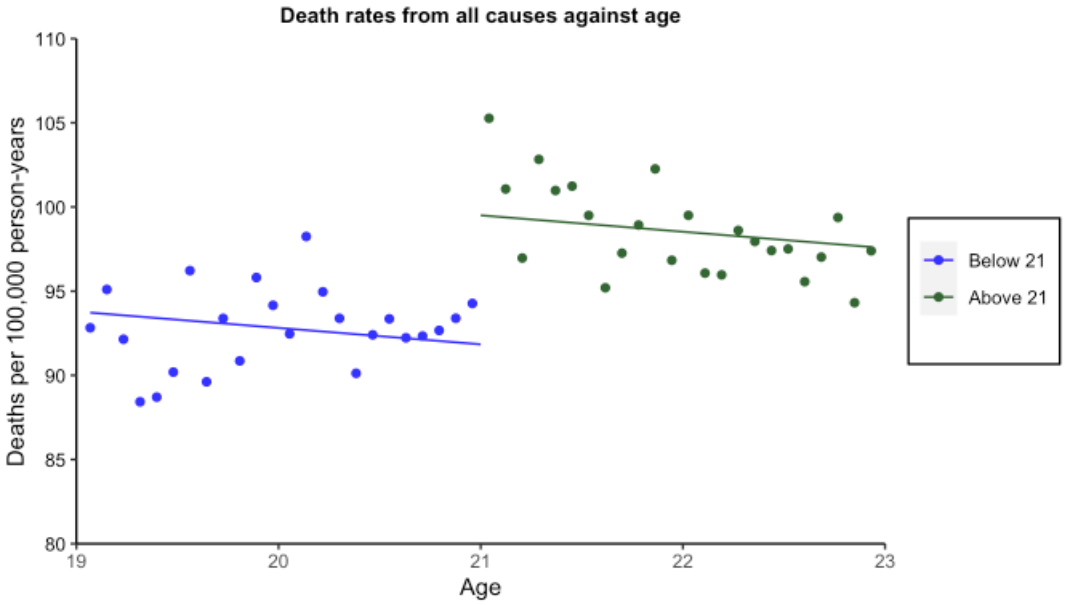


Figura 2: Taxas de mortalidade (todas as causas) versus idade.

(ii) Execute uma regressão descontínua para "todas" as mortes por idade (não me refiro a todas as variáveis, apenas à variável chamada `all`). Analise os resultados. Como você usa esses resultados para estimar a relação entre consumo de álcool e mortalidade? ***Nota****: O conjunto de dados fornecido possui menos do que 50 observações. Sendo assim, não espere reproduzir exatamente os resultados das tabelas do artigo, que usa um conjunto de dados completo de 1.500 observações. Além disso, você também não tem as mesmas variáveis.*

(iii) Produza o gráfico todas as variáveis por idade e adicione as linhas de regressão definidas pelo resultado da regressão (não há problema se as linhas se estenderem por toda a figura. O gráfico que você deve produzir aqui deve ser semelhante ao apresentado na Figura 2.

# O que deve ser entregue

Você pode desenvolver esse trabalho em duas linguagens alternativas, `R` ou `Python`. Independente da linguagem que escolher, você deve preparar explicar sua implementação, análise e conclusões de cada parte desse trabalho. Além disso, certifique-se de fornecer respostas para cada uma das perguntas formuladas em cada parte deste trabalho.

Seu trabalho deve ser necessariamente produzido como um único *notebook* Jupyter[8]. Como sugestão, você pode usar a plataforma Google Colab[9] para produzir seu trabalho. Essa plataforma permite criar *notebooks* em ambas as linguagens.

Você deve necessariamente organizar seu *notebook* em seções que reflitam as seções apresentadas no enunciado deste trabalho. Sendo assim, use como ponto de partida o exemplo apresentado na Figura 3. Repare que, para cada item do trabalho, você deve transcrever o enunciado correspondente para o notebook. Deve também criar duas células, uma de código e outra de texto, para apresentar a solução para o item.

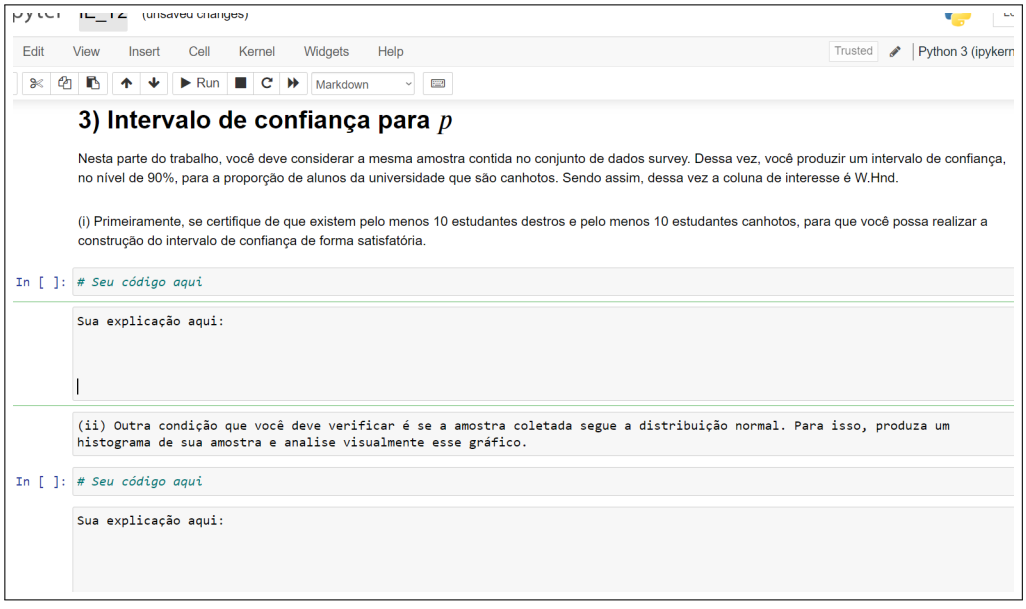


Figura 3: Modelo a ser seguido para apresentação da solução de cada parte do trabalho.

**IMPORTANTE**: Tão relevante quanto a implementação (seja em R ou Python) de cada parte deste trabalho é sua explicação sobre ela. Nesse sentido, você deve também apresentar suas análises e conclusões para cada item do trabalho.

[8] `http://jupyter.org/`
[9] `https://colab.research.google.com`

- Um item que apresente apenas código (em R ou em Python), sem a explicação do mesmo, não receberá a totalidade da pontuação correspondente.
- Um item que apresente apenas um valor numérico como resposta (ou apenas um "sim" ou "não"), sem uma descrição sobre como a resposta foi obtida, não receberá a totalidade da pontuação correspondente.

O *notebook* Jupyter resultante do seu trabalho deve ser necessariamente definido com nome que siga o padrão `IE_T4_SEU_NOME_COMPLETO.ipynb`. Um exemplo: `IE_T4_EDUARDO_BEZERRA_DA_SILVA.ipynb`. Siga à risca esse padrão de nomenclatura.

# Referências


David Card and Alan B Krueger. Minimum Wages and Employment: A Case Study of the Fast-Food Industry in New Jersey and Pennsylvania. *American Economic Review*, 84(4):772–793, September 1994. URL `https://ideas.repec.org/a/aea/aecrev/v84y1994i4p772-93.html`.

Christopher Carpenter and Carlos Dobkin. The effect of alcohol consumption on mortality: Regression discontinuity evidence from the minimum drinking age. *American Economic Journal: Applied Economics*, 1(1):164–82, January 2009. doi: 10.1257/app.1.1.164. URL `https://www.aeaweb.org/articles?id=10.1257/app.1.1.164`.

Bradley Efron and Robert J Tibshirani. *An Introduction to the Bootstrap*. CRC press, 1994.