Lego 2 - 1ª Lista de Exercícios

2025 – 2º Semestre

Data para entrega: 03/10/2025, até 23:00hs, via Google Classroom (impreterivelmente)

Professor: Rogério J Barbosa

Monitor: Rodrigo Roll

Instruções Gerais:

i. Esse exercício é dividido em **duas partes**: uma conceitual-matemática; outra prática, com exercícios de análise de dados e programação.
ii. A parte conceitual deve ser feita à mão
iii. A parte prática deve ser feita em RMarkdown impreterivelmente:
    a. Os códigos e resultados devem ser incluídos
    b. Suprimir "messages" e "warnings" de todos os chunks de código
    c. Trata-se de um documento feito para seres humanos. Então, então deve ser bem formatado e incluir apontamentos, comentários e complementos das respostas em português, interpretando os resultados.
iv. Aceitaremos exercícios incompletos. Mas não haverá chance de refazer após a data da entrega
v. Esse exercício é obrigatório inclusive para os "ouvintes" do curso. Para os matriculados, não entregar significa não receber a nota. Para ouvintes, não entregar significa abandonar o curso.

Parte 1: Conceitual

Instruções para a parte conceitual:

i. Como já dito, deve ser feita à mão em letra legível e com boa organização do espaço. Se possível usando lápis e canetas com cores diferentes, para destacar ênfases, quando necessário
ii. Numere as páginas
iii. Digitalize e organize, com páginas em ordem, num único arquivo PDF. Não serão aceitos exercícios que não sigam esse formato.

1) **Utilizando as propriedades do somatório, mostre que a soma dos desvios de uma variável aleatória qualquer X com respeito à sua média $\overline{X}$ é igual à zero.**

2) **Explique com suas palavras, em, no máximo, 2 ou 3 frases curtas, o que é:**
   a. Processo Gerador de Dados (PGD)
   b. A diferença entre parâmetros e estatísticas
   c. A diferença entre erros e resíduos
   d. Pressuposto de Linearidade do modelo de regressão (ML.1)
   e. Pressuposto de que os indivíduos amostrados são independentes e identicamente distribuídos (iid), o segundo pressuposto do modelo de regressão (ML.2)
   f. Pressuposto de que não há colinearidade perfeita entre os regressores (ML.3)
   g. Pressuposto de exogeneidade e média condicional zero (ML.4)
   h. O conceito de homocedasticidade
   i. Pressuposto de Homocedasticidade (ML.5)

3) **Vamos tratar agora da importância da cláusula *ceteris paribus* e do pressuposto de exogeneidade (ML.4). Responda cada item em, no máximo, um parágrafo.**
   a. O que significa a cláusula *ceteris paribus* e por que ela é importante?
   b. Por que a satisfação da exogeneidade permite a interpretação dos coeficientes de regressão como efeitos causais?
   c. Por que a regressão múltipla torna mais fácil satisfazer o pressuposto de exogeneidade?
   d. O que acontece quando violamos o pressuposto de exogeneidade?

4) **Vamos tratar agora da correlação entre os regressores de uma regressão múltipla. Em todas as letras desse exercício, a resposta se relaciona com um mesmo problema de fundo: a noção de colinearidade perfeita, de Álgebra Linear. Responda sempre de forma breve, com no máximo 3 linhas.**

a. Por que não podemos inserir a mesma variável duas vezes como regressor em uma regressão múltipla?
b. Suponha que você tenha uma variável explicativa com duas categorias (tal como sexo). Por que numa regressão múltipla que contém um intercepto/constante não podemos inserir, ao mesmo tempo, como regressores as duas dummies que podem ser geradas a partir dessa variável binária? Em outras palavras, por que devemos deixar uma das variáveis como "categoria de referência"?
c. Se omitimos ou suprimimos a constante passa a ser possível inserir uma dummy indicadora do sexo feminino e, ao mesmo tempo, uma dummy indicadora de sexo masculino. Por que isso ocorre? Por que a omissão da constante torna desnecessária a existência de uma categoria de referência?
d. Qual a diferença entre colinearidade perfeita e multicolinearidade? Por que essa última não é um grande problema e também pode ser chamada de "micronumerosidade"?
e. O que é o fator de inflação da variância (ou Variance Inflation Factor – VIF) e porque ele é um bom diagnóstico para multicolinearidade?

5) **Derive o estimador da Regressão Linear <u>Simples</u> utilizando o Método dos Momentos (i.e. demonstre o passo a passo matemático, demonstre como se chega à formula). <u>Não use vetores e matrizes</u> (objetos típicos de Álgebra Linear) – use apenas da Álgebra que aprendemos no Ensino Médio e das Propriedades do Somatório. Faça todos os passos à mão e indique quando cada um dos pressupostos do Modelo Linear foi utilizado.**

6) **Derive o estimador da Regressão Linear <u>Múltipla</u> utilizando o Método dos Momentos (i.e. demonstre o passo a passo matemático, demonstre como se chega à formula). Use vetores e matrizes. Faça todos os passos à mão e indique quando cada um dos pressupostos do Modelo Linear foi utilizado.**

7) **Derive o estimador da Regressão Linear <u>Simples</u> utilizando o Método dos Mínimos Quadrados. Use Cálculo, derivadas e propriedades do somatório. Faça todos os passos à mão.**

8) **Os estimadores obtidos nos exercícios 5 e 7 são idênticos. Mas quais as diferenças conceituais entre o Método dos Momentos e o Método dos Mínimos quadrados? Responda em, no máximo, um parágrafo.**

Parte 2: Prática

Instruções para a parte prática:

i. Como já dito, deve ser feita em RMarkdown.
ii. Preze pela boa formatação do documento.
    a. Preencha adequadamente o cabeçalho com seu nome, data, nome do documento etc
    b. Use adequadamente os marcadores de títulos e seções (os hashtags #). Jamais escreva tudo com hashtag (sim, já teve quem me entregou trabalho assim...). Diferencie títulos do corpo do texto
    c. Não deixe o output "poluídos" desnecessariamente. Desligue os warnings e messages de todos os chunks de código.
iii. O documento precisa expressar uma análise que é 100% replicável. Isso significa que tudo o que consta como output/resultado foi gerado via código de R que efetivamente está no documento
iv. Você deve entregar o arquivo PDF compilado – e NÃO O RMD. Se você não conseguir gerar o PDF diretamente, você pode gerar um DOCX e depois salvar em PDF.
v. Não serão aceitos meros scripts de R, nem trabalhos feitos diretamente em Word (colando os outputs).

9) Prevendo a Expectativa de Vida ao Nascer nos municípios brasileiros no ano de 1991. Neste exercício utilizaremos o arquivo dados_expectativaVida.csv. **Nesse banco de dados, cada linha representa um município brasileiro, tal como observado no ano de 1991.**

Variável	Significado	Tipo	Unidade de Medida
esperanc	Esperança de Vida ao Nascer	Variável Contínua	Anos
populaca	População do Município	Variável Contínua	Pessoas
ginitrab	Índice de Gini (desigualdade) da renda do trabalho	Variável Contínua	Escala de 0 a 1
renda_me	Renda Domiciliar *per capita* Média do Município	Variável Contínua	R$ de 2010
escoa2_p	Percentual de domicílios que possui esgotamento sanitário no município (escoadouro ligado à rede geral)	Variável Contínua	0 a 100
urbano_p	Percentual de pessoas residentes em áreas urbanas	Variável Contínua	0 a 100
med1000	Número de Médicos por 1000 habitantes	Variável Contínua	Número de Médicos por 1000 habitantes
norte	Variável indicadora da localização do município na região Norte	Variável Categórica	0 ou 1
centrooe	Variável indicadora da localização do município na região Centro-Oeste	Variável Categórica	0 ou 1
sul	Variável indicadora da localização do município na região Sul	Variável Categórica	0 ou 1
nordeste	Variável indicadora da localização do município na região Nordeste	Variável Categórica	0 ou 1
pop15a_f	Percentual de pessoas com 15 anos ou mais que possui ensino fundamental completo	Variável Contínua	0 a 100

a. Construa um histograma para todas as variáveis contínuas do banco de dados. Alguma das variáveis apresentou comportamento estranho?

b. Construa uma única tabela de estatísticas descritivas para todas as variáveis. Em cada, uma das variáveis. Nas colunas: a média, a mediana, o desvio-padrão e o número de casos válidos (i.e. aqueles que não contém *missing*)

c. Vamos rodar as regressões!

   i. Rode o modelo abaixo usando apenas operações de Álgebra Linear, com matrizes. Depois rode novamente, com o comando lm(.) e compare os resultados. Eles devem ser idênticos.

Y	**X**
Expectativa de Vida	Desigualdade de renda

   Interprete os resultados: intercepto, efeito dos coeficientes, erro-padrão dos coeficientes, teste-t, significância estatística dos parâmetros, R2, RMSE (desvio-padrão dos resíduos)

   ii. Rode o modelo abaixo e salve-o num objeto chamado **reg1**

Y	**X**
Expectativa de Vida	Desigualdade de renda x 100

   A diferença deste modelo para o anterior é unicamente a mudança na escala do Coeficiente de Gini. Agora ele varia de 0 a 100 (ao invés de 0 a 1) – e passa a ser chamado de **Índice** de Gini. Cada unidade dessa nova variável é chamada de "Pontos no Índice de Gini".

   Responda: por que às vezes pode fazer sentido multiplicar uma variável contínua que varia de 0 a 1 por 100?

   Interprete os resultados deste modelo e compare-os com os do item anterior.

   iii. Rode o modelo abaixo usando apenas operações de Álgebra Linear, com matrizes. Depois rode novamente, com o comando lm(.) e compare os resultados. Eles devem ser idênticos.
   Salve o resultado do comando lm(.) num objeto chamado **reg2**

Y	**X**
Expectativa de Vida	Desigualdade de renda x 100
	Renda domiciliar per capita média

Interprete os resultados deste modelo e compare-os com os do item anterior.

Responda: Por que o efeito do indicador de desigualdade mudou? Qual é a nova interpretação?

iv. A partir de agora, use apenas a função lm(.).Rode o modelo abaixo e salve-o num objeto chamado **reg3**

Y	**X**
Expectativa de Vida	Desigualdade de renda x 100
	Renda domiciliar per capita média
	log(população)

Responda: Por que acha que a variável população entrou na equação como log? Qual a interpretação desse coeficiente?

Interprete os resultados deste modelo e compare-os com os do item anterior.

v. Rode o modelo abaixo e salve-o num objeto chamado **reg4**:

Y	**X**
Expectativa de Vida	Desigualdade de renda x 100
	Renda domiciliar per capita média
	log(população)
	Perc. Domicílios com rede de esgoto
	Perc. população urbana
	Nível educacional da localidade
	Médicos na localidade
	Regiões do Brasil

Interprete os resultados deste modelo e compare-os com os do item anterior.

vi. Apresente os resultados dos modelos reg1, reg2, reg3 e reg4 numa única tabela, usando comando stargazer, do pacote de mesmo nome.

10) Você acha que a regressão reg4, que rodamos no exercício anterior, tem interpretação causal? Justifique sua resposta