Lego 2 - 2ª Lista de Exercícios

2024 – 2º Semestre

Data para entrega: 27/10/2025, até 20:00hs, via Google Classroom (impreterivelmente)

Professor:
Rogério J Barbosa

Monitor:
Rodrigo Roll

Parte 1: Conceitual

Instruções para a parte conceitual:

i. Você pode fazer à mão ou no computador. Se for fazer no computador (recomendado), utilize as ferramentas adequadas para escrever fórmulas/notação matemática (seja no Word, seja no RMarkdown etc.)
ii. Numere as páginas

1) **Explique com suas palavras, em, no máximo, 2 ou 3 frases curtas, o que é:**
   a. O conceito homocedasticidade
   b. O conceito de heterocedasticidade
   c. A idéia de que pode haver correlação entre os erros das observações
   d. Pressuposto de Homocedasticidade (ML.5)
   e. Pressuposto de Normalidade dos Erros (ML.6)
   f. Erros robustos à heterocedasticidade, de Huber-White
   g. Erros clusterizados
   h. Qual a diferença entre colinearidade perfeita e multicolinearidade
   i. Fator de inflação da variância (ou Variance Inflation Factor – VIF)

2) **Sobre o uso dos diferentes tipos de estimadores de variância das estimativas da regressão:**
   a. Quando devemos usar erros homocedásticos (ou "clássicos")?
   b. Quando devemos usar erros heterocedásticos (Huber-White)?
   c. Por que, às vezes, a diferença entre eles e os erros clássicos podem ser um indicativo de má especificação funcional?
   d. Quando devemos usar erros clusterizados?

3) **Derive a fórmula <u>mais geral</u> da matriz de variância-covariância dos coeficientes de regressão. Ou seja, mostre como chegamos nisso aqui:**

$$\text{Var}(\widehat{\boldsymbol{\beta}}) = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\Omega}\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}$$

Isso envolverá:

- Tomar o estimador de $\widehat{\boldsymbol{\beta}}$ (i.e. a equação/fórmula que nos dá os betas) já em forma matricial.
- Substituir $\boldsymbol{y}$ pela equação que o gera (i.e. aquela dada por ML.1, já em forma matricial). Agora a expressão será função do vetor de erros ($\boldsymbol{\epsilon}$) e da matriz $\boldsymbol{X}$
- Aplicar o operador de variância condicional em $\boldsymbol{X}$ dos dois lados da equação $\mathrm{Var}(.|\boldsymbol{X})$. Por que fazemos isso? Por que estamos estimando coeficientes e erros dado que temos o conjunto de dados efetivamente observados. Os erros poderiam ser um pouquinho diferentes, se tivéssemos outro conjunto de dados.
- Podemos sempre "tirar da variância" o que é considerado "constante"/"não variante". E o que é constante sai da variância ao quadrado. No caso de vetores e matrizes, o "quadrado" tem uma forma bem específica. Nesse caso, assumimos que $\boldsymbol{X}$ é constante, pois condicionamos em $\boldsymbol{X}$. Eu sei... $\boldsymbol{X}$ é nossa *design matrix*, a matriz de variáveis explicativas. Pode parecer estranho entendê-la como constante. Se tiver dúvidas, procure um material ou vídeo sobre "variância condicional" ou "esperança condicional".
- Com mais um pouquinho de contas, você chega na fórmula acima

4) **Derive a fórmula dos erros homocedásticos a partir da expressão anterior. Ela envolve aplicar a suposição de homocedasticidade à matriz Ω. E então a expressão simplifica.**

Parte 2: Prática

Instruções para a parte prática:

i. Faça 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 um arquivo PDF compilado diretamente a partir do R – NÃO ENTREGUE O RMD.
v. Se você não conseguir gerar o PDF diretamente, você pode compilar um DOCX a partir do próprio R, ao invés disso. E então depois salve em PDF. NAÕ ENTREGUE EM FORMATO DOCX.
vi. Não serão aceitos meros scripts de R, nem trabalhos feitos diretamente em Word (colando os outputs). Dá pra perceber claramente a diferença

Nos próximos exercícios, utilizaremos os dados do BH *Area Survey*, pesquisa sobre a Região Metropolitana de Belo Horizonte, que foi a campo em 2002, sob coordenação da importantíssima socióloga Neuma Aguiar, então docente do Departamento de Sociologia da UFMG. O "BH Survey", como se convencionou chamar, era parte de um projeto maior, internacional e comparativo, o *Social Hubble* – em alusão ao pesquisador (e, posteriormente, ao telescópio) que revolucionou a astrofísica.

A proposta era investigar temas específicos das Ciências Sociais em profundidade, de modo que as pesquisas oficiais como a PNAD ou a PME, vigentes à época, não cobriam. Para possibilitar o BH, Neuma fundou o Centro de Pesquisas Quantitativas em Ciências Sociais (CPEQS), em 1999, e, associado a ele, o Programa de Metodologia Quantitativa (o famoso "MQ", que era uma Escola de Inverno focada em Metodologia de Survey: amostragem, aspectos cognitivos da elaboração de questionários, estatística básica, modelos multivariados, demografia social etc.).

Os temas investigados eram bastante variados: qualidade de vida e capital social, participação política e associativismo, cultura e valores, religião, classificação racial, polícia e criminalidade, trabalho, estratificação social. Seu questionário era feito para tornar possível a comparação com os outros países do *Social Hubble*. Alguns módulos inspiravam-se ou replicavam formatos de questões já canônicos.

Sua amostra é probabilística em todas as etapas: sorteio de municípios, setores censitários, domicílios e pessoas dentro dos domicílios. Trata-se então de uma amostra complexa. Sua aplicação era presencial, realizada no domicilio. Foi planejada para ter 1270 questionários; porém, em função de recusas e outros problemas de coleta, apenas 1029 casos constam no banco de dados. Ajustes foram feitos nos pesos amostrais para corrigir eventuais vieses de não resposta.

Forneço para vocês, juntamente com esta lista de exercícios, os dados (banco original e uma versão que contém algumas variáveis adicionais, fruto de recodificações – é com essa última que vamos trabalhar), documentação, questionários e dicionário de variáveis.

Desta vez, para abrir os dados, vamos utilizar o pacote foreign, ao invés do rio. Ele permite abrir arquivos SPSS (.sav) com labels – o que às vezes é o que desejamos. Contudo a função `read.spss(.)` gera uma lista, ao invés de um `data.frame`. Temos assim que especificar uma opção adicional:

```
library(foreign)
p <- read.spss(file = paste0(wd, "PRMBH_2002_DadosRecodificados.sav"),
               use.value.labels = T,
               to.data.frame    = T)
```

Como sou uma pessoa legal, já vou passar aqui pra vocês as principais recodificações com que vamos trabalhar:

```
p <- p %>%
      mutate(idade  = as.numeric(E2),
             mulher = ifelse(E1 %in% "Mulher", 1, 0),

             brancos = case_when(RC4A %in% "Branco (a)"  ~ 1,
                                 RC4A %in% "Amarelo (a)" ~ 1,

                                 RC4B %in% "Branco (a)"  ~ 1,
                                 RC4B %in% "Amarelo (a)" ~ 1,
                                 RC4B %in% "Claro(a)"    ~ 1,

                                 TRUE ~ 0),

             anosEstudo = ifelse(ANOSEST %in% "17 anos ou mais", "17", as.character(ANOSEST)),
             anosEstudo = as.numeric(anosEstudo),

             religiao = case_when(
                    R4 %in% "Católico (a) não praticante" ~ "1. Catolico nao-praticante",
                    R4 %in% "Católico (a) praticante"     ~ "2. Catolico praticante",
                    R4 %in% "Evangélico(a)"               ~ "3. Evangelico",
                    R4 %in% "Não tem religião?"           ~ "5. Sem Religiao",
                    TRUE                                  ~ "4. Outra"),

             classes_rendaDom = case_when(
                    RENDA %in% "Até 1 salário mínimo (SM) (até R$ 200,00)" ~ "1. Até 1SM",
                    RENDA %in% "Mais de 1 a 3 SM"                          ~ "2. 1 a 3 SM",
                    RENDA %in% "Mais de 3 a 5 SM"                          ~ "3. 3 a 5 SM",
                    RENDA %in% "Mais de 5 a 10 SM"                         ~ "4. 5 a 10 SM",
                    RENDA %in% "Mais de 10 a 20 SM"                        ~ "5. 10 a 20 SM",
                    RENDA %in% "Mais de 20 SM"                             ~ "6. 20+ SM",
                    TRUE ~ NA_character_),


             aborto  = V9,
             aborto  = ifelse(V9 %in% "Nunca aceitável",  1,  aborto),
             aborto  = ifelse(V9 %in% "Sempre aceitável", 10, aborto),
             aborto  = ifelse(V9 %in% c("NS", "NR", "NA"), NA_character_, aborto),
             aborto  = as.numeric(aborto))
```

- A idade está medida em anos.
- "Mulher" é uma variável *dummy* indicadora do sexo feminino.
- "Brancos" é uma variável *dummy* indicadora da raça autodeclarada, "Branca" ou "Amarela" são 1.
- anosEstudo é a variável de anos de escolarização formal.
- A variável de religião agrupa categorias pequenas, com poucos casos.

- A renda domiciliar total está medida de forma categórica, em classes/faixas.

O propósito do exercício será o de tentar compreender a posição individual quanto ao tema do aborto. Trata-se de um exercício que busca compreender valores que uma pessoa pode ter.

A informação foi coletada da seguinte forma no questionário:

V9. E em relação ao aborto? Você pensa que o aborto é sempre aceitável, nunca é aceitável, ou você tem alguma opinião intermediária? Nesta escala de resposta, 1 significa nunca aceitável e 10 sempre aceitável.

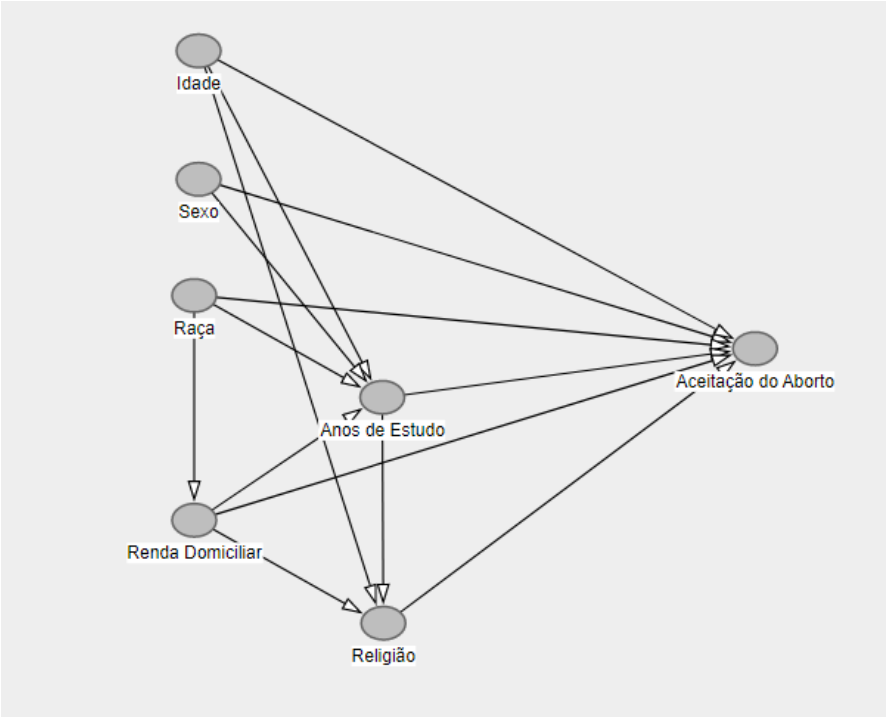
```
|_____|_____|_____|_____|_____|_____|_____|_____|_____|_____|   NR. 77
1     2     3     4     5     6     7     8     9     10
Nunca                                                Sempre      NS. 99
aceitável                                            aceitável
```

Ela gera então uma escala discreta, com 10 pontos, em que quantias maiores significam maior aceitação do aborto. A questão é: o que determina maiores patamares de aceitação?

Algumas variáveis/informações possíveis são aquelas já aventadas acima:

- Suspeito que a idade importe, pois pessoas mais velhas, de gerações anteriores, possivelmente cresceram em ambientes mais conservadores poderão tem menor aceitação
- Homens e mulheres muito provavelmente terão opiniões distintas – uma vez que a interrupção de uma gravidez diz respeito ao corpo das mulheres e também porque são movimentos feministas os maiores responsáveis pela defesa do "direito de decidir".
- Quanto à raça, não tenho palpite muito informado... Suspeito que mulheres negras estão entre as maiores vítimas de fatalidades provocadas por métodos precários de aborto. Mulheres brancas possivelmente acessam clínicas ou fármacos que tornam o procedimento mais seguro pra si. Mas seria esse um efeito de raça ou de classe!?
- A educação está associada tanto ao maior acesso à informação (eventualmente sobre métodos contraceptivos e de abroto), como também a eventuais mudanças no sistema de valores culturais.
- O efeito da religião parece bastante óbvio: indivíduos mais praticantes ou adeptos de denominações mais conservadoras tendem a ser contra o aborto – trata-se de um conteúdo discutido e tratado nas próprias reuniões de grupos religiosos.
- Para tentar dissociar o que é efeito da renda do que seria o da raça ou da escolarização, vale incluir uma informação sobre o *status* socioeconômico dos domicílios.

Talvez um possível esquema de processo gerador seria o seguinte:



5) **Apresente estatísticas descritivas univariadas de todas as variáveis. Alguma requer recodificação adicional? Alguma variável requer transformação logarítmica?**

6) **Estimaremos a regressão múltipla para a aceitação do aborto, operacionalizando o modelo hipotético (i.e. o Processo Gerador de Dados que imaginamos ter operado). Lembre-se apenas as setas diretas na variável dependente entram na equação. Utilize a variável PESOFIM como penso amostral. E use a função `feols(.)` do pacote `fixest`.**

   a. Primeiramente, salve uma cópia do banco de dados, que contém apenas as as variáveis do modelo (dependente e independentes), além do peso amostral e a a variável F3, que utilizaremos logo abaixo. Depois, filtre esse data.frame, para manter apenas os casos completos.
   b. Estime 3 modelos: com erros homocedásticos, com erros heterocedásticos, com erros clusterizados (tomando F3, variável indicadora do setor censitário sorteado) como cluster.Apresente as estimativas lado a lado, usando o `modelsummary`. Ative a opção de mostrar com asterisco os coeficientes estatisticamente significativos. Os erros-padrão e as significâncias se alteram?
   c. Qual o tipo de erro-padrão adequado para essa situação?

7) **Interprete os coeficientes do modelo que você considerou adequado no exercício 10c. Indique se a direção dos efeitos vai na direção que inicialmente suspeitávamos.**

8) **Estime novamente o modelo. Mas operacionalizaremos religião de duas formas:**
   - **Da forma usual, já utilizada no exercício anterior**
   - **Mantendo apenas a variável *dummy* indicadora de evangélicos. Todos os demais grupos serão categoria de referência.**

   Assim, primeiramente crie variáveis *dummy* para cada religião. Isso não foi necessário antes pois R automaticamente transforma variáveis character em conjuntos de *dummies*.

   a. Estime a regressão com as *dummies* de religião (deixando católicos não praticantes como referência). Estime, em seguida, a regressão apenas com a *dummy* de evangélicos. Apresente-as lado a lado com o `modelsummary`.

   b. Lendo as estatísticas e medidas de ajuste, decida sobre qual modelo é melhor. Justifique sua resposta.

9) **Seria adequado usar idade ao quadrado ou ao cubo?**
   a. Construa um gráfico com a média da variável sobre aborto ao longo das idades. Adicione uma camada com a regressão não-paramétrica automaticamente gerada pelo ggplot (`geom_smooth`). Qual a sua opinião? Algum polinômio faria bem?
   b. Estime uma regressão (com a especificação que você decidiu que é a melhor no exercício 12.c), adicionando um termo quadrático para idade (além do linear)
   c. Agora adicione também o termo cúbico
   d. Lendo as estatísticas e medidas de ajuste, decida sobre qual modelo é melhor. Justifique sua resposta.