Atividade1_MichelleBouhid

MichelleBouhid

2024-09-18

Exercicio 1 e 2

Q1: Primeiramente, vocês devem acessar os dados da PNAD referente à cross-section do seu grupo. Isso pode ser feito diretamente pelo site, ou através do pacote `pnadcibge` no R. Em todos os casos, utilizem os dados referentes ao terceiro trimestre da amostra.

Q2: A PNAD possui uma quantidade enorme de dados, porém iremos utilizar apenas algumas variáveis tradicionais da literatura. Como o dataset guarda os dados através de seus códigos, vocês precisam consultar o dicionário da PNAD (disponível no site) e selecionar as colunas com os códigos corretos. As variáveis são:

```
- Rendimento mensal efetivo de todos os trabalhos para pessoas de 14 anos ou mais de idade;
- Horas habitualmente trabalhadas em todos os trabalhos para pessoas de 14 anos ou mais;
- Unidade da Federação;
- Rural/Urbano;
- Sexo;
- Idade do morador na data de referência;
- Cor ou raça;
- Curso mais elevado que frequentou anteriormente?;
- Anos de estudo;
- Condição em relação à força de trabalho na semana de referência para pessoas de 14 anos o
u mais de idade;
- Condição de ocupação na semana de referência para pessoas de 14 anos ou mais de idade;
- Setor do trabalho.
```

O primeiro passo do trabalho foi baixar os dados da PNAD referente ao terceiro trimestre do ano de 2019. Podemos fazer isso de duas formas: pelo site da PNAD ou utilizando o pacote pnadcibge no R. Nesta análise, escolhemos utilizar o pacote R para facilitar o acesso direto aos dados e a manipulação posterior.

Utilizamos a função get_pnadc() do pacote pnadcibge para baixar os dados. Após baixar os dados, selecionamos as variáveis de interesse para garantir que vamos trabalhar apenas com as colunas necessárias para a análise.

Fizemos a seleção das variáveis de interesse, como rendimento, horas trabalhadas, UF, situação de domicílio, sexo, idade, cor/raça, curso mais elevado, anos de estudo, força de trabalho, ocupação e setor de trabalho. Salvamos os dados em CSV para posterior acesso.

```r
# Código referente a resposta do Exercício 1 e 2

# Baixando os dados da PNAD Contínua para o ano de 2019, terceiro trimestre
# A função get_pnadc() é utilizada para baixar os dados com as variáveis de interesse.
df_pnad <- get_pnadc(year = 2019, quarter = 3,
                     vars = c(
                       "VD4020", # Rendimento mensal efetivo
                       "VD4031", # Horas habitualmente trabalhadas
                       "UF", # Unidade da federação
                       "V1022", # Situação do Domicílio (Rural/Urbano)
                       "V2007", # Sexo
                       "V2009", # Idade
                       "V2010", # Cor ou raça
                       "V3009A", # Curso mais elevado que frequentou
                       "VD3005", # Anos de estudo
                       "VD4001", # Condição em relação à Força de Trabalho
                       "VD4002", # Condição do ocupação
                       "VD4009" # Setor
                     ), design = FALSE) ## Definindo design como FALSE para não carregar o pacote de survey (opcional nesse caso)

# Selecionando apenas as colunas de interesse que foram baixadas para assegurar que as variáveis são mantidas
df_pnad <- df_pnad[,c(
  "VD4020", # Rendimento mensal efetivo
  "VD4031", # Horas habitualmente trabalhadas
  "UF", # Unidade da federação
  "V1022", # Situação do Domicílio (Rural/Urbano)
  "V2007", # Sexo
  "V2009", # Idade
  "V2010", # Cor ou raça
  "V3009A", # Curso mais elevado que frequentou
  "VD3005", # Anos de estudo
  "VD4001", # Condição em relação à Força de Trabalho
  "VD4002", # Condição do ocupação
  "VD4009" # Setor
)]

# Gerando uma tabela para examinar as categorias da variável 'V3009A' que representa o curso mais elevado que frequentou.
table(df_pnad$V3009A)
```

```
##
## Creche (disponível apenas no questionário anual de educação)
##                                                              0
##                                                      Pré-escola
##                                                            869
##                                   Classe de alfabetização - CA
##                                                           7000
##                               Alfabetização de jovens e adultos
##                                                           2066
##                                    Antigo primário (elementar)
##                                                          42805
##                                 Antigo ginásio (médio 1º ciclo)
##                                                           5852
##                   Regular do ensino fundamental ou do 1º grau
##                                                         105311
##   Educação de jovens e adultos (EJA) ou supletivo do 1º grau
##                                                           4121
##           Antigo científico, clássico, etc. (médio 2º ciclo)
##                                                           4272
##                         Regular do ensino médio ou do 2º grau
##                                                         120812
##   Educação de jovens e adultos (EJA) ou supletivo do 2º grau
##                                                           5935
##                                          Superior - graduação
##                                                          44243
##                               Especialização de nível superior
##                                                          12985
##                                                       Mestrado
##                                                           1833
##                                                      Doutorado
##                                                            739
```

```
# Salvando os dados em um arquivo CSV para garantir que podemos reutilizar os dados mais ta
rde
# Especificamos o caminho "Input/microdados_pnad.csv" onde o arquivo será salvo
write.csv(df_pnad, "Input/microdados_pnad.csv")

# Carregando os dados salvos em CSV novamente para verificar e manipular posteriormente
# df_pnad <- read.csv("Input/microdados_pnad.csv")
```

Exercicio 3

Q3: Baixadas as variáveis, tomem cuidado para realizar quaisquer limpezas que possam ser necessárias. Todas as informações sobre valores de variáveis estão disponíveis no dicionário. Nesta parte classificamos cada variável de acordo com o tipo que seria melhor para os calculos das proximos questões:

Conversão de tipos de variáveis:

VAL_RENDIMENTO: Foi convertida para numérico porque queremos calcular a média, valores de log, etc. Essa variável representa o rendimento mensal efetivo, o que faz sentido ser numérica.

NUM_HORAS: Foi convertida para inteiro porque são horas trabalhadas, que só fazem sentido em valores inteiros.

NOM_UF: Foi convertida para string (character) porque estamos tratando com nomes de estados, que são categóricos.

Criação de variáveis binárias:

IN_URBANA: Criada para representar se o indivíduo vive em área urbana (1 para urbano, 0 para rural). Vamos incluir da variável como dummy em modelos de regressão, sendo uma das variáveis explicativas.

IN_FEMININO: Foi criada para indicar o gênero, com 1 para feminino e 0 para masculino. Vamos usar análise de gênero como fator explicativo em regressões.

IN_PRETA_PARDA: Foi criada para agrupar pessoas de cor preta ou parda, de acordo com as categorias presentes na variável original. Essa variável é útil em análises de discriminação ou diferenças de rendimento por cor/raça.

Criação de variáveis categóricas:

IN_EDUC_1, IN_EDUC_2, IN_EDUC_3: Cada uma dessas variáveis categóricas representa um nível de escolaridade: básico, médio e superior. Criar variáveis separadas para cada nível de educação facilita a modelagem estatística e análise de diferenças entre esses grupos.

GRUPO_EDUC: Esta variável agrupa os três níveis de educação (IN_EDUC_1, IN_EDUC_2, IN_EDUC_3) em uma única variável categórica, permitindo uma análise mais simplificada ao invés de usar três variáveis separadas, além de possibilitar comparações entre grupos.

Conversão de variáveis:

NUM_ANOS_ESTUDO: Essa variável foi criada a partir de várias categorias de anos de estudo (VD3005). A variável representa os anos de estudo em formato numérico para facilitar a análise quantitativa.

Variáveis relacionadas à ocupação e força de trabalho:

IN_FORCA_TRABALHO: Foi criada para indicar se o indivíduo está ou não na força de trabalho. Variáveis binárias como esta são úteis para simplificar a inclusão dessa característica em modelos.

IN_OCUPADO: Outra variável binária para indicar se o indivíduo está ocupado ou não, importante para separar os que estão no mercado de trabalho dos que não estão.

COD_POSICAO_OCUPACAO: Esta variável categórica foi criada para detalhar a posição ocupacional (se tem carteira assinada, trabalhador doméstico, etc.). As categorias são baseadas em posições ocupacionais específicas. Isso é importante para análise de setores específicos de trabalho e suas influências no rendimento.

Filtragem:

Os dados foram filtrados para incluir apenas os residentes do estado do Amapá (NOM_UF==“Amapá”).

```
# Código referente a resposta do Exercício 3


# Questão 3:

# Baixadas as variáveis, tomem cuidado para realizar quaisquer limpezas que possam ser nece
ssárias.
# Todas as informações sobre valores de variáveis estão disponíveis no dicionário.

# Aplicando as transformações para limpar e preparar os dados com o pacote dplyr:
df_pnad <- df_pnad %>% mutate(

  # Convertendo a variável de rendimento para numérico
  VAL_RENDIMENTO = as.numeric(VD4020),

  # Convertendo a variável de horas trabalhadas para inteiro
  NUM_HORAS = as.integer(VD4031),

  # Convertendo a variável de Unidade Federativa para string
  NOM_UF = as.character(UF),

  # Criando uma variável binária indicando se o indivíduo vive em área urbana
  IN_URBANA = ifelse(V1022=="Urbana", 1, 0),

  # Criando uma variável binária indicando o sexo (1 para feminino, 0 para masculino)
  IN_FEMININO = ifelse(V2007=="Mulher", 1, 0),

  # Convertendo a variável de idade para inteiro
  NUM_IDADE = as.integer(V2009),

  # Criando uma variável binária para identificar pessoas de cor preta ou parda
  IN_PRETA_PARDA = ifelse(V2010=="Preta"|V2010=="Parda",1,0), # & para interseção "e" e | p
ara união "ou"


  # Criando variáveis binárias para nível de educação:
  # IN_EDUC_1: Indivíduos com nível educacional básico (creche até supletivo de 1º grau)
  IN_EDUC_1 = ifelse(V3009A == "Creche (disponível apenas no questionário anual de educaçã
o)"|
                       V3009A == "Pré-escola"|
                       V3009A == "Classe de alfabetização - CA"|
                       V3009A == "Alfabetização de jovens e adultos"|
                       V3009A == "Antigo primário (elementar)"|
                       V3009A == "Antigo ginásio (médio 1º ciclo)"|
                       V3009A == "Regular do ensino fundamental ou do 1º grau"|
                       V3009A == "Educação de jovens e adultos (EJA) ou supletivo do 1º gra
u",1,0),

  # IN_EDUC_2: Indivíduos com nível educacional médio (científico até EJA de 2º grau)
  IN_EDUC_2 = ifelse(V3009A == "Antigo científico, clássico, etc. (médio 2º ciclo)"|
                       V3009A == "Regular do ensino médio ou do 2º grau"|
                       V3009A == "Educação de jovens e adultos (EJA) ou supletivo do 2º gra
u",1,0),
```

```
  # IN_EDUC_3: Indivíduos com nível educacional superior (graduação até doutorado)
  IN_EDUC_3 = ifelse(V3009A == "Superior - graduação"|
                       V3009A == "Especialização de nível superior"|
                       V3009A == "Mestrado"|
                       V3009A == "Doutorado",1,0),

  # Agrupando os três níveis de educação em uma única variável categórica
  GRUPO_EDUC = 1*IN_EDUC_1+2*IN_EDUC_2+3*IN_EDUC_3,

  # Criando uma variável de anos de estudo
  NUM_ANOS_ESTUDO = as.integer(case_when(
    VD3005 == "Sem instrução e menos de 1 ano de estudo" ~ "00",
    VD3005 == "1 ano de estudo" ~ "01",
    VD3005 == "2 anos de estudo" ~ "02",
    VD3005 == "3 anos de estudo" ~ "03",
    VD3005 == "4 anos de estudo" ~ "04",
    VD3005 == "5 anos de estudo" ~ "05",
    VD3005 == "6 anos de estudo" ~ "06",
    VD3005 == "7 anos de estudo" ~ "07",
    VD3005 == "8 anos de estudo" ~ "08",
    VD3005 == "9 anos de estudo" ~ "09",
    VD3005 == "10 anos de estudo" ~ "10",
    VD3005 == "11 anos de estudo" ~ "11",
    VD3005 == "12 anos de estudo" ~ "12",
    VD3005 == "13 anos de estudo" ~ "13",
    VD3005 == "14 anos de estudo" ~ "14",
    VD3005 == "15 anos de estudo" ~ "15",
    VD3005 == "16 anos ou mais de estudo" ~ "16"
  )),

  # Criando uma variável binária para a força de trabalho (1 se está na força de trabalho)
  IN_FORCA_TRABALHO = ifelse(VD4001=="Pessoas na força de trabalho",1,0),

  # Criando uma variável binária para ocupação (1 se está ocupado)
  IN_OCUPADO = ifelse(VD4002=="Pessoas ocupadas",1,0),

  # Criando uma variável categórica para a posição ocupacional
  COD_POSICAO_OCUPACAO = case_when(
    VD4009 == "Empregado no setor privado com carteira de trabalho assinada" ~ "01",
    VD4009 == "Empregado no setor privado sem carteira de trabalho assinada" ~ "02",
    VD4009 == "Trabalhador doméstico com carteira de trabalho assinada" ~ "03",
    VD4009 == "Trabalhador doméstico sem carteira de trabalho assinada" ~ "04",
    VD4009 == "Empregado no setor público com carteira de trabalho assinada" ~ "05",
    VD4009 == "Empregado no setor público sem carteira de trabalho assinada" ~ "06",
    VD4009 == "Militar e servidor estatutário" ~ "07",
    VD4009 == "Empregador" ~ "08",
    VD4009 == "Conta-própria" ~ "09",
    VD4009 == "Trabalhador familiar auxiliar" ~ "10")
) %>%

  # Selecionando apenas as variáveis de interesse
  select(VAL_RENDIMENTO,
         NUM_HORAS,
         NOM_UF,
         IN_URBANA,
```

```
        IN_FEMININO,
        NUM_IDADE,
        IN_PRETA_PARDA,
        IN_EDUC_1, #incluindo as classes novas
        IN_EDUC_2,
        IN_EDUC_3,
        GRUPO_EDUC,
        NUM_ANOS_ESTUDO,
        IN_FORCA_TRABALHO,
        IN_OCUPADO,
        COD_POSICAO_OCUPACAO) %>%

# Filtrando apenas os dados do Amapá
filter(NOM_UF=="Amapá")
```

Exercicio 4

Q4: Filtre a base de dados, mantendo apenas a população entre 20 e 60 anos. Calcule a proporção de pessoas economicamente ativas, condicional ao nível de escolaridade, sexo, e idade, separadamente. Depois, façam o mesmo exercício apenas para a região de vocês (Amapá). Como a probabilidade de estar no mercado de trabalho varia em função dessas características? Analisem também as possíveis causas de diferenças entre o resultado nacional e o resultado específico do grupo.

Proporção geral de trabalhadores economicamente ativos:

A tabela mostra que cerca de 76,19% das pessoas entre 20 e 60 anos estão na força de trabalho. Este valor está dentro do esperado, considerando que esta faixa etária inclui a maioria das pessoas em idade economicamente ativa.

Proporção condicional ao nível de escolaridade:

O grupo com educação superior (GRUPO_EDUC 3) apresenta uma maior fração de participação no mercado de trabalho, com 90% dos indivíduos. Isso é coerente com a teoria econômica, que sugere que indivíduos com maior nível educacional têm mais oportunidades de emprego e maior propensão a participar do mercado de trabalho. Indivíduos com nível básico de educação (GRUPO_EDUC 1) têm uma proporção de participação menor (71%). Isso pode ser explicado pelas dificuldades que esses indivíduos enfrentam no mercado de trabalho, muitas vezes em empregos informais ou de baixa qualificação.

Proporção condicional ao sexo:

Homens têm uma participação significativamente maior no mercado de trabalho (87,83%) em comparação às mulheres (65,37%). Este resultado reflete uma tendência ainda presente em muitas regiões do Brasil, onde as mulheres enfrentam barreiras maiores para entrar e se manter no mercado de trabalho, como dupla jornada (trabalho doméstico e trabalho remunerado) e discriminação de gênero.

Proporção condicional à idade:

A participação no mercado de trabalho aumenta gradualmente a partir dos 20 anos, atinge um pico em torno dos 35-40 anos e começa a diminuir novamente após os 50 anos. Isso reflete o ciclo típico da vida laboral, onde jovens ainda estão ingressando no mercado e pessoas mais velhas começam a se retirar ou se preparar para a aposentadoria.

Proporção condicional à idade e sexo:

Para todas as faixas etárias, os homens têm uma maior participação no mercado de trabalho do que as mulheres. A diferença é maior nas idades mais jovens e diminui à medida que a idade aumenta. Essa diferença entre homens e mulheres pode estar relacionada a fatores sociais, como a maior propensão de mulheres jovens a cuidar de filhos pequenos ou realizar tarefas domésticas, enquanto homens mais jovens tendem a entrar mais cedo no mercado de trabalho.

Conclusão:

Esses resultados são coerentes com a literatura econômica e demográfica, que mostra como fatores como educação, gênero e idade influenciam as chances de uma pessoa estar no mercado de trabalho. Indivíduos mais qualificados têm mais oportunidades, homens tendem a ter maior participação, e a idade é um fator importante na trajetória laboral, com picos de participação no meio da vida adulta.

```
# Código referente a resposta do Exercício 4

# Monitoria 03/09

# Questão 4:

# (a) Filtre a base de dados, mantendo apenas a população entre 20 e 60 anos.
# (b) Calcule a proporção de pessoas economicamente ativas, condicional ao nível de escolar
idade, sexo, e idade, separadamente.
# (c) Depois, façam o mesmo exercício apenas para a região de vocês (Amapá). DISPENSADO
# (d) Como a probabilidade de estar no mercado de trabalho varia em função dessas caracterí
sticas?
# (e) Analisem também as possíveis causas de diferenças entre o resultado nacional e o resu
ltado específico do grupo.DISPENSADO

# (a) Filtrando os dados para manter apenas pessoas entre 20 e 60 anos de idade
df_pnad <- df_pnad %>% filter(NUM_IDADE>=20 & NUM_IDADE<=60)
# Outra opção de filtro usando todas as colunas:
# df_pnad <- df_pnad[df_pnad$NUM_IDADE >= 20 & df_pnad$NUM_IDADE<=60,]



# (b) Proporção geral de trabalhadores economicamente ativos (na força de trabalho)
# Calcula a proporção de pessoas na força de trabalho (IN_FORCA_TRABALHO == 1) no total
nrow(df_pnad[df_pnad$IN_FORCA_TRABALHO==1,])/nrow(df_pnad)
```

```
## [1] 0.7619709
```

```r
# (b) Proporção condicional ao nível de escolaridade (agrupando por nível de educação)
# Agrupa pelo nível de educação e calcula a média de pessoas na força de trabalho
tab_educ <- df_pnad %>% group_by(GRUPO_EDUC) %>% summarise(
  FRAC_EDUC = mean(IN_FORCA_TRABALHO, na.rm = T) #excluindo NA
)

# (b) Proporção condicional ao sexo (agrupando pela variável IN_FEMININO)
# Agrupa por sexo (IN_FEMININO, onde 0 = masculino e 1 = feminino) e calcula a proporção na força de trabalho
tab_sexo <- df_pnad %>% group_by(IN_FEMININO) %>% summarise(
  FRAC_SEXO = mean(IN_FORCA_TRABALHO, na.rm = T)
)

# (b) Proporção condicional à idade (entre 20 e 60 anos)
# Agrupa por idade e calcula a proporção de pessoas na força de trabalho por idade
tab_idade <- df_pnad %>% group_by(NUM_IDADE) %>% summarise(
  FRAC_IDADE = mean(IN_FORCA_TRABALHO, na.rm = T)
)

# OBS: (d) Relatório: Agrupando por idade e sexo dentro da força de trabalho
# Usamos o pacote tidyverse (que inclui dplyr) para fazer o agrupamento por idade e sexo,
# e calcular a proporção de cada grupo na força de trabalho.


# Agrupando por idade e sexo (IN_FEMININO), pivotando os dados para ter as colunas de fraçõ
es
tab_sexo_idade <- df_pnad %>% group_by(NUM_IDADE, IN_FEMININO) %>% summarise(
  FRAC_SEXO_IDADE = mean(IN_FORCA_TRABALHO, na.rm = T)
) %>% pivot_wider(id_cols = "NUM_IDADE", #coluna que nao muda, coluna de identificação
                  names_from = "IN_FEMININO", #coluna pivotada
                  values_from = "FRAC_SEXO_IDADE",#identifica valores imputados nas obs
                  names_prefix = "SEXO_") # sexo_ 0 ou 1
```

```
## `summarise()` has grouped output by 'NUM_IDADE'. You can override using the
## `.groups` argument.
```

```r
#usamos o pipe nesse formato %>%

# Renomeando as colunas da tabela pivotada
colnames(tab_sexo_idade) <- c("Idade", "Fração Masculina", "Fração Feminina")
```

Exercicio 5

Q5: Após gerar a base tratada, filtre os dados para manter apenas trabalhadores do setor privado, empregados na data da amostra. Vocês devem reportar uma tabela com estatísticas descritivas (mínimo, máximo, média, desvio-padrão e quantis) das variáveis contínuas, e as proporções de cada grupo para as variáveis de nível (trate idade como contínua). Para realizar nossa avaliação, é importante que nenhuma variável tenha todas as observações iguais. Explique qual seria o problema caso contrário

Após filtrar os dados do setor Privado chegamos as seguintes estatísticas nas **variaveis numéricas:**

Rendimento (VAL_RENDIMENTO):

O rendimento varia de um mínimo de 998 até um máximo de 15.000, com a média sendo 1.435. A mediana está em 1.100, mostrando que a maioria dos rendimentos é concentrada em torno desse valor, indicando uma distribuição assimétrica, com alguns valores extremos elevados.

Horas Trabalhadas (NUM_HORAS):

A maioria das pessoas trabalha 40 horas semanais (mediana e média), o que é consistente com o padrão de jornada semanal de trabalho no Brasil. O mínimo é 8 horas, enquanto o máximo é 70 horas, indicando variações significativas nas jornadas de trabalho, possivelmente devido a diferentes tipos de emprego ou contratos de trabalho.

Idade (NUM_IDADE):

A idade das pessoas varia de 20 a 60 anos, como esperado devido ao filtro aplicado anteriormente. A média de idade é de aproximadamente 34 anos, sugerindo que a amostra inclui majoritariamente pessoas em fase intermediária de suas carreiras profissionais.

Anos de Estudo (NUM_ANOS_ESTUDO):

A média de anos de estudo é de 11,76, o que equivale a aproximadamente ensino médio completo. Os valores variam de 11 a 16 anos de estudo, indicando que a maioria das pessoas na amostra completou pelo menos o ensino médio, com algumas tendo ensino superior.

Frequências das Variáveis Categóricas:

**Área Urbana (IN_URBANA):

Cerca de 530 pessoas vivem em áreas urbanas, enquanto 29 vivem em áreas rurais. Isso reflete a concentração populacional em áreas urbanas, o que é comum no Brasil.

Sexo (IN_FEMININO):

A amostra em termos de gênero, observa-se uma distribuição desbalanceada, com mais homens do que mulheres, contendo 205 pessoas do sexo feminino e 354 do sexo masculino.

Cor/Raça (IN_PRETA_PARDA):

Aproximadamente 462 pessoas se identificam como pretas ou pardas e 97 outros na amostra. Essa é uma característica demográfica importante, dada a desigualdade racial observada no mercado de trabalho.

Nível Educacional (GRUPO_EDUC):

A maioria das pessoas (261) tem nível educacional médio (2), seguido por pessoas com nível básico (98) e, por último, aquelas com nível superior (130). A predominância de trabalhadores com ensino médio e superior é compatível com as exigências educacionais do setor privado, que geralmente requer um nível mais alto de qualificação.

Posição Ocupacional (COD_POSICAO_OCUPACAO):

A maior parte das pessoas (382) está empregada com carteira assinada (grupo 01), enquanto um número significativo (177 pessoas) está empregada sem carteira.

Conclusões:

Estatísticas Contínuas: As variáveis contínuas, como rendimentos, horas trabalhadas, anos de estudo e idade, mostram uma variação significativa entre os trabalhadores do setor privado. A distribuição de renda, em particular, revela uma disparidade considerável, com valores extremos mais elevados aumentando a

média em comparação com a mediana. Isso sugere desigualdades dentro do setor privado em termos de remuneração. A variação nas horas trabalhadas também é notável, com algumas pessoas trabalhando muito mais do que outras, refletindo uma diversidade de jornadas de trabalho.

Estatísticas Categóricas: As frequências das variáveis categóricas indicam um mercado de trabalho predominantemente urbano e masculino, com uma boa quantidade de pessoas pretas ou pardas e uma distribuição equilibrada de nível educacional. A maior parte da população na amostra tem educação de nível médio ou superior, o que está alinhado com as expectativas do setor privado.

** Problema de Observações Iguais:** Se isso ocorrer, essa variável não contribui para a análise, pois não oferece variabilidade suficiente para ser estatisticamente relevante. Ela seria redundante e não explicaria a variabilidade do resultado (como rendimento), o que poderia distorcer as conclusões de um modelo estatístico ou econométrico.

```
# Código referente a resposta do Exercício 5

# Questão 5:
# (a) Filtre os dados para manter apenas trabalhadores do setor privado, empregados na data da amostra.
# (b) Vocês devem reportar uma tabela com estatísticas descritivas (mínimo, máximo, média, desvio-padrão e quantis)
# (c) Proporções de cada grupo para as variáveis de nível (trate idade como contínua). Falt
a
# (d) Para realizar nossa avaliação, é importante que nenhuma variável tenha todas as obser
vações iguais.
#Explique qual seria o problema caso contrário - Falta

# (a)

# (a) Filtrando os trabalhadores do setor privado com carteira assinada ou sem carteira ass
inada
# (Varivel SETOR VD4009) - 1 carteira assinada / 2 sem carteira assinada
df_pnad_setor_privado <- df_pnad %>% filter(
  COD_POSICAO_OCUPACAO %in% c("01","02")
  # COD_POSICAO_OCUPACAO=="01"|COD_POSICAO_OCUPACAO=="02" (2ºmaneira)
)

# (b) Estatísticas descritivas das variáveis contínuas (salário, horas trabalhadas, idade, anos de estudo)
# Utilizando a função summary para gerar mínimo, 1º quartil, mediana, média, 3º quartil, má
ximo e valores ausentes (NA)

# Estatísticas descritivas do valor de rendimento (salário)
# fç summary, das variáveis contínuas
summary(df_pnad_setor_privado$VAL_RENDIMENTO) #Fotografia do resultado abaixo
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##       0     998    1100    1435    1600   15000
```

```
#  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
# 0     998    1100    1435    1600   15000       6

# Estatísticas descritivas do número de horas trabalhadas por semana
summary(df_pnad_setor_privado$NUM_HORAS)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    8.00   40.00   40.00   39.76   44.00   70.00
```

```
#  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
# 8.00   40.00   40.00   39.76   44.00   70.00

# Estatísticas descritivas da idade dos indivíduos
summary(df_pnad_setor_privado$NUM_IDADE)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   20.00   26.00   32.00   34.14   41.50   60.00
```

```
# Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
# 20.00   26.00   32.00   34.14   41.50   60.00

# Estatísticas descritivas dos anos de estudo
summary(df_pnad_setor_privado$NUM_ANOS_ESTUDO)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00   11.00   12.00   11.76   14.00   16.00
```

```
# Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
# 0.00   11.00   12.00   11.76   14.00   16.00

# (c) Frequência das variáveis categóricas
# Utilizando a função table para contar as ocorrências de cada nível das variáveis categóri
cas

# Contagem da variável "IN_URBANA" (se o domicílio é urbano ou rural)
# Exemplo de saída esperada: Frequência para 0 (rural) e 1 (urbano)
table(df_pnad_setor_privado$IN_URBANA)
```

```
##
##   0   1
##  29 530
```

```
# 0   1
# 29 530

# Contagem da variável "IN_FEMININO" (gênero, sendo 0 = masculino e 1 = feminino)
# Exemplo de saída esperada: Frequência para 0 (homem) e 1 (mulher)
table(df_pnad_setor_privado$IN_FEMININO)
```

```
##
##   0   1
## 354 205
```

```
#  0   1
# 354 205

# Contagem da variável "IN_PRETA_PARDA" (se a pessoa se declara preta ou parda)
# Exemplo de saída esperada: Frequência para 0 (não preta/parda) e 1 (preta/parda)
table(df_pnad_setor_privado$IN_PRETA_PARDA)
```

```
##
##   0   1
##  97 462
```

```
# 0   1
# 97 462

# Contagem da variável "GRUPO_EDUC" (nível de educação: 1 = baixa, 2 = média, 3 = alta)
# Exemplo de saída esperada: Frequência para 1 (baixa), 2 (média), 3 (alta)
table(df_pnad_setor_privado$GRUPO_EDUC)
```

```
##
##   1   2   3
##  98 261 130
```

```
#   1    2    3
#  98 261 130

# Contagem da variável "COD_POSICAO_OCUPACAO" (posição de ocupação: 01 = carteira assinada,
02 = sem carteira assinada)
# Exemplo de saída esperada: Frequência para 01 (com carteira) e 02 (sem carteira)
table(df_pnad_setor_privado$COD_POSICAO_OCUPACAO)
```

```
##
##  01  02
## 382 177
```

```
# 01   02
# 382 177

# (d) Explicação sobre o problema de variáveis com observações iguais:
# Se uma variável tem todas as observações iguais, ela não contribui para o modelo de regre
ssão.
# Isso ocorre porque uma variável constante não tem variabilidade e, portanto, não pode exp
licar a
# variação na variável dependente (rendimento, por exemplo).
```

Exercicio 6

Q6: Calcule o salário de cada indivíduo e gere os histogramas dos salários condicionais ao seu nível mais elevado de educação. Faça o mesmo com o logaritmo dos salários. Qual conjunto de histogramas mais se assemelha com uma distribuição normal? Comente a relação desses histogramas com o que sabemos sobre a distribuição condicional de y dadas todas as hipóteses de MQO.

Objetivo da questão foi a análise dos salários condicionados ao nível educacional, tanto com os salários absolutos quanto com os logaritmos dos salários. Após excluir valores Ausentes e tratar outliers, as conclusões foram:

Salários condicionados ao nível educacional (gráfico 1):

O gráfico de rendimentos absolutos mostra uma grande concentração de salários abaixo de 1000 reais, especialmente para os grupos de educação baixa e média. Isso sugere que a maior parte dos trabalhadores com menor escolaridade ganha salários mais baixos. Para o grupo de educação superior, há uma distribuição mais dispersa, com alguns salários mais altos, mas ainda concentrados em níveis intermediários.

A distribuição não segue um padrão de normalidade e apresenta uma assimetria significativa, principalmente nos extremos.

Logaritmo dos salários condicionados ao nível educacional (gráfico 2):

Ao calcular o logaritmo dos salários, a distribuição se torna mais próxima de uma distribuição normal, especialmente para os grupos de educação média e alta. Esse comportamento é esperado, já que o logaritmo de variáveis com assimetria positiva (como a renda) tende a produzir distribuições mais simétricas.

Para os indivíduos com menor nível educacional, a distribuição do logaritmo dos salários ainda apresenta certa concentração em níveis mais baixos, o que indica desigualdade salarial significativa entre os diferentes grupos educacionais. Comparação com MQO:

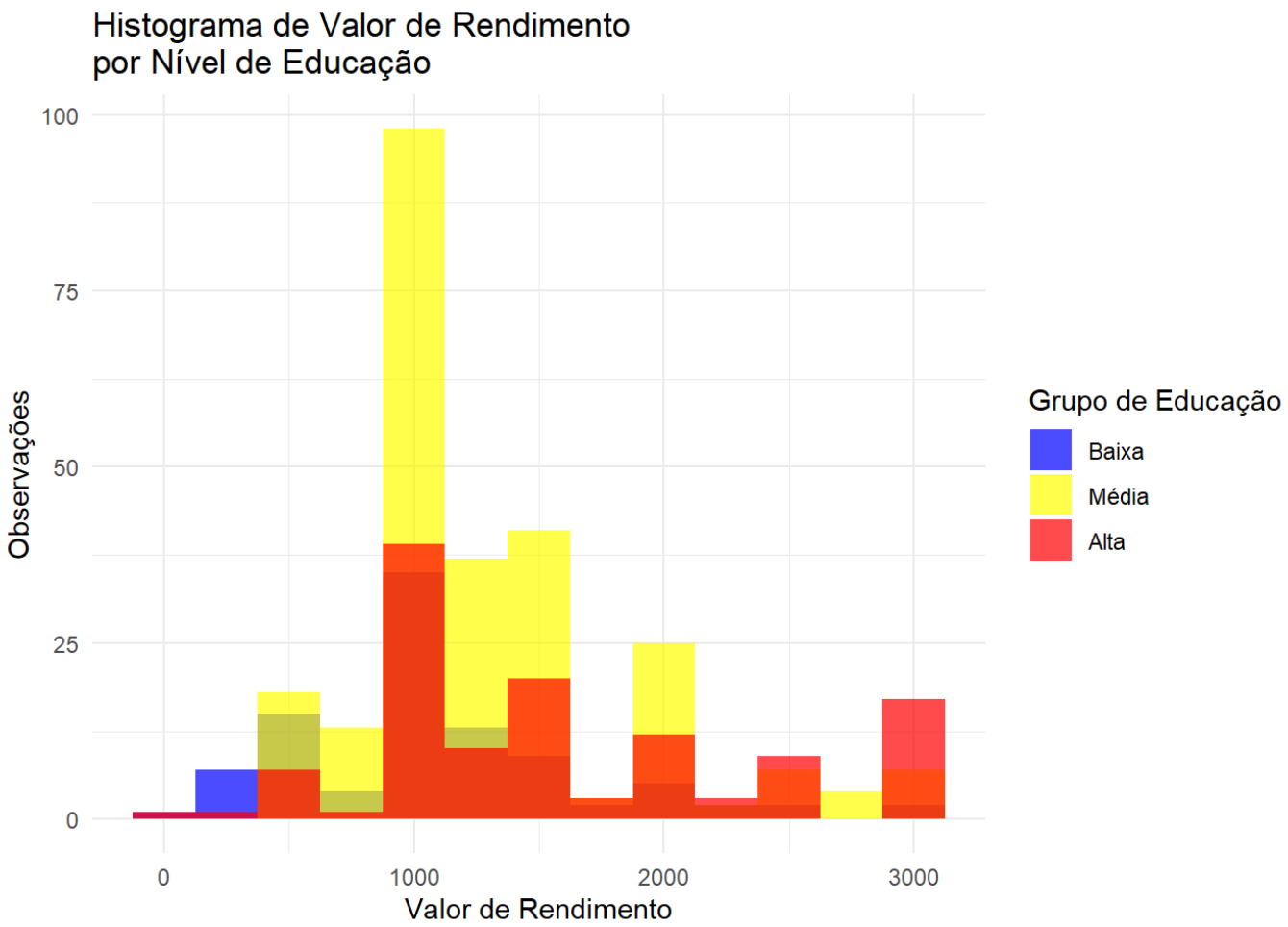
Uma das hipóteses do modelo de MQO é que os erros seguem uma distribuição normal. Ao observar os histogramas do logaritmo dos salários, vemos que eles se aproximam mais de uma distribuição normal, indicando que o uso do logaritmo dos salários pode ser mais adequado para atender às premissas de normalidade dos erros.

Portanto, ao modelar o rendimento com MQO, é mais apropriado usar o logaritmo dos salários para melhorar a adequação do modelo às hipóteses teóricas.

```r
# Código referente a resposta do Exercício 6

# Monitoria 04/09

# Questão 6:
# (a) Calcule o salário de cada indivíduo e gere os histogramas dos salários condicionais a
o seu nível mais elevado de educação.
# (b) Faça o mesmo com o logaritmo dos salários. Qual conjunto de histogramas mais se assem
elha com uma distribuição normal?
# (c) Comente a relação desses histogramas com o que sabemos sobre a distribuição condicion
al de y dadas todas as hipóteses de MQO.
# A distribuição é aproximadamente Normal portanto podemos usar MQO.
# (a)

# Removendo observações com valores NA para o nível de educação
df_plot <- df_pnad_setor_privado[!is.na(df_pnad_setor_privado$GRUPO_EDUC),]

# Calculando o quantil de 95% do salário para eliminar outliers
quantile(df_plot$VAL_RENDIMENTO, 0.95, na.rm = T) # 3000
```

```
##  95%
## 3000
```

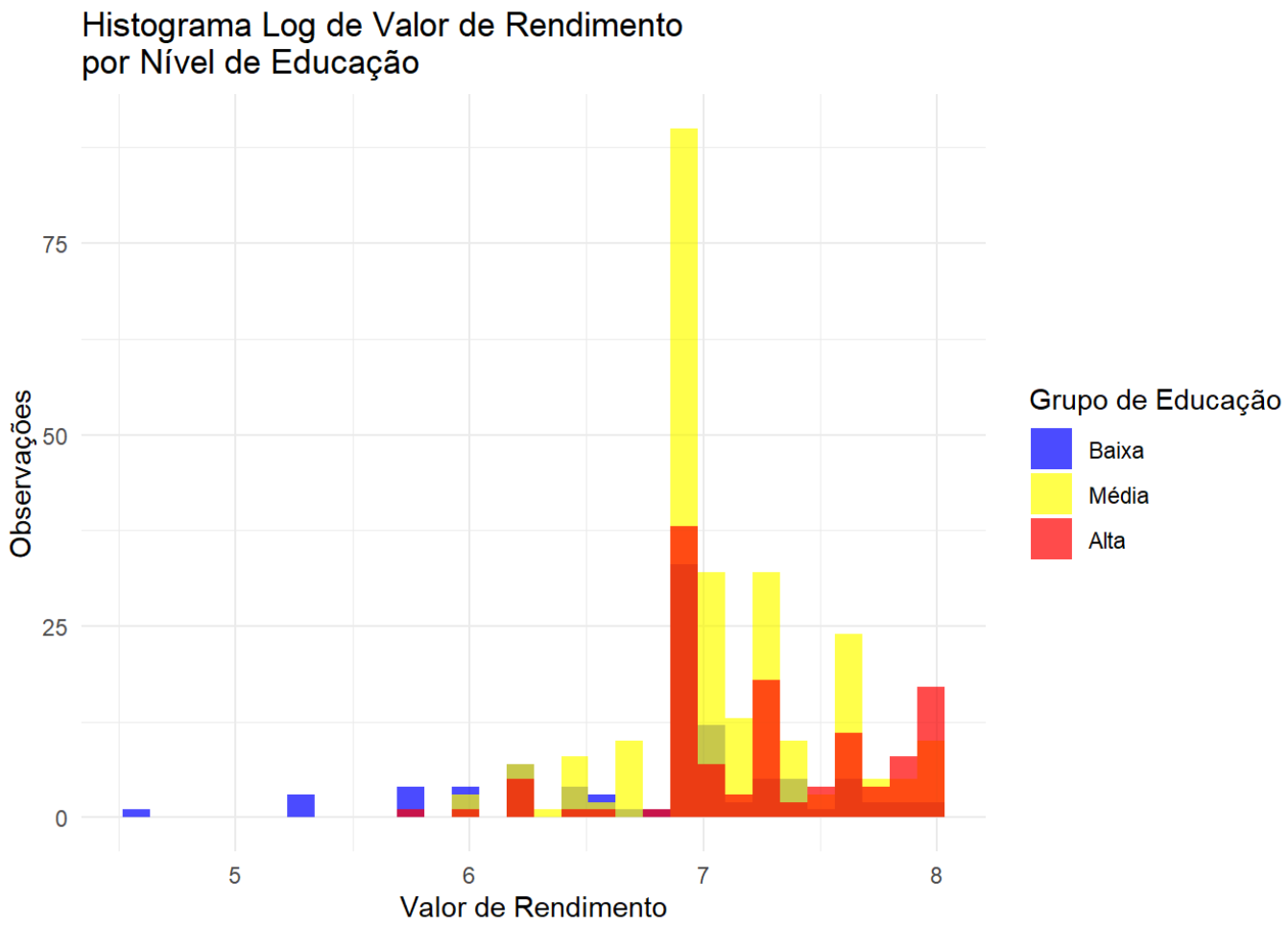
```r
# Filtrando os dados para manter apenas indivíduos com salários menores ou iguais ao 95º pe
rcentil
df_plot <- df_plot[df_plot$VAL_RENDIMENTO<=quantile(df_plot$VAL_RENDIMENTO, 0.95, na.rm =
T),]

# Gerando o histograma dos salários condicionados ao nível mais elevado de educação
hist_plot <- ggplot(data = df_plot,
                    aes(x = VAL_RENDIMENTO,
                        fill = as.factor(GRUPO_EDUC))) +
  geom_histogram(alpha = 0.7, binwidth = 250, position=position_identity()) +
  labs(x = "Valor de Rendimento",
       y = "Observações",
       title = "Histograma de Valor de Rendimento \npor Nível de Educação") +
  scale_fill_manual(
    name = "Grupo de Educação",
    values = c("1" = "blue", "2" = "yellow", "3" = "red"),
    breaks = c("1","2","3"),
    labels = c("Baixa", "Média", "Alta")
  ) + theme_minimal()

# Exibindo o histograma
hist_plot
```




```
# (b)

# Gerando o histograma do logaritmo dos salários condicionados ao nível mais elevado de edu
cação
hist_plot_log <- ggplot(data = df_plot,
                        aes(x = log(VAL_RENDIMENTO),
                            fill = as.factor(GRUPO_EDUC))) +
  geom_histogram(alpha = 0.7, position=position_identity()) +
  labs(x = "Valor de Rendimento",
       y = "Observações",
       title = "Histograma Log de Valor de Rendimento \npor Nível de Educação") +
  scale_fill_manual(
    name = "Grupo de Educação",
    values = c("1" = "blue", "2" = "yellow", "3" = "red"),
    breaks = c("1","2","3"),
    labels = c("Baixa", "Média", "Alta")
  ) + theme_minimal()

# Exibindo o histograma
hist_plot_log
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

```
## Warning: Removed 1 row containing non-finite outside the scale range
## (`stat_bin()`).
```




```
# y = alpha + beta X + epsilon


# Interpretação dos coeficientes:
# d VAL_RENDIMENTO/ d NUM_ANOS_ESTUDO (Homens) = beta_1
# d VAL_RENDIMENTO/ d NUM_ANOS_ESTUDO (Mulheres) = beta_1 + beta_3
# Gender Gap Educacional = d VAL_RENDIMENTO/ d NUM_ANOS_ESTUDO (Homens) -  d VAL_RENDIMENT
O/ d NUM_ANOS_ESTUDO (Mulheres) = - beta_3
```

Exercicio 7

Q7: Faça uma regressão simples do logaritmo dos salários sobre anos de escolaridade e reporte o resultado, comentando os testes de hipótese usuais. Se estiver usando o R, o pacote stargazer pode ser útil para a geração de tabelas.

Na regressão 1 (linear), os coeficientes representam variações absolutas em reais no rendimento.

Na regressão 2 (logarítmica), os coeficientes representam variações percentuais no rendimento. Portanto, enquanto a regressão linear indica o impacto em termos absolutos de reais, a regressão logarítmica indica o impacto em termos percentuais.

Regressão1

Fórmula do Modelo: Fórmula: lm(VAL_RENDIMENTO ~ 1 + NUM_ANOS_ESTUDO + IN_FEMININO + NUM_ANOS_ESTUDO * IN_FEMININO)

VAL_RENDIMENTO: O rendimento é a variável dependente.

NUM_ANOS_ESTUDO: Anos de estudo da pessoa variavel independente

IN_FEMININO: Uma variável dummy, onde 1 representa mulher e 0 representa homem.

Interação (NUM_ANOS_ESTUDO * IN_FEMININO): O efeito combinado entre anos de estudo e sexo, para ver como a escolaridade afeta homens e mulheres de forma diferente.

Resíduos:

Min: -1829.3 ; 1Q: -449.5 (primeiro quartil) ; Median: -225.5 ; 3Q: 204.7 (terceiro quartil) ; Max: 13170.7 ;

Esses valores representam a distribuição dos resíduos (erros) do modelo, ou seja, a diferença entre os valores observados e os preditos. Idealmente, a mediana dos resíduos deveria ser próxima de zero, o que sugere um bom ajuste do modelo. O fato de a mediana dos resíduos ser negativa (-225.5) com os resíduos estão centrados um pouco abaixo de zero, sugerindo que o modelo pode estar subestimando ligeiramente os rendimentos médios. Isso pode significar que o modelo tem um viés, o que indica que há fatores não capturados pelas variáveis incluídas (anos de estudo, gênero e sua interação).

Coeficientes:

Aqui estão os coeficientes estimados para cada variável, o erro-padrão, o valor t e o p-valor para teste de significância.

Intercepto (906.34):

Este é o valor médio estimado de rendimento para homens com zero anos de escolaridade (baseline). Significa que o salário de um homem sem escolaridade é estimado em R$ 906,34.O p-valor extremamente baixo indica que o intercepto é estatisticamente significativo, ou seja, é muito improvável que o valor verdadeiro seja zero. **Erro padrão (180.71):** A incerteza na estimativa do intercepto. **t valor (5.015):** Estatística t para testar a hipótese nula de que o coeficiente é igual a zero. Neste caso, o valor t é alto, sugerindo que o intercepto é significativamente diferente de zero. **p-valor (7.13e-07):** Um p-valor muito baixo (<0.001), que rejeita a hipótese nula, sugerindo que o intercepto é altamente significativo.

NUM_ANOS_ESTUDO (44.94):

Este coeficiente significa que, para cada ano adicional de estudo, o salário de um homem (baseline) aumenta em média R$ 44,94.

Erro padrão (15.44): A incerteza associada a essa estimativa. **t valor (2.910):** O valor t para testar se o coeficiente é significativamente diferente de zero. O valor t é maior que 2, o que sugere significância. **p-valor (0.00376):** Com um p-valor abaixo de 0,01, rejeitamos a hipótese nula e concluímos que o coeficiente de anos de estudo é significativo.

IN_FEMININO (-834.39):

Este coeficiente significa que, em média, as mulheres ganham R$ 834,39 a menos do que os homens, controlando pelos anos de escolaridade.

Erro padrão (384.05): A incerteza associada a essa estimativa. **t valor (-2.173):** O valor t é negativo, pois o coeficiente é negativo, indicando que o salário das mulheres é significativamente menor que o dos homens. **p-valor (0.03023):** Com um p-valor menor que 0,05, rejeitamos a hipótese nula e concluímos que há uma diferença estatisticamente significativa entre os rendimentos de homens e mulheres.

####** Interação (NUM_ANOS_ESTUDO * IN_FEMININO) (64.89):** Este coeficiente indica que o impacto dos anos de estudo nas mulheres é maior em R$ 64,89 em comparação aos homens.

Erro padrão (30.00): A incerteza na estimativa da interação. **t valor (2.163):** O valor t mostra que a interação é estatisticamente significativa. **p-valor (0.03097):** Com um p-valor menor que 0,05, a interação entre gênero e escolaridade é significativa, sugerindo que os anos de estudo têm um efeito maior nas mulheres.

Significância dos Coeficientes: Os asteriscos indicam o nível de significância: ***: p < 0,001 (muito significativo) ;*** *: p < 0,01 (significativo) ;* : p < 0,05 (moderadamente significativo)

Estatisticas do Modelo1:

Erro padrão residual: Residual standard error: 1051 Este valor representa o erro-padrão dos resíduos, ou seja, a variabilidade dos erros do modelo. Quanto menor, melhor o ajuste.

Multiple R-squared-R² (0.04709): Aproximadamente 4,71% da variabilidade no rendimento é explicada pelas variáveis incluídas no modelo.

Adjusted R-squared (0.04194): O R² ajustado penaliza pela inclusão de variáveis não explicativas, sendo levemente inferior ao R² simples. Ainda assim, o modelo tem um baixo poder explicativo, o que sugere que outras variáveis além da escolaridade e do gênero explicam o salário.

Estatística F: F-statistic (9.143): Testa se o modelo como um todo é significativo, comparado a um modelo sem preditores.

p-value (6.498e-06): Um p-valor extremamente pequeno indica que o modelo geral é altamente significativo, ou seja, pelo menos uma das variáveis independentes tem um efeito significativo sobre o salário.

Regressão 2

Formula Usada: log(1+VAL_RENDIMENTO) ~ NUM_ANOS_ESTUDO

Neste modelo, o logaritmo do rendimento (com adição de 1 para evitar log de zero) é a variável dependente. NUM_ANOS_ESTUDO: Os anos de estudo são a variável independente.

Resíduos:

Min: -7.2318 1Q (Primeiro Quartil): -0.1933 Mediana: -0.0092 (próximo de zero, o que é bom e indica que o modelo pode estar bem ajustado nesse sentido). 3Q (Terceiro Quartil): 0.2797 Máximo: 2.3841

Comparação entre Reg1(sem log) e a Reg2 (com log) - Resíduos

Amplitude dos resíduos: Na primeira regressão (sem o logaritmo), os resíduos têm uma amplitude muito maior, variando de -1829.3 até 13170.7. Isso reflete que os valores previstos pelo modelo divergem significativamente dos valores reais, indicando um ajuste menos eficiente. Já na segunda regressão (com o logaritmo), os resíduos variam de -7.23 a 2.38, o que é uma amplitude muito menor. Isso sugere que o modelo com o logaritmo conseguiu reduzir a variação dos erros, proporcionando um ajuste mais eficiente.

Mediana: A mediana dos resíduos na regressão 1 é -225.5, enquanto na regressão 2 é -0.0092, o que está muito mais próximo de zero na segunda regressão. Isso indica que, na regressão logarítmica, os erros estão mais balanceados em torno de zero, o que é um sinal de melhor ajuste.

Conclusão geral: O uso do logaritmo parece melhorar a distribuição dos erros, o que é esperado quando se trabalha com dados de rendimento, já que a distribuição de salários tende a ser assimétrica e concentrada em valores mais baixos, com poucos salários muito altos puxando a média. O modelo com logaritmo ajusta melhor essa assimetria, resultando em resíduos menores e mais bem distribuídos.

Coeficientes:

Intercepto (6.704994):

O intercepto indica o valor logarítmico médio do rendimento estimado para uma pessoa com zero anos de escolaridade.

Erro padrão (0.101551): A incerteza associada a essa estimativa. **t valor (66.026):** Um valor t muito alto, sugerindo que o intercepto é significativamente diferente de zero. **p-valor (<2e-16):** Um p-valor extremamente baixo, rejeitando a hipótese nula e indicando que o intercepto é altamente significativo.

NUM_ANOS_ESTUDO (0.032923):

Este coeficiente indica que, para cada ano adicional de estudo, o logaritmo do salário (ou rendimento) aumenta em média 0.0329. Isso se traduz em um aumento percentual do salário (rendimento) com cada ano de estudo adicional.

Erro padrão (0.008282): A incerteza associada a essa estimativa é relativamente pequena. **t valor (3.975):** Um valor t significativo, indicando que o coeficiente de anos de estudo é significativamente diferente de zero. **p-valor (7.95e-05):** Um p-valor muito baixo, rejeitando a hipótese nula, sugerindo que a variável “anos de estudo” é altamente significativa para explicar o log do rendimento.

Comparação entre Reg1(sem log) e a Reg2 (com log) - Coeficientes

Regressão 1 (sem log): O intercepto é 906.34, o que significa que, para um homem com zero anos de estudo, o rendimento estimado é de R$ 906,34.

O coeficiente de anos de estudo é 44.94, o que significa que, para cada ano adicional de estudo, o rendimento de um homem aumenta em média R$ 44,94.

Regressão 2 (com log): O intercepto é 6.704994, o que indica que o valor médio do logaritmo do salário (log(1 + salário)) para um homem sem anos de estudo é de aproximadamente 6,70. Este valor, quando retirado do logaritmo, corresponderia a um salário de aproximadamente R$ 812,43 (usando exp(6.704994) - 1), o que está próximo do valor na regressão linear simples, sem log.

O coeficiente de anos de estudo é 0.032923, o que indica que, para cada ano adicional de estudo, há um aumento percentual de aproximadamente 3,29% no salário. Este valor é a interpretação no contexto de uma transformação logarítmica, onde pequenos coeficientes são aproximados pelo percentual de variação.

Estatisticas do Modelo2:

Erro padrão residual: 0.6775 A variação dos resíduos em torno da linha de regressão.

R-quadrado múltiplo: 0.02759 Apenas cerca de 2.75% da variação no logaritmo do rendimento pode ser explicada pela variável anos de estudo.

R-quadrado ajustado: 0.02584 Ajustado para o número de preditores no modelo.

Estatística F: 15.8 O valor de F significativo sugere que o modelo, como um todo, é melhor que um modelo sem preditores.

p-valor: 7.952e-05 Esse valor é extremamente baixo, sugerindo que o modelo geral é estatisticamente significativo.

Comparação entre Reg1(sem log) e a Reg2 (com log) - Estatísticas

Resíduo padrão (Residual standard error):

Reg1: O resíduo padrão é 1051. ; Reg2: O resíduo padrão é 0.6775.

Conclusão: O resíduo padrão da reg2 (com o log) é significativamente menor que o da reg1. Isso sugere que a reg2 está ajustando melhor os dados, ou seja, o modelo com o logaritmo dos rendimentos se aproxima mais dos valores observados. Usar o logaritmo costuma suavizar distribuições altamente assimétricas, como pode ser o caso de rendimentos salariais.

R² e R² ajustado: Reg1: R² = 0.04709, R² ajustado = 0.04194. Reg2: R² = 0.02759, R² ajustado = 0.02584.

Conclusão: O R² e o R² ajustado da reg1 são maiores que os da reg2, indicando que a reg1 explica uma parcela ligeiramente maior da variação no rendimento em função das variáveis explicativas. No entanto, o ajuste dos resíduos (observado no erro padrão dos resíduos) é consideravelmente melhor na reg2, apesar do R² ser menor.

F-statistic: Reg1: 9.143 com p-valor 6.498e-06. Reg2: 15.8 com p-valor 7.952e-05.

Conclusão: A estatística F da reg2 é maior, o que sugere que a variável independente (anos de estudo) tem um impacto mais significativo sobre o rendimento quando transformamos os salários em log. Mesmo com a transformação logarítmica, o modelo explica uma variação significativa.

Comparando os Resultados

Primeira regressão (rendimento direto):

O coeficiente de NUM_ANOS_ESTUDO foi 44.94, ou seja, a cada ano adicional de estudo, o rendimento aumenta em média R$ 44,94. Contudo, o ajuste do modelo foi fraco, com um $R2$ R2 muito baixo (~0.047), indicando que ele explica pouco da variabilidade dos rendimentos.

Segunda regressão (logaritmo do rendimento):

O coeficiente de NUM_ANOS_ESTUDO foi 0.032923, o que significa que, para cada ano adicional de estudo, o salário aumenta, em média, 3.29%. O$R2$R2 continua baixo, mas o uso do logaritmo pode ser mais apropriado para capturar os efeitos relativos no salário, especialmente quando há uma alta dispersão nos rendimentos, como é o caso aqui.

Na primeira regressão (reg_1): Estamos modelando o rendimento direto, sem aplicar o logaritmo à variável dependente, mas incluindo uma interação entre os anos de estudo e o sexo.

Fórmula: VAL_RENDIMENTO ~ 1 + NUM_ANOS_ESTUDO + IN_FEMININO + NUM_ANOS_ESTUDO * IN_FEMININO

Na segunda regressão (reg_2): Aplicamos o logaritmo à variável dependente, o que significa que está modelando a taxa de crescimento do rendimento em relação aos anos de estudo, ou seja, analisando como os rendimentos crescem percentualmente à medida que os anos de estudo aumentam.

Fórmula: log(1 + VAL_RENDIMENTO) ~ NUM_ANOS_ESTUDO

A transformação logarítmica melhorou a distribuição dos resíduos (menor erro padrão residual) e melhorado o ajuste do modelo de forma a suavizar a assimetria dos salários. Porém, o poder explicativo (R²) diminuiu um pouco, o que é esperado ao reduzir os efeitos de outliers. Isso mostra que o modelo reg2 captura melhor a relação percentual entre anos de estudo e rendimentos, mas que a relação não explica tanto a variação do rendimento como na reg1, que inclui o sexo e a interação com anos de estudo.

```r
# Código referente a resposta do Exercício 6

# Questão 7:
# Faça uma regressão simples do logaritmo dos salários sobre anos de escolaridade e reporte o resultado,comentando os testes de hipótese usuais.
# Se estiver usando o R, o pacote stargazer pode ser útil para a geração de tabelas.- usa LaTex

# Estimando uma regressão linear para o rendimento, considerando anos de estudo e sexo
reg_1 <- lm(VAL_RENDIMENTO ~ 1 + NUM_ANOS_ESTUDO + IN_FEMININO + NUM_ANOS_ESTUDO*IN_FEMININO, data = df_pnad_setor_privado)

# Regressão simples: logaritmo dos salários sobre anos de escolaridade
# Log coloca 1+,1 ano a mais de estudo, como impacta; log percentual de rendimento
reg_2 <- lm(log(1+VAL_RENDIMENTO) ~ NUM_ANOS_ESTUDO, data = df_pnad_setor_privado)

# Visualizando os resultados
summary(reg_1)
```

```
##
## Call:
## lm(formula = VAL_RENDIMENTO ~ 1 + NUM_ANOS_ESTUDO + IN_FEMININO +
##     NUM_ANOS_ESTUDO * IN_FEMININO, data = df_pnad_setor_privado)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1829.3  -449.5  -225.5   204.7 13170.7
##
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                    906.34     180.71   5.015 7.13e-07 ***
## NUM_ANOS_ESTUDO                 44.94      15.44   2.910  0.00376 **
## IN_FEMININO                   -834.39     384.05  -2.173  0.03023 *
## NUM_ANOS_ESTUDO:IN_FEMININO    64.89      30.00   2.163  0.03097 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1051 on 555 degrees of freedom
## Multiple R-squared:  0.04709,    Adjusted R-squared:  0.04194
## F-statistic: 9.143 on 3 and 555 DF,  p-value: 6.498e-06
```

```r
summary(reg_2)
```

```
##
## Call:
## lm(formula = log(1 + VAL_RENDIMENTO) ~ NUM_ANOS_ESTUDO, data = df_pnad_setor_privado)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.2318 -0.1933 -0.0092  0.2797  2.3841
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)     6.704994   0.101551  66.026  < 2e-16 ***
## NUM_ANOS_ESTUDO 0.032923   0.008282   3.975 7.95e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6775 on 557 degrees of freedom
## Multiple R-squared:  0.02759,    Adjusted R-squared:  0.02584
## F-statistic:  15.8 on 1 and 557 DF,  p-value: 7.952e-05
```

Exercicio 8

Enunciado Q8: Faça duas regressões simples, de anos de escolaridade sobre sexo e raça, respectivamente. O que os resultados dessa regressão nos dizem sobre a validade da regressão feita no item anterior? Analise o possível viés que pode existir, se baseando na teoria de MQO, e a direção desse viés.

Na Questão 8, foram realizadas duas regressões simples para analisar a relação entre os anos de escolaridade e as variáveis de raça (IN_PRETA_PARDA) e sexo (IN_FEMININO).

Regressão 1: Escolaridade sobre Raça (IN_PRETA_PARDA)

Intercepto (12.4536):

Esse valor indica que, para indivíduos que não se identificam como pretos ou pardos (IN_PRETA_PARDA = 0), o número médio de anos de estudo é 12.45.

Coeficiente da variável IN_PRETA_PARDA (-0.8346):

Indica que indivíduos que se identificam como pretos ou pardos têm, em média, 0.83 anos a menos de escolaridade em relação a indivíduos que não se identificam como pretos ou pardos.

Significância:

O p-valor (0.0308) é significativo, sugerindo que a diferença na escolaridade entre pretos/pardos e não pretos/pardos é estatisticamente relevante.

R^2 ajustado (0.0066): Muito baixo, indicando que a variável raça explica muito pouco da variação nos anos de estudo.

Regressão 2: Escolaridade sobre Sexo (IN_FEMININO)

Intercepto (11.1271): Refere-se ao número médio de anos de estudo para homens (IN_FEMININO = 0), que é de 11.13.

Coeficiente da variável IN_FEMININO (1.7363): Indica que, em média, as mulheres têm 1.74 anos a mais de escolaridade em comparação com os homens.

Significância: O p-valor (< 0.001) é altamente significativo, indicando que a diferença entre a escolaridade de homens e mulheres é relevante.

R^2 ajustado (0.0568): Melhor que o da variável raça, mas ainda relativamente baixo. Isso sugere que o sexo tem uma influência maior na escolaridade em relação à raça, mas ainda não explica muito da variabilidade.

Análise de Viés:

As regressões revelam que tanto raça quanto sexo afetam significativamente os anos de escolaridade, sendo que as mulheres, em média, possuem mais anos de estudo do que os homens, e indivíduos que se identificam como pretos/pardos têm menos escolaridade.

Esses fatores podem indicar viés no modelo anterior (regressão de rendimento sobre anos de estudo e sexo), pois essas variáveis (raça e sexo) afetam o nível educacional e, consequentemente, podem influenciar o rendimento.

Portanto, o modelo anterior pode estar omitindo variáveis relevantes (como raça), introduzindo viés de variável omitida, o que pode distorcer a interpretação dos coeficientes, especialmente se essas variáveis estão correlacionadas com os anos de estudo.

A direção do viés seria negativa para pretos/pardos, visto que eles tendem a ter menos anos de estudo, e positiva para mulheres, que tendem a ter mais anos de estudo.

```
# Código referente a resposta do Exercício 7

# Questão 8:
# Faça duas regressões simples, de anos de escolaridade sobre sexo e raça, respectivamente.
# O que os resultados dessa regressão nos dizem sobre a validade da regressão feita no item anterior?
# Analise o possível viés que pode existir, se baseando na teoria de MQO, e a direção desse viés.

# Regressão de anos de escolaridade sobre raça - Preconceito racial
reg_3 <- lm(NUM_ANOS_ESTUDO ~ IN_PRETA_PARDA, data = df_pnad_setor_privado)
summary(reg_3)
```

```
##
## Call:
## lm(formula = NUM_ANOS_ESTUDO ~ IN_PRETA_PARDA, data = df_pnad_setor_privado)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -11.619  -0.619   0.381   2.381   4.381
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)     12.4536     0.3505  35.532   <2e-16 ***
## IN_PRETA_PARDA  -0.8346     0.3855  -2.165   0.0308 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.452 on 557 degrees of freedom
## Multiple R-squared:  0.008343,   Adjusted R-squared:  0.006562
## F-statistic: 4.686 on 1 and 557 DF,  p-value: 0.03083
```

```r
# Regressão de anos de escolaridade sobre gênero (sexo)
reg_4 <- lm(NUM_ANOS_ESTUDO ~ IN_FEMININO, data = df_pnad_setor_privado)
summary(reg_4)
```

```
##
## Call:
## lm(formula = NUM_ANOS_ESTUDO ~ IN_FEMININO, data = df_pnad_setor_privado)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -12.8634  -0.8634   0.8729   2.1366   4.8729
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  11.1271     0.1788  62.243  < 2e-16 ***
## IN_FEMININO   1.7363     0.2952   5.882 7.01e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.364 on 557 degrees of freedom
## Multiple R-squared:  0.05848,    Adjusted R-squared:  0.05679
## F-statistic: 34.59 on 1 and 557 DF,  p-value: 7.007e-09
```

```r
# Em media mulheres estudam xxx mais que Homem
```

Exercicio 9

Enunciado Q9: Faça duas regressões, de logaritmo dos salários sobre anos de educação, sendo uma apenas com mulheres e outra apenas com homens. Os coeficientes estimados são os mesmos? O que isso nos diz sobre o que precisamos incluir na regressão completa? Além do modelo, vocês devem gerar um gráfico com os pontos da amostra (ou uma subamostra deles, se o código ficar lento), e as duas retas de regressão.

Regressão 6 (Mulheres)

Intercepto (6.43642): Isso sugere que, para uma mulher com zero anos de estudo, o logaritmo do salário é de aproximadamente 6.44. Esse valor não é diretamente interpretável como o salário em si, mas como o logaritmo do salário, devido à transformação logarítmica.

NUM_ANOS_ESTUDO (0.04917): Para mulheres, um ano adicional de escolaridade está associado a um aumento de aproximadamente 4.92% no salário, quando o salário é expresso em termos logarítmicos. Isso significa que o retorno da educação para mulheres é positivo, embora relativamente pequeno.

Regressão 6 (Homens)

Intercepto (6.753705): Para homens com zero anos de estudo, o logaritmo do salário é de 6.75, ligeiramente superior ao das mulheres.

NUM_ANOS_ESTUDO (0.031649): Para homens, um ano adicional de escolaridade está associado a um aumento de aproximadamente 3.16% no salário. Esse valor é menor do que o observado para mulheres, sugerindo que o retorno da escolaridade para homens é menor em comparação às mulheres.

Comparação dos Coeficientes

Os coeficientes para NUM_ANOS_ESTUDO são diferentes para homens e mulheres, com o retorno educacional sendo maior para mulheres (0.049 vs. 0.031). Essa diferença sugere que, para o mesmo aumento em anos de estudo, as mulheres têm um aumento maior no salário percentual (considerando o logaritmo) em relação aos homens. Portanto, a inclusão de uma interação entre o gênero e os anos de estudo pode ser necessária para capturar essa diferença entre homens e mulheres na regressão completa.

Validade da Regressão Completa

Dado que os coeficientes são diferentes entre homens e mulheres, isso nos indica que uma única regressão que não leve em consideração essa diferença pode estar subestimando ou superestimando o impacto dos anos de estudo sobre os salários, dependendo do gênero. Para evitar viés na análise, seria importante incluir uma interação entre gênero e escolaridade no modelo completo, como foi feito anteriormente na regressão 1, o que permitiria modelar as diferenças entre os gêneros corretamente.

Além disso, os gráficos das duas retas de regressão, um para cada gênero, também ajudariam a visualizar essa diferença no impacto dos anos de escolaridade sobre os salários.

Interpretação do Gráfico:

Mostra a relação entre os anos de estudo e o logaritmo do salário, separados por gênero (homens e mulheres). Podemos observar que:

Diferença entre as Retas de Regressão:

As retas de regressão para homens (azul) e mulheres (vermelho) são muito próximas, mas há uma leve diferença. Para mulheres, a inclinação é um pouco maior, indicando que o aumento nos anos de estudo tem um impacto ligeiramente maior nos salários das mulheres em comparação aos homens, o que está alinhado com os coeficientes obtidos nas regressões. Distribuição dos Dados:

A maior concentração de dados está em torno de 5 a 15 anos de estudo, com menos dados para extremos (muito poucos anos de estudo ou muitos anos). A dispersão é maior para homens com poucos anos de estudo. Variação:

Embora as inclinações sejam similares, a variação dos dados é considerável, principalmente para salários menores, onde existem mais outliers. Para salários maiores, a variabilidade é menor.

Interpretação:

O gráfico sugere que os anos de estudo afetam os salários de maneira semelhante para homens e mulheres, embora o impacto seja um pouco maior para as mulheres. No entanto, outras variáveis (como raça ou posição no mercado de trabalho) podem estar influenciando os salários e precisam ser incluídas para capturar o efeito completo, como discutido nas perguntas anteriores. Esse gráfico visualiza a diferença no impacto dos anos de estudo sobre os salários entre gêneros, confirmando a necessidade de incluir a interação entre gênero e escolaridade em uma regressão mais completa, para capturar essa variação de maneira mais precisa.

```r
# Código referente a resposta do Exercício 8

# Questão 9: Regressões separadas por gênero (homens e mulheres), verificando os coeficientes.
# Faça duas regressões, de logaritmo dos salários sobre anos de educação, sendo uma apenas com mulheres e outra apenas com homens.
# Os coeficientes estimados são os mesmos? O que isso nos diz sobre o que precisamos incluir na regressão completa?
# Além do modelo, vocês devem gerar um gráfico com os pontos da amostra (ou uma subamostra deles, se o código ficar lento), e as duas retas de regressão.


# Criando subsets para homens e mulheres
df_pnad_homens <- df_pnad_setor_privado[df_pnad_setor_privado$IN_FEMININO==0,]
df_pnad_mulheres <- df_pnad_setor_privado[df_pnad_setor_privado$IN_FEMININO==1,]

# Regressões para homens e mulheres (logaritmo dos salários)
reg_5.homens <- lm(VAL_RENDIMENTO ~ NUM_ANOS_ESTUDO, data = df_pnad_homens)
reg_6.homens <- lm(log(1+VAL_RENDIMENTO) ~ NUM_ANOS_ESTUDO, data = df_pnad_homens)

# Visualizando
summary(reg_5.homens)
```

```
##
## Call:
## lm(formula = VAL_RENDIMENTO ~ NUM_ANOS_ESTUDO, data = df_pnad_homens)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1280.5  -447.7  -222.9   174.5 10374.5
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)       906.34     156.12   5.805 1.43e-08 ***
## NUM_ANOS_ESTUDO    44.94      13.34   3.368  0.00084 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 908.1 on 352 degrees of freedom
## Multiple R-squared:  0.03123,    Adjusted R-squared:  0.02847
## F-statistic: 11.35 on 1 and 352 DF,  p-value: 0.00084
```

```r
summary(reg_6.homens)
```

```
##
## Call:
## lm(formula = log(1 + VAL_RENDIMENTO) ~ NUM_ANOS_ESTUDO, data = df_pnad_homens)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.29683 -0.22674 -0.03971  0.24489  2.13266
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)     6.753705   0.090011  75.032  < 2e-16 ***
## NUM_ANOS_ESTUDO 0.031649   0.007693   4.114 4.85e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5235 on 352 degrees of freedom
## Multiple R-squared:  0.04588,    Adjusted R-squared:  0.04316
## F-statistic: 16.92 on 1 and 352 DF,  p-value: 4.847e-05
```

```
reg_5.mulheres <- lm(VAL_RENDIMENTO ~ NUM_ANOS_ESTUDO, data = df_pnad_mulheres)
reg_6.mulheres <- lm(log(1+VAL_RENDIMENTO) ~ NUM_ANOS_ESTUDO, data = df_pnad_mulheres)

summary(reg_5.mulheres)
```

```
##
## Call:
## lm(formula = VAL_RENDIMENTO ~ NUM_ANOS_ESTUDO, data = df_pnad_mulheres)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1829.3  -531.0  -290.0   239.5 13170.7
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)        71.95     406.62   0.177 0.859735
## NUM_ANOS_ESTUDO   109.83      30.86   3.559 0.000463 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1261 on 203 degrees of freedom
## Multiple R-squared:  0.05874,    Adjusted R-squared:  0.0541
## F-statistic: 12.67 on 1 and 203 DF,  p-value: 0.0004633
```

```
summary(reg_6.mulheres)
```

```
##
## Call:
## lm(formula = log(1 + VAL_RENDIMENTO) ~ NUM_ANOS_ESTUDO, data = df_pnad_mulheres)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.2231 -0.1688  0.0278  0.3520  2.3928
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)      6.43642    0.28408   22.66   <2e-16 ***
## NUM_ANOS_ESTUDO  0.04917    0.02156    2.28   0.0236 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8811 on 203 degrees of freedom
## Multiple R-squared:  0.02498,    Adjusted R-squared:  0.02017
## F-statistic:   5.2 on 1 and 203 DF,  p-value: 0.02362
```

```
# Criando gráfico para visualizar a regressão de log do salário por anos de estudo, separad
o por gênero
plot_2 <- ggplot(data = df_pnad_setor_privado,
                 aes(x = NUM_ANOS_ESTUDO, y = log(1+VAL_RENDIMENTO), color = as.factor(IN_F
EMININO))) +
  geom_point() +
  geom_smooth(method = "lm") +
  scale_color_manual(name = "Gênero",
                     values = c("0" = "blue", "1" = "red"),
                     labels = c("Homem", "Mulher")) +
  labs(
    title = "Log do Valor de Rendimentos por \nAnos de Estudo",
    x = "Anos de Estudo",
    y = "Log do Valor de Rendimentos"
  )

# Exibindo o gráfico
plot_2
```

```
## `geom_smooth()` using formula = 'y ~ x'
```

Exercicio 10

Enunciado Q10: Com base na sua conclusão do item anterior, façam uma regressão para explicar o logaritmo dos salários, incluindo todas as variáveis vistas até aqui, e interpretem os valores dos coeficientes estimados, assim como a significância deles e de quaisquer interações incluídas. Qual é o valor marginal de um ano adicional de escolaridade para um homem? E para uma mulher? Eles são estatisticamente iguais?

Interpretação dos coeficientes:

Intercepto (6.753705): Este valor representa o logaritmo do salário esperado para homens (pois IN_FEMININO = 0) com zero anos de estudo. O salário inicial de um homem, quando expresso no formato logarítmico, é 6.75.

IN_FEMININO (−0.317288): O coeficiente negativo sugere que, em média, as mulheres têm um salário logarítmico menor que os homens, ou seja, as mulheres ganham menos que os homens, quando controlamos pelo número de anos de estudo. No entanto, o p-valor (0.19989) indica que essa diferença não é estatisticamente significativa.

NUM_ANOS_ESTUDO (0.031649): Esse coeficiente significa que, para cada ano adicional de escolaridade, o salário logarítmico de um homem aumenta em aproximadamente 0.0316. Isso sugere que os homens ganham, em média, 3.16% a mais no salário para cada ano adicional de estudo. Esse efeito é significativo, já que o p-valor (0.00154) é muito pequeno.

IN_FEMININO: NUM_ANOS_ESTUDO(0.017516): O coeficiente de interação positivo sugere que o efeito dos anos de estudo sobre o salário é maior para mulheres do que para homens. No entanto, o p-valor (0.36479) é alto, o que indica que essa interação não é estatisticamente significativa. Isso significa que, embora o efeito dos anos de estudo possa ser um pouco maior para as mulheres, não há evidência suficiente para afirmar que isso é uma diferença real e significativa.

Conclusão sobre os valores marginais:

Para homens, o valor marginal de um ano adicional de escolaridade é 0.031649, o que significa que um ano extra de estudo aumenta o salário em 3.16% (aproximadamente), de forma estatisticamente significativa.

Para mulheres, o valor marginal seria a soma de 0.031649 e o coeficiente de interação 0.017516, resultando em 0.049165 (ou seja, 4.91% de aumento no salário logarítmico para cada ano adicional de estudo). Porém, como o coeficiente de interação não é significativo, essa diferença pode não ser real.

** Conclusão sobre a igualdade dos coeficientes:**

Os coeficientes não são estatisticamente iguais, já que o efeito dos anos de estudo é maior para as mulheres do que para os homens, embora a diferença não seja significativa. Isso sugere que, ao construir um modelo de regressão completo, pode ser necessário incluir a interação entre gênero e anos de escolaridade para capturar melhor essas variações.

```
# Código referente a resposta do Exercício 10

# Questão 10: Gráfico de regressão por gênero.

# Com base na sua conclusão do item anterior, façam uma regressão para explicar o logaritmo dos salários,incluindo todas as variáveis vistas até aqui,
# e interpretem os valores dos coeficientes estimados, assim como a significância deles e de quaisquer interações incluídas.
# Qual é o valor marginal de um ano adicional de escolaridade para um homem? E para uma mulher? Eles são estatisticamente iguais?
# Tem que ver o pvalue pra responder

# Regressão completa para explicar o logaritmo dos salários incluindo anos de estudo e interação com o gênero
reg_7 <- lm(log(1 + VAL_RENDIMENTO) ~ IN_FEMININO + NUM_ANOS_ESTUDO + IN_FEMININO:NUM_ANOS_ESTUDO, data = df_pnad_setor_privado)

# Visualizando o resumo da regressão para verificar a significância e interpretação dos coeficientes
summary(reg_7)
```

```
##
## Call:
## lm(formula = log(1 + VAL_RENDIMENTO) ~ IN_FEMININO + NUM_ANOS_ESTUDO +
##     IN_FEMININO:NUM_ANOS_ESTUDO, data = df_pnad_setor_privado)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.2231 -0.2267 -0.0259  0.2875  2.3928
##
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                  6.753705   0.116327  58.058  < 2e-16 ***
## IN_FEMININO                 -0.317288   0.247225  -1.283  0.19989
## NUM_ANOS_ESTUDO              0.031649   0.009942   3.183  0.00154 **
## IN_FEMININO:NUM_ANOS_ESTUDO  0.017516   0.019312   0.907  0.36479
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6766 on 555 degrees of freedom
## Multiple R-squared:  0.03367,    Adjusted R-squared:  0.02845
## F-statistic: 6.447 on 3 and 555 DF,  p-value: 0.0002698
```

Exercicio 11

Enunciado Q11: É possível que o nível de experiência tenha efeito importante sobre os salários, porém há dúvidas sobre como deve ser a forma funcional desse efeito. Como poderíamos incorporar efeitos não-lineares no modelo? Supondo que o efeito seja quadrático, qual seria o nível de experiência que maximiza o salário, tudo mais constante?

Com base nos resultados das regressões rodadas:

Regressão Quadrática (reg_8):

Interpretação dos Coeficientes:

Intercepto (5.17): Representa o valor esperado do log do salário quando todas as outras variáveis são zero.

NUM_ANOS_ESTUDO (0.045): Cada ano adicional de estudo aumenta o log do rendimento em cerca de 4.5%, mantidas as outras variáveis constantes.

NUM_IDADE (0.066): Cada ano adicional de idade aumenta o log do rendimento em aproximadamente 6.6%.

NUM_IDADE_QUADRADO (-0.00065): O coeficiente quadrático de idade é negativo, indicando que o impacto da idade no rendimento diminui após um certo ponto, o que sugere uma relação côncava entre idade e salário.

Idade Ótima (50.51):

Calculada a partir da função quadrática, a idade de 50 anos é onde o salário é maximizado. Após essa idade, o efeito da idade no salário diminui.

Regressão Adicional (reg_9):

Interpretação dos Coeficientes:

NUM_ANOS_ESTUDO (0.041): Cada ano adicional de estudo aumenta o log do rendimento por hora em 4.1%.

Interação IN_FEMININO*NUM_ANOS_ESTUDO (-0.006): A interação entre ser mulher e o número de anos de estudo não foi significativa, o que sugere que o efeito dos anos de estudo sobre o salário por hora é semelhante para homens e mulheres.

NUM_IDADE (0.032) e NUM_IDADE_QUADRADO (-0.00024): O efeito quadrático da idade continua presente, sugerindo que o salário por hora atinge um pico com a idade e depois diminui.

Conclusão para Q11:

A inclusão do termo quadrático para a idade no modelo se mostrou adequada para capturar a relação não linear entre idade e salário. A idade ótima calculada, em torno de 50 anos, faz sentido à luz de considerações teóricas e práticas sobre o mercado de trabalho, e o modelo fornece uma boa base para entender como a idade influencia os rendimentos ao longo da vida profissional.

```
# Código referente a resposta do Exercício 11

# Questão 11:Incorporando efeito quadrático de experiência.

# É possível que o nível de experiência tenha efeito importante sobre os salários,porém há
dúvidas sobre como deve ser a forma funcional desse efeito.
# Como poderíamos incorporar efeitos não-lineares no modelo?
# Supondo que o efeito seja quadrático, qual seria o nível de experiência que maximiza o sa
lário, tudo mais constante?

# Criando uma variável quadrática para idade (proxy de experiência)
df_pnad_setor_privado$NUM_IDADE_QUADRADO <- df_pnad_setor_privado$NUM_IDADE**2

# Regressão com o efeito quadrático da idade e interação com anos de estudo e gênero
reg_8 <- lm(log(1+VAL_RENDIMENTO) ~ IN_FEMININO*NUM_ANOS_ESTUDO + NUM_IDADE + NUM_IDADE_QUA
DRADO, data = df_pnad_setor_privado)

# Calculando o ponto de máximo da idade, onde o salário é maximizado
b1 <- reg_8$coefficients["NUM_IDADE"]
b2 <- reg_8$coefficients["NUM_IDADE_QUADRADO"]
idade_otima <- -b1 / (2 * b2)

# Exibindo a idade ótima (máximo da função quadrática)
idade_otima
```

```
## NUM_IDADE
##   50.5144
```

```r
# Criando uma nova variável para o rendimento por hora
df_pnad_setor_privado <- df_pnad_setor_privado %>% mutate(
  VAL_RENDIMENTO_HORA = VAL_RENDIMENTO/(NUM_HORAS*4)
)

df_pnad_setor_privado <- df_pnad_setor_privado %>% mutate(
  VAL_RENDIMENTO_HORA = VAL_RENDIMENTO / (NUM_HORAS * 4) # Calcular o rendimento por hora
)


# Verificando se a variável foi criada corretamente
head(df_pnad_setor_privado$VAL_RENDIMENTO_HORA)
```

```
## [1]  8.333333 13.888889  7.222222  8.333333  8.750000 11.111111
```

```r
# Regressão adicional para explicar o log do rendimento por hora, incluindo variáveis adici
onais
reg_9 <- lm(log(1+VAL_RENDIMENTO_HORA)~
              IN_URBANA +
              IN_FEMININO*NUM_ANOS_ESTUDO +
              IN_PRETA_PARDA*NUM_ANOS_ESTUDO+
              NUM_IDADE +
              NUM_IDADE_QUADRADO +
              as.factor(COD_POSICAO_OCUPACAO), data = df_pnad_setor_privado
)

# Visualizando o resumo da regressão com o efeito quadrático e interações
summary(reg_9)
```

```
##
## Call:
## lm(formula = log(1 + VAL_RENDIMENTO_HORA) ~ IN_URBANA + IN_FEMININO *
##     NUM_ANOS_ESTUDO + IN_PRETA_PARDA * NUM_ANOS_ESTUDO + NUM_IDADE +
##     NUM_IDADE_QUADRADO + as.factor(COD_POSICAO_OCUPACAO), data = df_pnad_setor_privado)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.33989 -0.27463 -0.04632  0.23870  2.07976
##
## Coefficients:
##                                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)                      1.0302555  0.3235228   3.184  0.00153 **
## IN_URBANA                        0.0091771  0.0893264   0.103  0.91821
## IN_FEMININO                     -0.1706021  0.1652756  -1.032  0.30242
## NUM_ANOS_ESTUDO                  0.0476010  0.0147989   3.217  0.00137 **
## IN_PRETA_PARDA                  -0.0554542  0.1962873  -0.283  0.77765
## NUM_IDADE                        0.0328151  0.0134385   2.442  0.01493 *
## NUM_IDADE_QUADRADO              -0.0002487  0.0001785  -1.393  0.16415
## as.factor(COD_POSICAO_OCUPACAO)02 -0.2615435  0.0415515  -6.294 6.32e-10 ***
## IN_FEMININO:NUM_ANOS_ESTUDO      0.0067433  0.0129073   0.522  0.60157
## NUM_ANOS_ESTUDO:IN_PRETA_PARDA  -0.0043996  0.0153984  -0.286  0.77520
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.447 on 549 degrees of freedom
## Multiple R-squared:  0.2261, Adjusted R-squared:  0.2134
## F-statistic: 17.82 on 9 and 549 DF,  p-value: < 2.2e-16
```