LEGO 2 - 1ª LISTA DE EXERCÍCIOS

PARTE 1 - CONCEITUAL

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

Seja uma variável aleatória $X = (X_1, X_2, \dots, X_n)$

QUEREMOS VERIFICAR: $\sum_{i=1}^{n} (X_i - \bar{X}) \overset{?}{=} 0$

PELA DEFINIÇÃO DE MÉDIA, SABEMOS: $\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i \Rightarrow \boxed{\sum_{i=1}^{n} X_i = n \cdot \bar{X}}$ (1)

PELAS PROPRIEDADES DO SOMATÓRIO:

$$\sum_{i=1}^{n} (X_i - \bar{X}) = \sum_{i=1}^{n} X_i - \sum_{i=1}^{n} \bar{X} = \boxed{\sum_{i=1}^{n} X_i - n \cdot \bar{X}} = 0$$

De fato, eu já poderia ter chegado à mesma conclusão usando a própria definição de média...

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)

a) O PGD é o mecanismo que gera os dados do mundo real.

b) PARÂMETROS SÃO QUANTIDADES POPULACIONAIS, EM GERAL DESCONHECIDAS; ESTATÍSTICAS SÃO QUANTIDADES

ESTIMADAS PARA OS PARÂMETROS, FREQUENTEMENTE A PARTIR DE AMOSTRAS.

c) ERROS SÃO COMPONENTES ALEATÓRIOS (não observável!) QUE FAZEM COM QUE PESSOAS IDÊNTICAS (i.e., COM AS MESMAS CARACTERÍSTICAS) APRESENTEM DIFERENÇAS NA VARIÁVEL DE INTERESSE. RESÍDUOS SÃO QUANTIDADES ESTIMADAS: VALOR OBSERVADO - ESTIMADO.

d) PODEMOS ESTIMAR A VARIÁVEL DE INTERESSE Y A PARTIR DE UMA COMBINAÇÃO LINEAR DAS COLUNAS DE UMA MATRIZ X.

e) SUPOSIÇÃO DE AMOSTRAGEM ALEATÓRIA, OU SEJA, A INCLUSÃO DE UM INDIVÍDUO NA AMOSTRA NÃO TEM QUALQUER RELAÇÃO COM A INCLUSÃO DE OUTRO.

f) NENHUM REGRESSOR PODE SER UMA COMBINAÇÃO LINEAR EXATA DE UM OU MAIS REGRESSORES.

g) O ERRO SE DISTRIBUI DE FORMA ALEATÓRIA, DE FORMA INDEPENDENTE DOS REGRESSORES X.

h) OS ERROS POSSUEM A MESMA VARIÂNCIA PARA TODOS OS VALORES DAS VARIÁVEIS EXPLICATIVAS.

i) $Var(u|X) = \sigma^2$, i.e., O ERRO TEM A MESMA VARIÂNCIA DADO QUALQUER VALOR DA VARIÁVEL EXPLICATIVA.

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?

a) A CLÁUSULA CETERIS PARIBUS SIGNIFICA "TUDO MAIS CONSTANTE", IMPORTANTE PORQUE QUEREMOS MEDIR O EFEITO ISOLADO DE UMA PARTICULAR COVARIÁVEL NA VARIÁVEL RESPOSTA — OU SEJA, SABER COM Y VARIA DADA UMA VARIAÇÃO EM X.

b) SATISFAZER A EXOGENEIDADE SIGNIFICA DIZER QUE ESTAMOS CONTROLANDO POR TODAS AS VARIÁVEIS QUE AFETAM Y, DE MODO QUE O ERRO SE DISTRIBUIRÁ ALEATORIAMENTE, DE FORMA INDEPENDENTE DOS REGRESSORES. NÃO HAVENDO OUTROS FATORES SISTEMÁTICOS AFETANDO Y, PODEMOS INTERPRETAR OS COEFICIENTES COMO EFEITOS CAUSAIS.

c) UMA ÚNICA VARIÁVEL, SOZINHA, DIFICILMENTE EXPLICARÁ TODA A VARIAÇÃO DE Y. AO CONTROLARMOS POR MAIS VARIÁVEIS, ESTAMOS MAIS PRÓXIMOS DE SATISFAZER A EXOGENEIDADE.

d) NOS CASOS EM QUE VIOLAMOS O PRESSUPOSTO DA EXOGENEIDADE (quase sempre), DEVEMOS INTERPRETAR OS COEFICIENTES DA REGRESSÃO NÃO COMO EFEITOS CAUSAIS, MAS COMO UM RESUMO DA CORRELAÇÃO ENTRE AS VARIÁVEIS.

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?

a) A INCLUSÃO DE DUAS VARIÁVEIS IDÊNTICAS INTRODUZ COLINEARIDADE PERFEITA. COM ISSO, (X^TX) SE TORNA UMA MATRIZ SINGULAR E NÃO CONSEGUIMOS ESTIMAR β.

b) QUANDO TEMOS UMA VARIÁVEL BINÁRIA COMO O SEXO, SÓ PODEMOS INCLUIR UMA DAS DUMMIES COMO REGRESSOR PARA EVITAR COLINEARIDADE PERFEITA COM O INTERCEPTO (COLUNA DE 1s). SE INCLUÍMOS AS DUAS DUMMIES, TERÍAMOS INTERCEPTO = $DUMMY_H + DUMMY_M$.

c) AO SUPRIMIR O INTERCEPTO, ELIMINAMOS O PROBLEMA DA COLINEARIDADE PERFEITA, PORQUE A SOMA DAS DUMMIES NÃO POSSUI MAIS UM CORRESPONDENTE EM X.

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.

ASSUMA A DISPONIBILIDADE DE UMA AMOSTRA ALEATÓRIA $\{(x_i, y_i) : i = 1, \dots, n\}$. → POR ML.2

A REGRESSÃO LINEAR SIMPLES PODE SER ESCRITA COMO: → por ML1

$$y_i = \beta_0 + \beta_1 x_i + u_i \quad (1)$$

ASSIM COMO EM WOOLDRIDGE (2020), VAMOS DERIVAR OS ESTIMADORES β_0 e β_1 PARTINDO DO PRESSUPOSTO DA EXOGENEIDADE, ML.4:

$\mathbb{E}[u] = 0$, E TAMBÉM $COV(x, u) = \mathbb{E}[xu] = 0$

PODEMOS REESCREVER (1) COMO $u = y - \beta_0 - \beta_1 x$, E então temos que

$$\mathbb{E}[y - \beta_0 - \beta_1 x] = 0 \quad (2) \quad \text{E} \quad \mathbb{E}[x(y - \beta_0 - \beta_1 x)] = 0 \quad (3)$$

As equivalências amostrais dos momentos (2) e (3) são:

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0 \quad (4) \quad \text{E} \quad \frac{1}{n}\sum_{i=1}^{n} x_i(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0 \quad (5)$$

PODEMOS REESCREVER (4) PARA ENCONTRAR $\hat{\beta}_0$:

$$\frac{1}{n}\sum_{i=1}^{n} y_i - \frac{1}{n}\sum_{i=1}^{n}\hat{\beta}_0 - \frac{1}{n}\sum_{i=1}^{n}\hat{\beta}_1 x_i = 0 \Rightarrow \bar{y} - \hat{\beta}_0 - \hat{\beta}_1\bar{x} = 0$$

$$\Rightarrow \boxed{\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x}} \quad (6)$$

AGORA BASTA SUBSTITUIR (6) EM (5) PARA ENCONTRAR $\hat{\beta}_1$:

$$\frac{1}{n}\sum_{i=1}^{n} x_i\left(y_i - (\bar{y} - \hat{\beta}_1\bar{x}) - \hat{\beta}_1 x_i\right) = 0 \Rightarrow \frac{1}{n}\sum_{i=1}^{n} x_i\left(y_i - \bar{y} + \hat{\beta}_1\bar{x} - \hat{\beta}_1 x_i\right) = 0$$

$$\Rightarrow \frac{1}{n}\sum_{i=1}^{n} x_i(y_i - \bar{y}) + \frac{1}{n}\sum_{i=1}^{n} x_i(\hat{\beta}_1\bar{x} - \hat{\beta}_1 x_i) = 0 \Rightarrow$$

$$\Rightarrow \cancel{\frac{1}{n}}\sum_{i=1}^{n} x_i(y_i - \bar{y}) = \hat{\beta}_1 \cancel{\frac{1}{n}} \sum_{i=1}^{n} x_i(x_i - \bar{x}) \Rightarrow \boxed{\hat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i(y_i - \bar{y})}{\sum_{i=1}^{n} x_i(x_i - \bar{x})}}$$

ou ainda $\hat{\beta}_1 = \dfrac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$, USANDO AS PROPRIEDADES DO SOMATÓRIO

↳ $x_i = (x_i - \bar{x}) + \bar{x}$

$$\sum_{i=1}^{n} x_i(x_i - \bar{x}) = \sum_{i=1}^{n}\left[(x_i - \bar{x}) + \bar{x}\right](x_i - \bar{x})$$

$$= \sum_{i=1}^{n}(x_i - \bar{x})^2 + \underbrace{\cancel{\sum_{i=1}^{n}\bar{x}(x_i - \bar{x})}}_{0} \rightarrow \text{(ver exercício 1)}$$

$$= \sum_{i=1}^{n}(x_i - \bar{x})^2.$$

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.

POR ML.1, ESCREVEMOS $\vec{y} = X\vec{\beta} + \vec{e}$. QUEREMOS ISOLAR $\vec{\beta}$.

POR ML.3, ASSUMIMOS QUE NÃO HÁ COLINEARIDADE PERFEITA E QUE X^TX, PORTANTO, É INVERTÍVEL.

$$\vec{y} = X\vec{\beta} + \vec{e} \Rightarrow X^Ty = X^TX\vec{\beta} + \underbrace{X^T\vec{e}}_{0,\ \text{por ML4 (exogeneidade)}}$$

$$\Rightarrow (X^TX)^{-1}X^Ty = \underbrace{\cancel{(X^TX)^{-1}X^TX}}_{I}\vec{\beta} \Rightarrow \boxed{\vec{\beta} = (X^TX)^{-1}X^Ty}$$

* como não estamos falando em variância, ainda não precisamos usar o pressuposto da homocedasticidade, ML.5.

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.

Em geral, queremos minimizar uma particular função de erro. No MQO, em particular, queremos minimizar a distância quadrática entre o y verdadeiro (y_i) e o y estimado ($\hat{y}_i$).

Por ML_1: $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i + u_i$, $i = 1, \dots, n$, numa amostra aleatória → ML 2 $\{(x_i, y_i) : i = 1, \dots, n\}$

Queremos minimizar a loss function $\mathcal{L}(\beta_0, \beta_1)$:

$$\mathcal{L}(\beta_0, \beta_1) = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2$$

Daí podemos extrair o seguinte sistema de equações:

$$\begin{cases} \dfrac{\partial f}{\partial \beta_0} = -\cancel{2}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i) = 0 \\[2ex] \dfrac{\partial f}{\partial \beta_1} = -\cancel{2}\sum_{i=1}^{n} x_i\,(y_i - \beta_0 - \beta_1 x_i) = 0 \end{cases}$$

$\Rightarrow$ ponto de mínimo erro com $\hat{\beta}_0$ e $\hat{\beta}_1$ (i.e., a derivada do erro com respeito aos dois parâmetros é zero)

Note que chegamos precisamente às mesmas equações do exercício 5 e, portanto, chegaríamos precisamente aos mesmos estimadores para $\hat{\beta}_1$ e $\hat{\beta}_0$. Dispensamos mais cálculos, portanto, já que eles foram realizados naquela oportunidade.

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.

AS DIFERENÇAS CONCEITUAIS PRINCIPAIS ESTÃO NA MOTIVAÇÃO PARA ENCONTRAR $\hat{\beta}$. O MÉTODO DOS MOMENTOS ENCONTRA $\hat{\beta}$ IGUALANDO MOMENTOS POPULACIONAIS A MOMENTOS AMOSTRAIS; O MQO, POR OUTRO LADO, ENCONTRA $\hat{\beta}$ MINIMIZANDO O ERRO QUADRÁTICO ENTRE O Y OBSERVADO E O Y ESTIMADO.

Lego 2 – 1ª Lista de Exercícios

Parte 2: Prática

Felipe Lamarca

2025-10-05

Agora, passemos à parte prática da lista, na qual desenvolveremos algumas análises em `R` utilizando um banco de dados com informações a respeito dos municípios brasileiros em 1991, como a expectativa de vida ao nascer, população do município, Índice de Gini etc. De saída, vamos começar importando as bibliotecas necessárias, organizando um setup e lendo o arquivo que será utilizado na análise:

```
# importa as libs
library(tidyverse)
library(stargazer)

# remove o uso de e^
options(scipen = 999)

# leitura do arquivo
expectativaVida = read.csv("dados/dados_expectativaVida.csv")
```

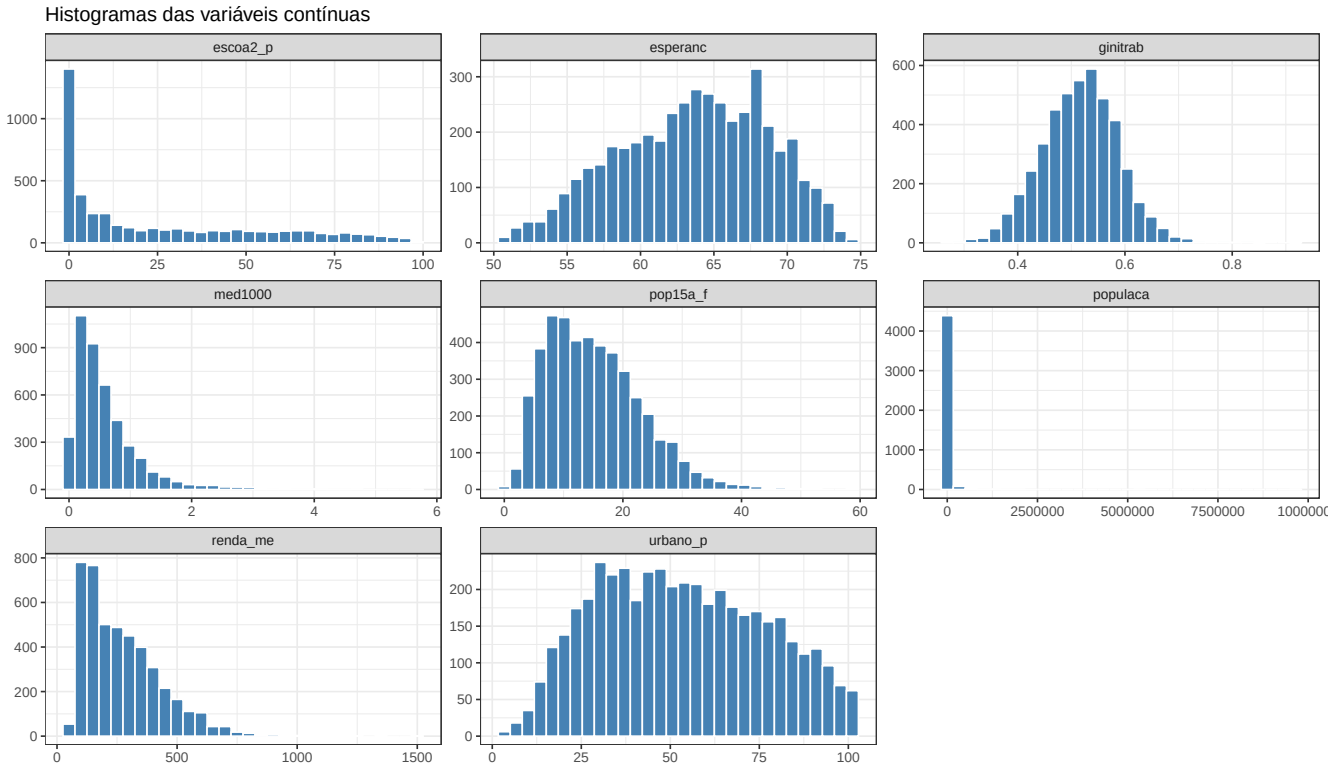
O *output* do código abaixo é uma série de histogramas, uma para cada variável contínua do banco de dados sob análise:

```
# seleciona apenas as variáveis contínuas
dados_cont <- expectativaVida %>%
  select(esperanc, populaca, ginitrab, renda_me,
         escoa2_p, urbano_p, med1000, pop15a_f)

# transforma em formato longo pro facet
dados_long <- dados_cont %>%
  pivot_longer(cols = everything(),
               names_to = "variavel",
               values_to = "valor")

# plota todos os histogramas em facets
  ggplot(dados_long, aes(x = valor)) +
  geom_histogram(bins = 30, fill = "steelblue", color = "white") +
  facet_wrap(~ variavel, scales = "free", ncol = 3) +
  labs(title = "Histogramas das variáveis contínuas",
```

```
       x = NULL, y = NULL) +
  theme_bw(base_size = 11)
```



Há alguns pontos que valem destaque. Em particular, está claro que a grande parte dos municípios brasileiros possui população (`populaca`) bem pequenas; além disso, boa parte também possui um baixo percentual de domicílios com esgotamento sanitário (`escoa2_p`). Vejamos na tabela abaixo algumas estatísticas descritivas para cada variável do banco:

```
# tabela com as estatisticas descritivas
desc = data.frame(
  variavel = names(expectativaVida),
  media = sapply(
    expectativaVida,
    function(x) if (is.numeric(x)) mean(x, na.rm = TRUE) else NA_real_
    ),
  mediana = sapply(
    expectativaVida,
    function(x) if (is.numeric(x)) median(x, na.rm = TRUE) else NA_real_
    ),
  desvio_padrao = sapply(
    expectativaVida,
    function(x) if (is.numeric(x)) sd(x, na.rm = TRUE) else NA_real_
    ),
  n_validos = sapply(
    expectativaVida,
```

```
    function(x) sum(!is.na(x))
    ),
  row.names = NULL
)

# numero de linhas
n = nrow(expectativaVida)

# proporcao de linhas validas (acho util, em particular)
desc$prop_validos = (desc$n_validos / n) * 100

# plot bonitinho da tabela
stargazer(
  desc,
  type = "text",
  summary = FALSE,
  title = "Estatísticas Descritivas",
  digits = 2,
  rownames = FALSE
)
```

```
Estatísticas Descritivas
=================================================================
variavel   media   mediana desvio_padrao n_validos prop_validos
-----------------------------------------------------------------
esperanc   63.63    64.05      5.08        4,491       100
populaca 32,693.27 12,407   187,705.20     4,491       100
ginitrab   0.52     0.52       0.07        4,491       100
renda_me  274.29   238.57     164.68       4,491       100
escoa2_p   25.42    11.68      28.52       4,491       100
urbano_p   53.14    51.67      23.26       4,491       100
med1000    0.60     0.44       0.55        4,332      96.46
norte      0.07       0        0.25        4,491       100
centrooe   0.08       0        0.28        4,491       100
sul        0.19       0        0.40        4,491       100
nordeste   0.34       0        0.47        4,491       100
pop15a_f   15.14    14.08      7.96        4,491       100
-----------------------------------------------------------------
```

De fato, como já havíamos observado no histograma da variável `populaca`, os municípios têm, em geral, pequena população. Note, por exemplo, que 50% dos municípios possui população menor que 12.407 habitantes; há, no entanto, municípios com alta população, o que se reflete no desvio padrão. A esperança de vida ao nascer, nossa variável resposta para os exercícios subsequentes, tinha média 63.63 naquele ano. A coluna `prop_validos` também nos ajuda a compreender o percentual de observações daquela coluna que possuem valor *missing*.

Agora vamos rodar as regressões! Para ajustar alguns dos modelos a seguir, vamos implementar uma função de regressão linear *from scratch* em `R`. A função retorna os coeficientes da regressão (i.e. o vetor β), os erros-padrão desses coeficientes, seus p-valores, e algumas estatísticas próprias do modelo, como o R^2, a variância dos resíduos e seu desvio-padrão (`RMSE`):

```
linear_model = function(formula, data = expectativaVida) {

  dependent_variable = as.character(formula[[2]])

  X = model.matrix(formula, data = data)
  y = data[[dependent_variable]]

  # beta estimado: \hat{\beta} = (X^T X)^{-1} X^T y
  beta_hat = solve( t(X) %*% X ) %*% t(X) %*% y

  # y estimado a partir do modelo: \hat{y} = X \hat{\beta}
  y_hat = X %*% beta_hat

  # residuals sao a diferenca entre y 'verdadeiro' e y estimado
  residuals = y - y_hat

  n = nrow(X)
  k = ncol(X)

  # erro quadratico medio
  var_residuals = sum(residuals^2) / (n - k)

  # raiz do erro quadratico medio
  RMSE = sqrt(var_residuals)

  # R-squared, ou proporcao da variancia explicada pelo modelo
  total_variance = var(y)
  R2 = 1 - var_residuals / total_variance

  # sob homocedasticidade (ML.5), var(\beta) = \sigma^2 (X^T X)^{-1}
  # onde a equivalencia de \sigma^2 para a amostra é a variância dos residuos
  varcov_beta_hat = var_residuals * solve( t(X) %*% X )
  se_beta_hat = sqrt(diag(varcov_beta_hat))

  # teste de hipotese para o p-valor
  H0 = 0 # teste contra a H0 de o efeito ser nulo
  t = (beta_hat - H0) / se_beta_hat
  p_valor = 2 * (1 - pnorm(abs(t)))

  coefs <- data.frame(
    Coeficientes = colnames(X),
    Estimativas = round(as.numeric(beta_hat), 4),
    Erros_padrao = round(se_beta_hat, 4),
```

```
    t = t,
    p_valor = p_valor,
    row.names = NULL,
    check.names = FALSE
  )

  list(
    coefs = coefs,
    R2 = R2,
    RMSE = RMSE,
    var_residuals = var_residuals
  )
}
```

A primeira regressão ajustada inclui apenas a variável de desigualdade de renda como variável explicativa (`genitrab`) da esperança de vida ao nascer. Vejamos:

```
print(linear_model(formula = esperanc ~ ginitrab))
```

```
$coefs
  Coeficientes Estimativas Erros_padrao          t p_valor
1  (Intercept)     58.4775       0.5485 106.606577       0
2     ginitrab      9.9553       1.0509   9.473201       0

$R2
[1] 0.01938121

$RMSE
[1] 5.030641

$var_residuals
[1] 25.30735
```

Vale a pena testar se nossos resultados são idênticos aos obtidos quando usamos a função nativa do `R` para ajustar modelos lineares, `lm()`:

```
summary(
  lm(
    formula = esperanc ~ ginitrab,
    data = expectativaVida
    )
  )
```

```
Call:
lm(formula = esperanc ~ ginitrab, data = expectativaVida)
```

```
Residuals:
     Min       1Q   Median       3Q      Max
-13.9188  -3.8036   0.3425   3.9275  11.8394

Coefficients:
            Estimate Std. Error t value            Pr(>|t|)
(Intercept)  58.4775     0.5485 106.607 <0.0000000000000002 ***
ginitrab      9.9553     1.0509   9.473 <0.0000000000000002 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.031 on 4489 degrees of freedom
Multiple R-squared:  0.0196,    Adjusted R-squared:  0.01938
F-statistic: 89.74 on 1 and 4489 DF,  p-value: < 0.00000000000000022
```

São idênticos, felizmente! Podemos interpretar nossos resultados com tranquilidade.

O intercepto é o valor da variável dependente quando as variáveis independentes são zero. Nesse caso, portanto, dizemos que, quando `ginitrab` é zero – isto é, não há desigualdade de renda do trabalho –, a expectativa de vida ao nascer é 58.5, em média. β_1, o coeficiente para `ginitrab`, indica que para o aumento em uma unidade de `ginitrab`, a expectativa de vida cresce em 9.95 anos. Esse resultado deveria ser estranho a qualquer cientista: estamos dizendo, em essência, que o aumento da desigualdade leva a um aumento na expectativa de vida. Trata-se, muito provavelmente, de um viés de variável omitida, mas voltaremos a essa questão oportunamente.

Os erros-padrão dos coeficientes são pequenos, o que nos leva a p-valores baixos, muito próximos de zero. Os efeitos, portanto, são estatisticamente significativos. O R-squared é baixo, então a variância explicada pelo modelo é pequena – ou seja, `ginitrab` sozinho explica pouco da variabilidade da variável resposta. Além disso, o erro médio das previsões é de 5 anos, conforme o valor do RMSE.

Agora, façamos uma pequena modificação na variável explicativa, multiplicando-a por 100.

```r
expectativaVida = expectativaVida %>%
  mutate(
    ginitrab100 = ginitrab * 100
  )

reg1 = lm(formula = esperanc ~ ginitrab100, data = expectativaVida)

summary(reg1)
```

```
Call:
lm(formula = esperanc ~ ginitrab100, data = expectativaVida)

Residuals:
     Min       1Q   Median       3Q      Max
```

```
-13.9188  -3.8036   0.3425   3.9275  11.8394

Coefficients:
            Estimate Std. Error t value            Pr(>|t|)
(Intercept) 58.47755    0.54854 106.607 <0.0000000000000002 ***
ginitrab100  0.09955    0.01051   9.473 <0.0000000000000002 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.031 on 4489 degrees of freedom
Multiple R-squared:  0.0196,    Adjusted R-squared:  0.01938
F-statistic: 89.74 on 1 and 4489 DF,  p-value: < 0.00000000000000022
```

A multiplicação não muda nada do ponto de vista estatístico, embora possa facilitar a interpretação dos resultados em certos casos. No caso em que não fizemos a multiplicação por 100, dizíamos que, aumentando `ginitrab` em uma unidade, aumentávamos a expectativa de vida em mais de 9 anos. Mas como `ginitrab` era uma variável contínua entre 0 e 1, estávamos indo de um cenário extremo para outro (de 0 para 1). Quando multiplicamos por 100, podemos dizer que o aumento em uma unidade do Índice de Gini leva a um incremento da expectativa de vida em 0.099 ano. Isso ainda não faz sentido do ponto de vista qualitativo, é claro, mas deixa a interpretação do coeficiente mais informativa.

Agora vamos lidar com o problema do viés de variável omitida. Em particular, vamos incluir a renda domiciliar per capita média como variável explicativa e analisar os coeficientes. Primeiro, vamos ajustar um modelo usando a função que criamos *from scratch*:

```r
reg2_mine = linear_model(formula = esperanc ~ ginitrab100 + renda_me)
print(reg2_mine)
```

```
$coefs
  Coeficientes Estimativas Erros_padrao          t              p_valor
1  (Intercept)     60.5965       0.4036 150.152844 0.000000000000000000
2  ginitrab100     -0.0574       0.0081  -7.081885 0.000000000001421974
3     renda_me      0.0219       0.0004  62.168859 0.000000000000000000

$R2
[1] 0.4730019

$RMSE
[1] 3.687888

$var_residuals
[1] 13.60052
```

Vamos checar se o resultado é o mesmo quando usamos `lm()`:

```r
reg2 = lm(formula = esperanc ~ ginitrab100 + renda_me, data = expectativaVida)
summary(reg2)
```

```
Call:
lm(formula = esperanc ~ ginitrab100 + renda_me, data = expectativaVida)

Residuals:
     Min       1Q   Median       3Q      Max
-26.1536  -2.5072   0.1249   2.6296  10.0669

Coefficients:
              Estimate Std. Error t value             Pr(>|t|)
(Intercept) 60.5965070  0.4035655 150.153 < 0.0000000000000002 ***
ginitrab100 -0.0574135  0.0081071  -7.082      0.00000000000164 ***
renda_me     0.0218646  0.0003517  62.169 < 0.0000000000000002 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.688 on 4488 degrees of freedom
Multiple R-squared:  0.4732,    Adjusted R-squared:  0.473
F-statistic:  2016 on 2 and 4488 DF,  p-value: < 0.00000000000000022
```

Tudo certo! Vejamos os resultados, então. Agora, finalmente, temos um coeficiente para `ginitrab100` que faz mais sentido (dessa vez, negativo). Quando aumentamos o Índice de Gini em uma unidade e mantemos o resto fixo, a expectativa de vida ao nascer decresce, em média, em 0.05 ano. Além disso, para uma unidade que incrementamos na renda per capita média do município – mantendo o resto fixo –, a expectativa de vida ao nascer cresce em 0.02 ano. Todos os coeficientes são estatisticamente significativos e possuem baixo erro-padrão. Além disso, o R-squared é substantivamente melhor do que o observado no primeiro modelo.

A interpretação do coeficiente para `ginitrab100` é diferente porque, antes, não controlávamos por outras variáveis que afetavam tanto a expectativa de vida quanto `ginitrab100`, como a renda per capita média. Nesse caso, havia pelo menos uma variável confundidora, `renda_me`, que enviesava a estimativa do beta para `ginitrab100`.

Vamos incluir mais variáveis no modelo para verificar como os resultados se comportam. Dessa vez, vamos incluir `log(populaca)`:

```r
reg3 = lm(
  formula = esperanc ~ ginitrab100 + renda_me + log(populaca),
  data = expectativaVida
  )

summary(reg3)
```

```
Call:
lm(formula = esperanc ~ ginitrab100 + renda_me + log(populaca),
    data = expectativaVida)

Residuals:
     Min       1Q   Median       3Q      Max
-29.5079  -2.2982   0.1263   2.5502  10.2874

Coefficients:
                Estimate Std. Error t value             Pr(>|t|)
(Intercept)   68.9810535  0.5866102 117.593 < 0.0000000000000002 ***
ginitrab100   -0.0447944  0.0078263  -5.724          0.0000000111 ***
renda_me       0.0235496  0.0003497  67.351 < 0.0000000000000002 ***
log(populaca) -0.9984062  0.0523691 -19.065 < 0.0000000000000002 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.547 on 4487 degrees of freedom
Multiple R-squared:  0.5127,    Adjusted R-squared:  0.5124
F-statistic:  1574 on 3 and 4487 DF,  p-value: < 0.00000000000000022
```

Ao incluir `log(populaca)`, as estimativas dos betas para `ginitrab100` e `renda_me` não se alteraram grosseiramente. Os coeficientes são todos estatisticamente significativos, e tivemos um ganho não-ignorável no R-squared. Ao incluir a variável da população na escala logarítmica, estamos comprimindo a escala dessa variável para evitar que o modelo seja muito sensível à alta variância dessa covaríavel. Vimos entre as estatísticas descritivas, por exemplo, que o desvio-padrão da população dos municípios é altíssima.

Para interpretar o coeficiente de `log(populaca)`, precisamos levar em consideração algumas propriedades do logaritmo. Em particular, devemos usar a seguinte aproximação, conforme demonstrado em…

$$\log(X_{\text{final}}) - \log(X_{\text{inicial}}) = \log\left(\frac{X_{\text{final}}}{X_{\text{inicial}}}\right) \approx \log\left(\frac{X_{\text{final}} - X_{\text{inicial}}}{X_{\text{inicial}}}\right) = \log\left(\frac{\Delta X}{X_{\text{inicial}}}\right),$$

presumindo que ΔX é suficientemente pequeno. Dito de outro modo, podemos interpretar o coeficiente da variável na escala log em termos de variação percentual. Portanto, a cada uma unidade de `log(populaca)` que aumentamos a população em 1%, controlando pelas demais variáveis, diminuímos a expectativa de vida ao nascer em aproximadamente 0.0099 ano, em média.

Por fim, vamos ajustar um modelo com todas as covariáveis disponíveis:

```r
formula_ = esperanc ~ ginitrab100 + renda_me + log(populaca) + escoa2_p +
  urbano_p + med1000 + pop15a_f + norte + centrooe + sul + nordeste

reg4 = lm(
  formula = formula_,
  data = expectativaVida
  )

summary(reg4)
```

```
Call:
lm(formula = formula_, data = expectativaVida)

Residuals:
    Min      1Q  Median      3Q     Max
-9.1851 -1.9687  0.1669  2.1229  9.1558

Coefficients:
                Estimate Std. Error t value             Pr(>|t|)
(Intercept)   67.6369495  0.5171727 130.782 < 0.0000000000000002 ***
ginitrab100   -0.0294642  0.0067134  -4.389  0.00001166656782484 ***
renda_me       0.0081743  0.0005242  15.595 < 0.0000000000000002 ***
log(populaca) -0.4789586  0.0505342  -9.478 < 0.0000000000000002 ***
escoa2_p       0.0212725  0.0025488   8.346 < 0.0000000000000002 ***
urbano_p      -0.0239271  0.0029920  -7.997  0.00000000000000162 ***
med1000       -0.1784613  0.0947582  -1.883               0.0597 .
pop15a_f       0.1467231  0.0115450  12.709 < 0.0000000000000002 ***
norte         -2.1534272  0.2196830  -9.802 < 0.0000000000000002 ***
centrooe      -1.2653372  0.1969555  -6.424  0.00000000014662483 ***
sul            1.1204752  0.1410866   7.942  0.00000000000000252 ***
nordeste      -4.5303204  0.1503863 -30.125 < 0.0000000000000002 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.879 on 4320 degrees of freedom
  (159 observations deleted due to missingness)
Multiple R-squared:  0.6785,    Adjusted R-squared:  0.6777
F-statistic:   829 on 11 and 4320 DF,  p-value: < 0.00000000000000022
```

Os ganhos são claros: temos um R-squared de 0.68, e praticamente todos os coeficientes são estatisticamente significativos, com exceção de `med1000`. O coeficiente da renda per capita média do município perdeu magnitude, um resultado provavelmente resultante da correlação entre essa variável e as demais variáveis do modelo. Quando as demais variáveis não eram incluídas, `renda_me` assumia parte desses efeitos. Algo semelhante ocorreu, inclusive, com a variável `log(populaca)`.

Observe também as variáveis categóricas do modelo. Note, em particular, que não foi incluída a variável `sudeste`, que de fato sequer existia no banco de dados. Como discutimos na parte conceitual desta lista, a inclusão da variável `sudeste` introduziria colinearidade perfeita no modelo e impediria o cálculo dos coeficientes. A consequência, aqui, é que Sudeste é a categoria de base. Do ponto de vista interpretativo, observamos que o fato de o município fazer parte da região Sul está, em média, associado a maiores expectativas de vida ao nascer do que no Sudeste; observamos o contrário, no entanto, para as regiões Centro-Oeste, Norte e sobretudo Nordeste.

A tabela abaixo sistematiza os resultados de todos os modelos ajustados em uma única tabela para facilitar a visualização dos resultados. Outra alternativa seria apresentar os coeficientes graficamente.

```r
stargazer(reg1, reg2, reg3, reg4,
          type = "text",
          title = "Modelos de Regressão Linear ajustados",
          dep.var.labels = "Esperança de Vida",
          column.labels = c("Modelo 1", "Modelo 2", "Modelo 3", "Modelo 4"),
          digits = 3,
          omit.stat = c("f", "adj.rsq", "ll", "aic"))
```

```
Modelos de Regressão Linear ajustados
================================================================================================
                                            Dependent variable:
                        ----------------------------------------------------------------------
                                                Esperança de Vida
                            Modelo 1          Modelo 2          Modelo 3          Modelo 4
                               (1)               (2)               (3)               (4)
------------------------------------------------------------------------------------------------
ginitrab100                 0.100***          -0.057***         -0.045***         -0.029***
                             (0.011)           (0.008)           (0.008)           (0.007)

renda_me                                      0.022***          0.024***          0.008***
                                              (0.0004)          (0.0003)           (0.001)

log(populaca)                                                   -0.998***         -0.479***
                                                                 (0.052)           (0.051)

escoa2_p                                                                          0.021***
                                                                                   (0.003)

urbano_p                                                                          -0.024***
                                                                                   (0.003)

med1000                                                                            -0.178*
                                                                                   (0.095)

pop15a_f                                                                          0.147***
                                                                                   (0.012)

norte                                                                             -2.153***
                                                                                   (0.220)

centrooe                                                                          -1.265***
                                                                                   (0.197)

sul                                                                               1.120***
                                                                                   (0.141)
```

nordeste				-4.530***
				(0.150)
Constant	58.478***	60.597***	68.981***	67.637***
	(0.549)	(0.404)	(0.587)	(0.517)
Observations	4,491	4,491	4,491	4,332
R2	0.020	0.473	0.513	0.679
Residual Std. Error	5.031 (df = 4489)	3.688 (df = 4488)	3.547 (df = 4487)	2.879 (df = 4320)
Note:				*p<0.1; **p<0.05; ***p<0.01

Por fim, um último comentário: embora o modelo 4 inclua uma série de covariáveis importantes que nos ajudam a compreender o comportamento da expectativa de vida, ele ainda não tem interpretação causal. De fato, estamos controlando por uma série de variáveis, mas não por todas; certamente há outros fatores que afetam não apenas a variável de interesse, como também as variáveis independentes – isso inclui, por exemplo, questões associadas à política local, repasses financeiros do governo federal e governos estaduais para os municípios, fatores conjunturais e assim por diante. Isso não altera o fato, no entanto, de que o modelo 4 está mais próximo de oferecer uma explicação causal do que o modelo 1.