



REGRESSÃO PARA PESQUISAS SOCIAIS

Aula 2 - Modelagem de dados de natureza contínua

Thiago Cordeiro Almeida

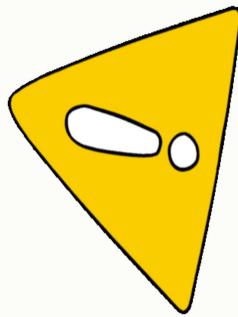
Doutorando, Centre d'Estudis Demogràfics (CED, Espanha)
Pesquisador Assistente, Cebrap

September 17, 2025



ANTES DE COMEÇAR....

lista de presença!





ANTES DE COMEÇAR (2)...

- Dúvidas gerais sobre a aula anterior?
- Exercício da aula anterior: dúvidas, comentários, considerações?





ESTRUTURA DA AULA

Tópicos que vamos cobrir hoje são:

- Fluxo de análise de dados contínuos com o uso de modelagem
- Conceitos, noções (e jargões) econométricos
- Exemplo empírico do dia
- Regressão linear simples
 - Intuição e formalização
 - Propriedade e pressupostos
 - Mão na massa!
- Regressão linear multivariada
 - Intuição e formalização
 - Novidades em relação à regressão linear simples
 - Mão na massa!: interpretação

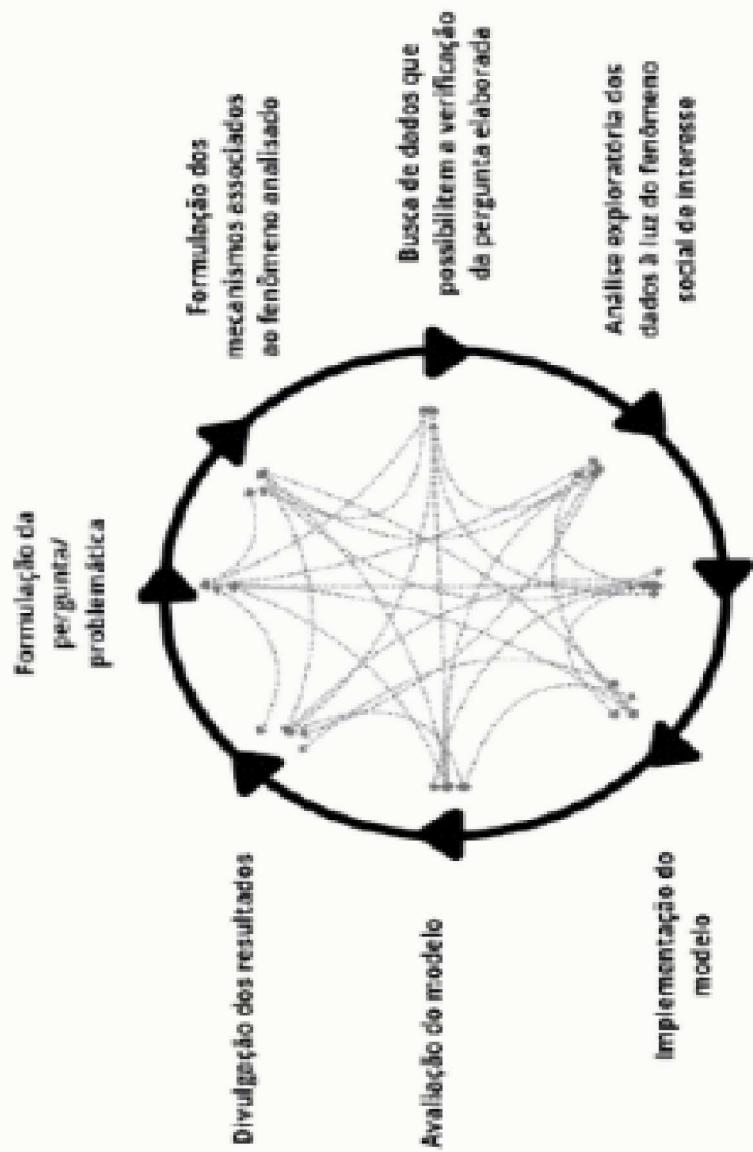


FLUXO DE ANÁLISE DE DADOS CONTÍNUOS

A partir do uso de modelos

FLUXO DE ANÁLISE DE DADOS

Fluxo 'pseudo'-hipotético-dedutivo de modelagem de dados



Desenho de um modelo populacional para estabelecimento das relações entre fenômenos
cebrap.lab - Introdução à análise de regressão para pesquisas sociais





CONCEITOS E DEFINIÇÕES ESTATÍSTICAS IMPORTANTES

Modelos probabilísticos

MODELOS PROBABILÍSTICOS

- Na aula anterior, exploramos diferentes formas de analisar descritivamente a distribuição de determinado fenômeno através de diferentes tipos de medida.
- Em termos estatísticos, para cada distribuição observada em uma amostra, há (*ou haveria*) um modelo teórico que descreveria estatisticamente essa mesma distribuição.





MODELOS PROBABILÍSTICOS

Exemplo: programa pé-de-meia.

Estamos interessados em saber a parcela dos estudantes que tiveram nota em matemática acima da média no ENEM dentro do universo de pessoas que participaram do programa.

ABORDAGEM EMPÍRICA

1. Calcularia a proporção de pessoas que tiveram a média acima como:

$$p(X) = \frac{e}{n}$$

1. Estabeleceria o evento de interesse e o universo amostral de análise

- **Evento:** proporção de estudantes que participaram do ENEM e do programa os quais tiveram nota acima da média em matemática.
 - **Universo amostral:** estudantes que participaram do ENEM e do programa.
2. Estabeleceria o modelo probabilístico que se associa ao evento de interesse
3. Estimaria a probabilidade desse evento de interesse ocorrer.
- e : evento, no caso, num. acima da média;
 - n : observações, no caso, num. estudantes no ENEM.

Em que,



MODELOS PROBABILÍSTICOS

Sob o pressuposto de que nossa amostra de estudantes que participaram do ENEM e do programa é i.i.d.¹, o que obteríamos calculando via abordagem empírica seria exatamente o mesmo ao calcular via abordagem probabilística!!!

Ou seja,

- Podemos estudar propriedades estatísticas dos fenômenos sociais em que trabalhamos não só restritos a uma determinada amostra em que temos, mas fazendo *generalizações* desses fenômenos;
- Utilizando modelos que representem esses fenômenos, conseguimos ter estimativas do comportamento deles que incorporem *incertezas* inerentes a eles.
- Modelagens são formas de adicionar complexidade ao estudo desses fenômenos a partir das suas interações com outros fenômenos sociais.



MODELOS PROBABILÍSTICOS

RELAÇÃO ENTRE MODELOS PROBABILÍSTICOS E MODELOS DE REGRESSÃO

O que fazemos em nossos modelos de regressão é, basicamente, estudar as propriedades da nossa **variável de estudo** através de relações associadas ao melhor modelo probabilístico que descreve essa variável.

Com isso, podemos:

- Prever resultados futuros condicionados a um conjunto de outras variáveis (probabilidade condicional a um conjunto de X)
- Estabelecer em que medida variando uma das variáveis, seus valores são afetados (probabilidade condicional a um X)
- Obter valores esperados e incerteza de seus resultados (esperança e variância)



CONCEITOS E DEFINIÇÕES ESTATÍSTICAS IMPORTANTES

Variável aleatória (v.a.)



VARIÁVEL ALEATÓRIA

Uma variável X é aleatória se cada um de seus valores – está associado a uma probabilidade $p(X)$.

EXEMPLO

Vamos voltar ao pé-de-meia.

Poderíamos dizer que para a variável X definida como ter nota de matemática no ENEM acima da média dos demais participantes do programa, há uma probabilidade $p(X)$, de modo que:

| x | p_x |
|--------------------|-------|
| Abaixo ($x = 0$) | 0.67 |
| Acima ($x = 1$) | 0.33 |



Variáveis aleatórias *v. a.* apresentam algumas propriedades que serão importantes nos modelos que utilizaremos:

ESPERANÇA DE X

Se uma determinada variável é aleatória e o seu seu modelo de probabilidade é conhecido, a sua *Esperança* – também chamada de média – pode ser conhecida.

Geralmente é apresentada enquanto: $E(X)$, sendo X uma variável aleatória.

VARIABILIDADE DE X

Do mesmo modo, se uma determinada variável é aleatória e é conhecido o seu modelo de probabilidade, a sua *variância* – ou sua variabilidade em torno da média – também pode ser conhecido.

Geralmente é apresentada enquanto: $V(X) = E[X - E(X)]^2$, sendo X uma variável aleatória.

RELAÇÃO ENTRE VARIÁVEIS ALEATÓRIAS

Conhecendo-se as propriedades de uma variável aleatória, pode-se relacionar duas ou mais variáveis aleatórias, de modo que:

$$Y = f(X_1, X_2, X_3, X_4, \dots, X_n) = X_1 + X_2 + X_3 + X_4 + \dots + X_n$$

Desse modo,

$$E(Y) = E(X_1 + X_2 + X_3 + X_4 + \dots + X_n)$$

$$E(Y) = E(X_1) + E(X_2) + E(X_3) + E(X_4) + \dots + E(X_n)$$

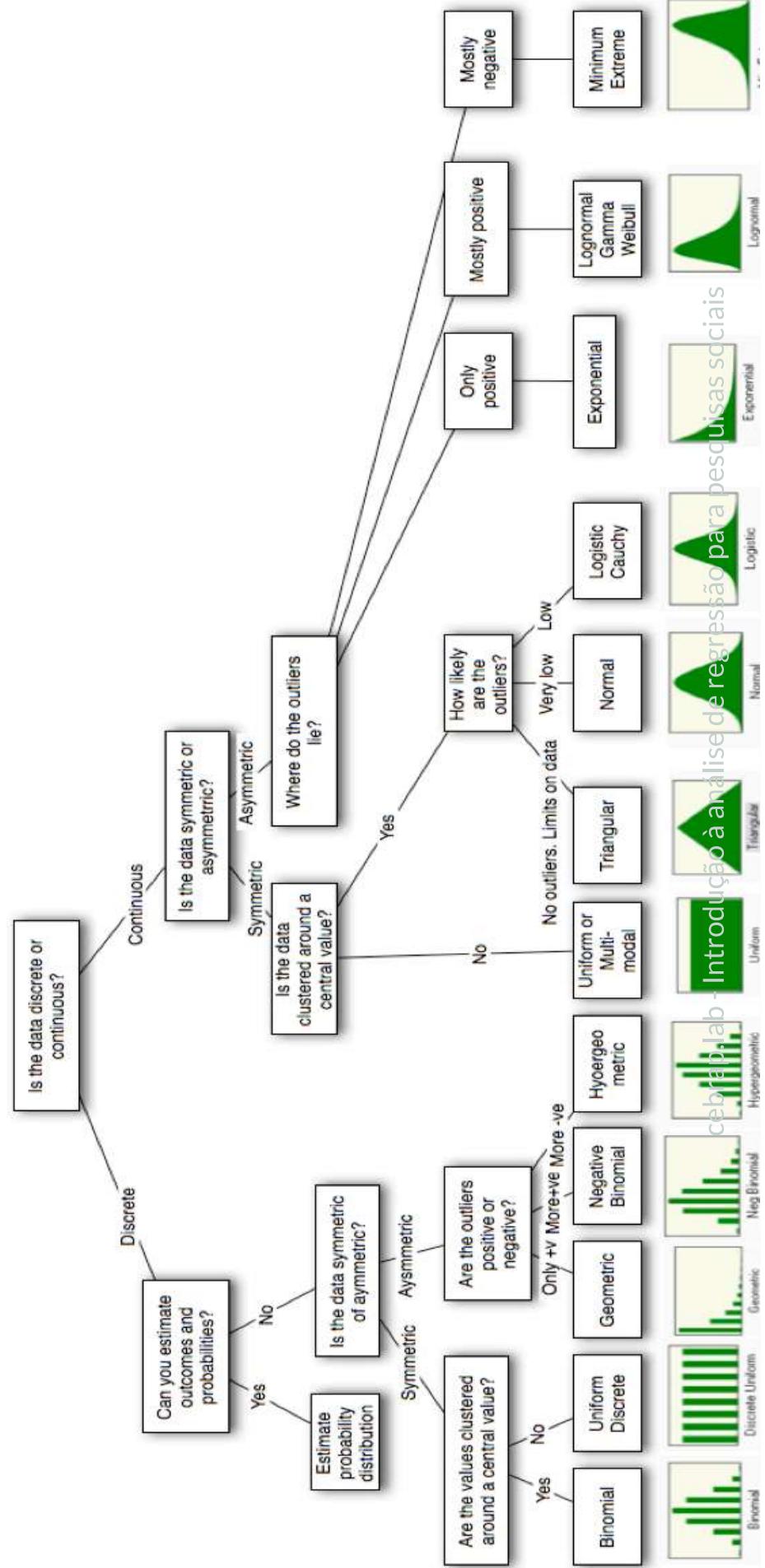
Portanto, tudo o que precisamos saber é qual o modelo probabilístico – $f(\cdot)$ – associado à determinado fenômeno social – que trataremos como uma *v. a.*. A partir disso, podemos relacionar esse fenômeno social de interesse com outros fenômenos – outras *v. a.*.



TIPOS DE VARIÁVEL ALEATÓRIA

As v. a. podem ser divididas entre **discretas** e **contínuas**

Figure 6A.15: Distributional Choices





CONCEITOS E DEFINIÇÕES ESTATÍSTICAS IMPORTANTES

Algumas nomenclaturas



ALGUMAS NOMENCLATURAS

VARIÁVEL A SER ESTUDADA

Y: Variável dependente - variável a ser medida - variável a ser testada - variável resposta - variável resultado

VARIÁVEIS QUE EXPLICAM A VARIÁVEL DE ESTUDO

X: Variável independente - variável explicativa - variável preditora

VARIÁVEIS INDEPENDENTES QUE NÃO ESTAMOS INTERESSADOS EM ANALISAR

Há muitos casos em que estamos interessados em uma – ou um conjunto de – variável(is) independente(s). Todavia, por razões das propriedades dos modelos, é importante que insiramos outras variáveis na análise. Estas são chamadas de **variáveis controle/confundidoras**.

- Em geral, seus resultados não são explicados, mas são explicitados.



MODELO DE REGRESSÃO LINEAR

Modelo de Regressão Linear Simples (RLS)



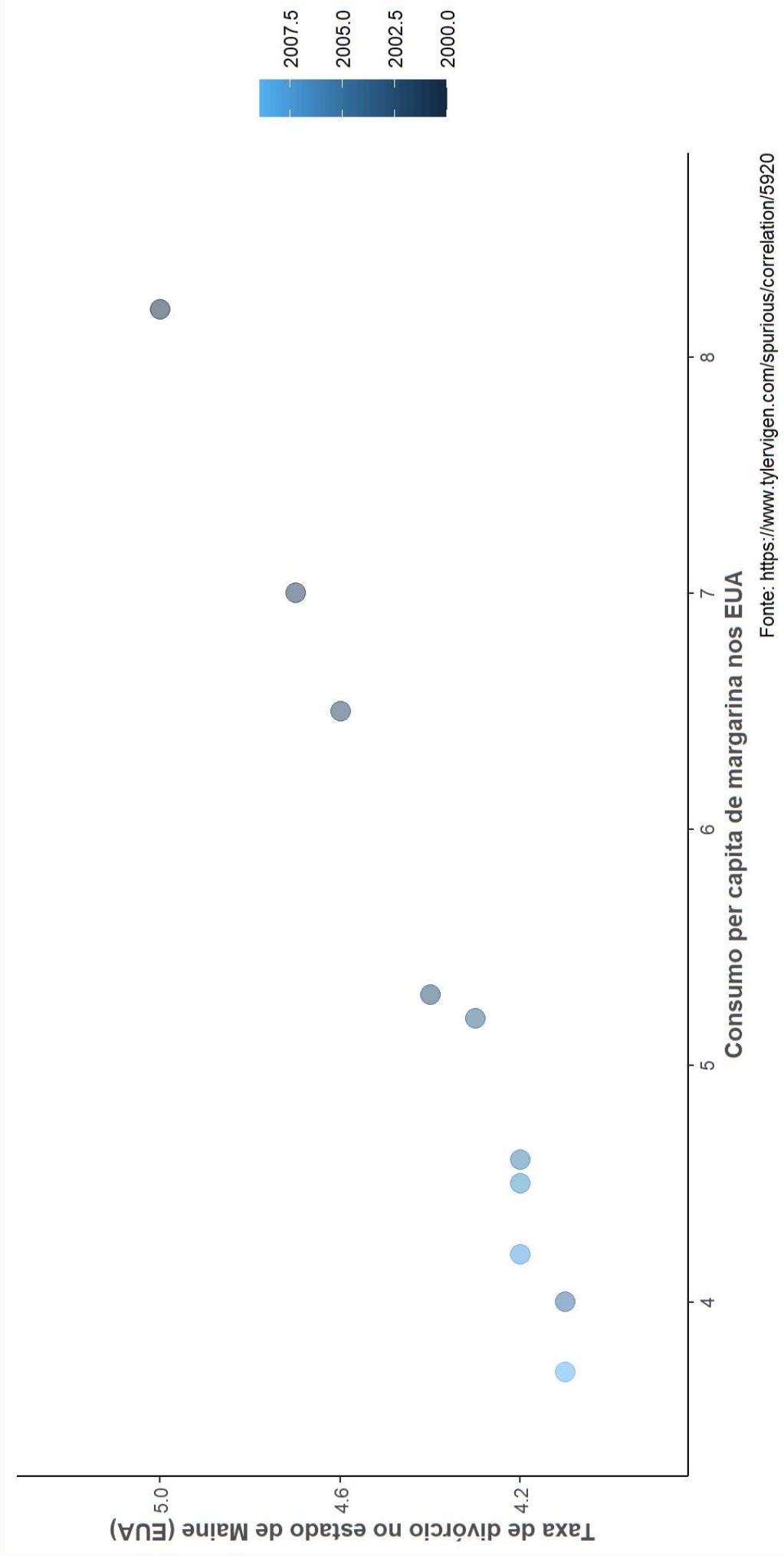
MODELO DE REGRESSÃO LINEAR SIMPLES (RLS)

Nos modelos de regressão em geral, estamos interessados em explicar como a **variação de uma determinada variável afeta a variação de uma outra variável (variável resposta)**.

No caso simplificado do modelo RLS, temos as seguintes características:

- Uma variável aleatória resposta/dependente de natureza contínua: Y
- Uma variável aleatória explicativa: X
- Um fator de erro do modelo que nos diz quanto da variabilidade de Y que continua não explicado por X : ϵ

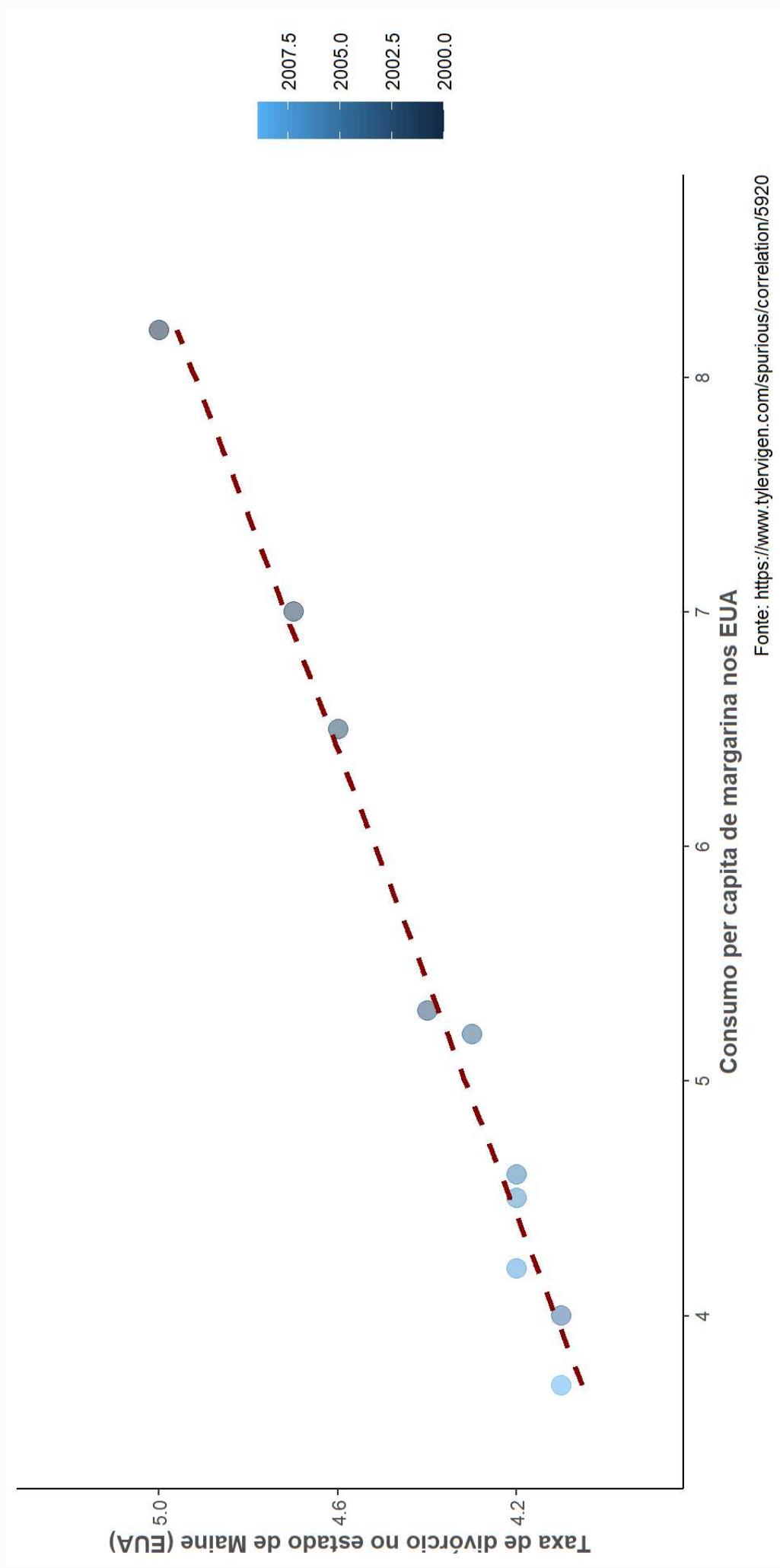
MODELO DE REGRESSÃO LINEAR SIMPLES (RLS)



Fonte: <https://www.tyverigen.com/spurious/correlation/5920>



MODELO DE REGRESSÃO LINEAR SIMPLES (RLS)



Fonte: <https://www.tyverigen.com/spurious/correlation/5920>





FORMALIZAÇÃO DE UM MRS

Dizemos que a representação geral (ou populacional) de um determinado modelo de regressão é dada por:

$$Y_i = \beta_0 + \beta_1 \cdot X_i + \epsilon_i$$

Em que:



HIPÓTESE: INDEPENDÊNCIA ENTRE ERRO E VARIÁVEIS OBSERVADAS

Partindo do nosso modelo RLS para a população,

$$Y_i = \beta_0 + \beta_1 \cdot X_i + \epsilon_i$$

A **hipótese-chave** é que a média dos erros condicionada na variável explicativa X_j , é igual a zero, ou seja:

$$E[\epsilon_i | X_i] = 0$$

Em outras palavras, estamos assumindo que não há nenhum fator contido em ϵ que seja correlacionado com X , de modo que β seja um efeito puro de X .



Modelo populacional:

$$Y_i = \beta_0 + \beta_1 \cdot X_i + \epsilon_i$$

Quando vamos *estimar* os nossos parâmetros para a nossa amostra, assumimos os valores médios para os nossos i casos observados, isto é:

$$\bar{Y} = \beta_0 + \beta_1 \cdot \bar{X} + \epsilon_i$$

Os únicos valores que não conhecemos são referentes a β_0 e β , uma vez que ϵ é obtido residualmente.

Tudo o que precisamos fazer é **estimar os parâmetros desconhecidos do modelo populacional**.



ESTIMANDO PARÂMETROS DO MODELO

O modelo a ser estimado consiste em:

$$\bar{Y} = \beta_0 + \beta_1 \cdot \bar{X} + \epsilon_i$$

ESTIMADORES DE MÍNIMOS QUADRADOS ORDINÁRIOS (MQO)

Os estimadores de MQO asseguram, via relações estatísticas, que tenhamos valores para nossos parâmetros desconhecidos que **minimizem o desvio dos resíduos**.

Isto é, temos a reta que melhor se ajusta aos nossos dados de modo que passe mais perto possível de cada ponto.

ESTIMANDO PARÂMETROS DO MODELO

ESTIMADORES DE MÍNIMOS QUADRADOS ORDINÁRIOS (MQO)

Acionando a hipótese descrita anteriormente, temos:

- $\hat{\beta}_0 = \bar{Y} - [\hat{\beta}_1 \cdot \bar{X}]$
- $\hat{\beta}_1 = \frac{Cov(x, y)}{Var(x)}$
- Valores preditos: $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 \cdot \bar{X}_i$ para $i = 1, \dots, n.$
- Resíduos: $u_i = y_i - \hat{Y}_i$, para $i = 1, \dots, n$



Para o exemplo anterior, se partimos do modelo a ser estimado:

$$\text{Divorcio} = \beta_0 + \beta_1 \cdot \text{Margarina} + \epsilon$$

Aplicando em R:

```
1 modelo <- lm(divorcio ~ margarina, data = base)
2 modelo
```

Call:
`lm(formula = divorcio ~ margarina, data = base)`

Coefficients:

| | |
|-------------|-----------|
| (Intercept) | margarina |
| 3.3086 | 0.2014 |



Para o exemplo anterior, se partimos do modelo a ser estimado:

— ^ ^ —

$$\hat{Divorcio} = \beta_0 + \beta_1 \cdot Margarina + \epsilon$$

Aplicando em R:

```
1 modelo <- lm(divorcio ~ margarina, data = base)
2 summary(modelo) # indo além nos resultados
```

Call:

`lm(formula = divorcio ~ margarina, data = base)`

Residuals:

| | Min | 1Q | Median | 3Q | Max |
|-----------|----------|----------|----------|---------|---------|
| margarina | -0.05583 | -0.01816 | -0.01452 | 0.03601 | 0.04625 |

Coefficients:

| | Estimate | Std. Error | t value | Pr(> t) |
|-------------|----------|------------|---------|---------------------|
| (Intercept) | 3.308626 | 0.048032 | 68.88 | 0.0000000000022 *** |
| margarina | 0.201386 | 0.008735 | 23.05 | 0.0000000132968 *** |

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.03841 on 8 degrees of freedom
 Multiple R-squared: 0.9852, Adjusted R-squared: 0.9833
 F-statistic: 531.5 on 1 and 8 DF, p-value: 0.0000000133



PAUSA!







MODELO DE REGRESSÃO LINEAR

Propriedades do MRL

PROPRIEDADES DO MRL

Para assegurar que, através do MQO, tenhamos estimadores **não viésados e mais eficientes**, devemos assegurar que 6 hipóteses sejam asseguradas:

1. O modelo é linear nos parâmetros;
2. Amostra é aleatória e segue o modelo populacional descrito em H1;
3. Inexistência variáveis constantes ou que não tenham uma relação linear exata com outra variável explicativa;
4. Independência entre erro e variáveis observadas;
5. Variância do erro constante para todas as variáveis explicativas;
6. O erro segue uma distribuição normal com média 0 e variância constante



H1: O MODELO É LINEAR NOS PARÂMETROS

Os parâmetros – β_S – se relacionam de forma linear.

$$Y_i = \beta_0 + \beta_1 \cdot X_i + \epsilon_i$$

É por essa propriedade que chamamos os modelos de **modelos lineares**.



H2: AMOSTRA É ALEATÓRIA E SEGUÉ O MODELO POPULACIONAL DESCrito EM H1

Para uma dada amostra de tamanho N qualquer, seu processo gerador seguiu uma distribuição aleatória, seguindo o modelo populacional descrito em H1.

Todavia, *em casos onde temos problemas de seletividade amostral*, isso pode afetar essa hipótese.



H3: INEXISTÊNCIA VARIÁVEIS CONSTANTES OU QUE NÃO TENHAM UMA RELAÇÃO LINEAR EXATA COM OUTRA VARIÁVEL EXPLICATIVA

Para qualquer variável explicativa X , há uma variabilidade de seus resultados em relação a Y .

Além disso, determinada variável X não pode estar totalmente correlacionada com outra variável X .

- Assegura que não haja perfeita colinearidade, i.e., que toda a variação em X_1 esteja associada a X_2 .



H4: INDEPENDÊNCIA ENTRE ERRO E VARIÁVEIS OBSERVADAS

Mencionado anteriormente: a média dos erros condicionada na variável explicativa X_i é igual a zero, ou seja:

$$E[\epsilon_i | X_i] = 0$$

Para um determinado valor de X , os outros fatores não observados no modelo não afetam o resultado de X .

Não é que ϵ não afete Y , e sim que ϵ não está condicionado a X .



H5: VARIÂNCIA DO ERRO CONSTANTE PARA TODAS AS VARIÁVEIS EXPLICATIVAS

$$\text{Var}[\epsilon | X] = \sigma^2$$

Chamamos essa hipótese de **homocedasticidade** ou **variância constante do erro**.

Pode ser violada em situações de:

- **Outliers para uma variável específica**
- **Distribuições de alguma variável específica que seguem um distinto padrão.**
- ...



H6: O ERRO SEGUÉ UMA DISTRIBUIÇÃO NORMAL COM MÉDIA 0 E VARIÂNCIA CONSTANTE

Assumindo isso, podemos também assumir, por *propriedades estatísticas*, que qualquer um de nossos parâmetros também têm distribuição normal.

Por que assumimos essa hipótese para o erro e não para os parâmetros?

- O erro pode ser estudado e avaliado, os parâmetros não.



ASSEGURANDO A EXISTÊNCIA DE H1-H4

Estimadores MQO não viesados. Isto é, representam o que seria o valor a ser obtido na população estudada, mesmo depois de implementar em N diferentes amostras.

\wedge

- $E[\beta_0 | X] = \beta_0$

\wedge

- $E[\beta_1 | X] = \beta_1$



ASSEGURANDO A EXISTÊNCIA DE H1-H4

Estimadores MQO não viesados. Isto é, representam o que seria o valor a ser obtido na população estudada, mesmo depois de implementar em N diferentes amostras.

\wedge

- $E[\beta_0 | X] = \beta_0$

\wedge

- $E[\beta_1 | X] = \beta_1$

ASSEGURANDO A EXISTÊNCIA DE H5

Estimadores MQO sendo os mais eficientes. Isto é, asseguram a menor variância possível dentre outros processos de estimação dos parâmetros.



ASSEGURANDO A EXISTÊNCIA DE H1-H4

Estimadores MQO não viesados. Isto é, representam o que seria o valor a ser obtido na população estudada, mesmo depois de implementar em N diferentes amostras.

\wedge

- $E[\beta_0 | X] = \beta_0$

\wedge

- $E[\beta_1 | X] = \beta_1$

ASSEGURANDO A EXISTÊNCIA DE H5

Estimadores MQO sendo os mais eficientes. Isto é, asseguram a menor variância possível dentre outros processos de estimação dos parâmetros.

ASSEGURANDO A EXISTÊNCIA DE H1-H5

Temos os melhores estimadores para o modelo populacional. À medida que aumenta o tamanho da amostra, aproxima-se mais do valor do parâmetro na população e com menor incerteza sobre ele.



ASSEGURANDO A EXISTÊNCIA DE H1-H4

Estimadores MQO não viesados. Isto é, representam o que seria o valor a ser obtido na população estudada, mesmo depois de implementar em N diferentes amostras.

\wedge

- $E[\beta_0 | X] = \beta_0$

\wedge

- $E[\beta_1 | X] = \beta_1$

ASSEGURANDO A EXISTÊNCIA DE H5

Estimadores MQO sendo os mais eficientes. Isto é, asseguram a menor variância possível dentre outros processos de estimação dos parâmetros.

ASSEGURANDO A EXISTÊNCIA DE H1-H5

Temos os melhores estimadores para o modelo populacional. À medida que aumenta o tamanho da amostra, aproxima-se mais do valor do parâmetro na população e com menor incerteza sobre ele.

ASSEGURANDO A EXISTÊNCIA DE H1-H6



•••

EXEMPLO DA AULA!



ACESSO AO ESGOTAMENTO SANITÁRIO

Pergunta: Quais os fatores associados ao acesso ao esgotamento sanitário de populações vulneráveis no Brasil?

Mais especificamente, qual a associação entre acesso ao Programa Bolsa Família para famílias com jovens entre 15-29 anos e o acesso ao esgotamento sanitário?

DADOS

Uma base de dados construída com dados do CadÚnico e Censo Demográfico 2010.



VAMOS LÁ!

...





MODELO DE REGRESSÃO LINEAR

Modelo de Regressão Linear Múltipla (RLM)



MODELO DE REGRESSÃO LINEAR MÚLTIPLA (RLM)

No caso dos modelos RLS, havia somente uma variável explicativa.

Todavia, pode ser que haja outros fenômenos sociais que estejam:

1. Associados com a nossa variável de interesse Y
2. Associados com a nossa variável explicativa X



NO CENÁRIO 1

Se temos na base de dados a variável explicativa que **sabemos que está associada com Y** , por que não a incluir?

GANHOS DE INCLUIR

PERDAS DE INCLUIR

- Reduz a variância não explicada do modelo – contida em ϵ
- Aumenta a capacidade de predição do modelo
- Parcimônia do modelo: *the simpler, the better!*



NO CENÁRIO 2

Se sabemos que está associada com X , ela deve ser incluída.

GANHOS DE INCLUIR PERDAS DE INCLUIR

- Não violação de H4 – independência entre o erro e as variáveis observadas.
- Neste caso, não há.
- Reduz a variância não explicada do modelo – contida em ϵ

FORMALIZAÇÃO

$$Y_i = \beta_0 + \beta_1 \cdot X_{1i} + \beta_2 \cdot X_{2i} + \dots + \beta_k \cdot X_{ki} + \epsilon_i$$

Em que:

- k é o número de parâmetros associados às variáveis explicativas incluídas no modelo.

O QUE MUDA ENTRE RLS E RLM?

Em RLS, temos o modelo populacional descrito como:

$$Y_i = \beta_0 + \beta_1 \cdot X_{1i} + \epsilon_i$$

Enquanto em RLM, temos:

$$Y_i = \beta_0 + \beta_1 \cdot X_{1i} + \beta_2 \cdot X_{2i} + \dots + \beta_k \cdot X_{ki} + \epsilon_i$$





O QUE MUDA ENTRE RLS E RLM?

Em RLS, temos o modelo populacional descrito como:

$$Y_i = \beta_0 + \beta_1 \cdot X_{1i} + \epsilon_i$$

Enquanto em RLM, temos:

$$Y_i = \beta_0 + \beta_1 \cdot X_{1i} + \beta_2 \cdot X_{2i} + \dots + \beta_k \cdot X_{ki} + \epsilon_i$$

Adicionar mais variáveis ao modelo nos dá maior confiança de que estaremos cumprindo os pressupostos para se ter estimadores não viésados e mais eficientes.

O QUE MUDA ENTRE RLS E RLM?

Em RLS, temos o modelo populacional descrito como:

$$Y_i = \beta_0 + \beta_1 \cdot X_{1i} + \epsilon_i$$

Enquanto em RLM, temos:

$$Y_i = \beta_0 + \beta_1 \cdot X_{1i} + \beta_2 \cdot X_{2i} + \dots + \beta_k \cdot X_{ki} + \epsilon_i$$

Adicionar mais variáveis ao modelo nos dá maior confiança de que estaremos cumprindo os pressupostos para se ter estimadores não viésados e mais eficientes. Aumenta a complexidade dos modelos também...





O QUE MUDA ENTRE RLS E RLM?

INTERPRETAÇÃO DOS MODELOS RLM

Nos modelos RLM, temos que invocar a noção de *ceteris paribus* para uma correta interpretação dos resultados.

CETERIS PARIBUS: “tudo o mais é constante”.

- Mantendo-se todos os outros fatores fixos, constantes, pode-se interpretar como a variação em determinada variável explicativa X_k se relaciona com Y .



EXEMPLO DA AULA!

Parte 2



VAMOS FAZER O NOSSO MODELO UM POUCO MAIS COMPLEXO

Vamos estimar 3 diferentes modelos:

MODELO 1

O mesmo estimado no exercício anterior - já temos código e tudo mais!

MODELO 2

Adicione ao **Modelo 1**, a variável informalidade (*informalidade*)

MODELO 3

Adicione ao **Modelo 1**, variável frequência ao ensino médio (*em*)

MODELO 4

Adicione ao **Modelo 1**, as variáveis que adicionamos no **Modelo 2 e 3**

VAMOS LÁ!

Dica (código em R para modelo com 2 variáveis explicativas):

```
modelo2 = lm(Var_Y ~ Var_X1 + Var_X2, data = dados)
```







ISSO É TUDO PARA HOJE!

Para próxima aula:

1. Exercício sobre Regressão Linear Múltipla (próximo slide)
2. Leitura de material para aula de hoje (caso não tenha lido)
3. Leitura de material para a aula seguinte (Ver ementa)



ORIENTAÇÕES SOBRE O EXERCÍCIO

Utilizaremos os mesmos dados que utilizamos na aula de hoje.

- Rode o Modelo 1 da aula de hoje e interprete os resultados
- Adicione ao modelo 1, a variável do Modelo 2 da aula de hoje e interprete os resultados de cada variável;
- Adicione ao Modelo 2, a variável do Modelo 3 da aula de hoje e interprete os resultados de cada variável;
- Adicione a variável `município_rural` e interprete os resultados dela;
- Volte às estatísticas do modelo – obtidas com `summary()` – e compare se houve muita mudança nos coeficientes de cada uma das variáveis quando incluída outras variáveis ao modelo. **O que isso nos diz?**
- Observe se o R^2 – obtidos com `summary()` – apresentou muita mudança entre os modelos. **O que isso nos diz?**



CEBRAP

Presidência Adrian Gurza Lavalle

Diretoria Administrativa Victor Callil

Diretoria Científica Arilson Favareto

Coordenação de Seminários Bianca Tavolari

Coordenação de Cursos Monise Fernandes Picanço

Curso

Introdução à análise de regressão para pesquisas sociais

Ministrante

Thiago Cordeiro Almeida

E-mail: thiagocordalmeida@gmail.com

Github: [@thiagocalm](https://github.com/thiagocalm)

