Introdução ao Software R e à Análise Econométrica

Junho a Setembro de 2018

Alexandre Xavier Ywata Carvalho
Geraldo Sandoval Góes

Curso de Introdução à Econometria

- **Objetivo:**
  - Transmitir os principais conceitos de:
    - Regressão linear
    - Inferência em modelos de regressão
    - Seleção de variáveis
    - Correlação versus causalidade
    - Regressão com dados binários
    - Modelos com dados de painel
    - Introdução aos métodos de *matching*
- **Metodologia:**
  - Aulas expositivas
  - Discussão de estudos aplicados
  - Introdução ao software livre R
  - Todas os exercícios serão feitos em grupos de 2 ou 3 alunos

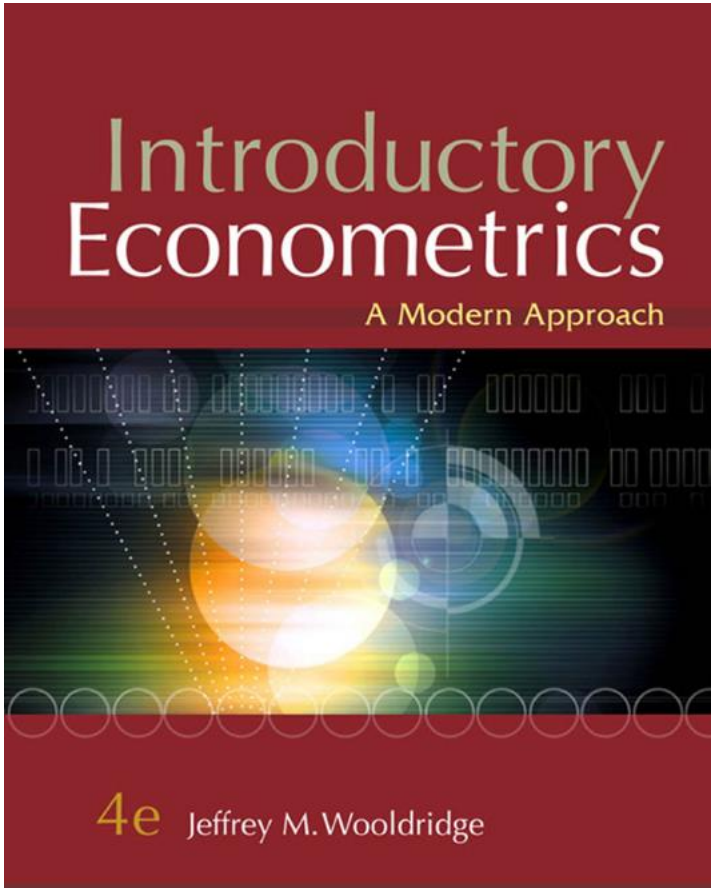
Trade-off nos Cursos de Econometria

Curso que seja útil para a vida prática do profissional

VERSUS

Curso que permita ao profissional se aprofundar no assunto futuramente

Referência básica



Revisão Básica de Medidas de Informação e Análise Gráfica

Tratamento das Informações Disponíveis

- Os primeiros passos para a análise de dados consistem em:
  - Limpeza dos dados
  - Análise exploratória para melhor descrever as informações disponíveis
- Para a limpeza e tratamento de dados, importante ter disponível um software adequado
  - Exemplos: R, Stata, SPSS, SAS, Pentaho Data Integration (PDI)
- Para a análise exploratória, diversas medidas estão disponíveis:
  - Medidas de localização
  - Medidas de dispersão
  - Análises gráficas
  - Medidas de relações
  - Medidas de desigualdade e concentração
- O processo de limpeza de dados e a análise exploratória podem ocorrer simultaneamente
  - Exemplo, a análise exploratória para ajudar a identificar outliers que precisam ser excluídos ou tratados na amostra

Medidas de Informação

- Na análise exploratória de dados, existe um conjunto consolidado de indicadores que são calculados para melhor conhecer os dados disponíveis
    - Medidas de localização
    - Medidas de dispersão
    - Análises gráficas
    - Medias de desigualdade e concentração
    - Medidas de relações
- Essas medidas estão programas em todos os programas estatísticos e em programas como o Excel
- Análise gráficas também são muito utilizadas
    - Histogramas
    - Gráficos de dispersão
    - Box plots
    - Outros ...

Medidas de Informação

- Medidas de Centralidade:
  - Conjunto de dados: $x_1, x_2, x_3, \ldots . \ x_n$
  - Exemplo: dados de renda per capita em domicílios em um determinado município

$$x_1 = 300, x_2 = 150, x_3 = 420, x_4 = 120, x_5 = 315, x_6 = 375 \ (n = 6)$$

  - Soma:

$$S_N = x_1 + x_2 + x_3 + \ \ldots + x_n = \sum_{i=1}^{n} x_i$$

  - Média:

$$\bar{x} = \frac{x_1 + x_2 + x_3 + \ \ldots + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n}$$

  - Mediana:
    - Divide o conjunto de dados (ordenados) em duas “metades”
    - Exemplo: 300, 150, 420, 120, 315, 375
    - Ordenados: 120, 150, 300, 315, 375, 420
    - Valor intermediário: Mediana = (300 + 315)/2 = 307,5
    - Menos sensível a valores extremos do que a média

Medidas de Informação

- Medidas de Centralidade:
    - Moda:
        - Corresponde ao valor mais frequente em um conjunto de dados
        - Exemplo: 300, 150, 300, 420, 120, 375, 315, 375, 300
        - Moda = 300
- No Excel ou no R:
    - Funções no Excel: soma(), média(), med(), modo.único()
    - Obs.: na função, modo.único(), se não houver valor mais frequente que outros, a função retorna #n/d.
    - Exercício rápido: verificar o que acontece com média e mediana quando multiplicamos por 1000 um dos valores nos dados

Medidas de Informação

- Medidas de Dispersão:
  - Conjunto de dados: $x_1, x_2, x_3, \ldots . \ x_n$
  - Exemplo: dados de renda per capita em domicílios em um determinado município

$$x_1 = 300, x_2 = 150, x_3 = 420, x_4 = 120, x_5 = 315, x_6 = 375 \ (n = 6)$$

  - **Variância Populacional**:

$$\sigma^2 = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + (x_3 - \bar{x})^2 + \ \ldots + (x_n - \bar{x})^2}{n} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n}$$

  - A variância não está na escala original dos dados (renda per capita em R$ e variância em R$ ao quadrado, por exemplo)

  - Corrigimos esse problema com o **Desvio-Padrão** (também **Populacional**):

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

Medidas de Informação

- Medidas de Dispersão:
  - **Variância Amostral**:

$$s^2 = \frac{(x_1-\bar{x})^2 + (x_2-\bar{x})^2 + (x_3-\bar{x})^2 + \ \ldots + (x_n-\bar{x})^2}{\boldsymbol{n-1}} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{\boldsymbol{n-1}}$$

  - **Desvio-Padrão Amostral**:

$$s = \sqrt{s^2} = \sqrt{\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{\boldsymbol{n-1}}}$$

  - Exemplo rápido – no Excel ou no R
    - Dados: $x_1 = 300, x_2 = 150, x_3 = 420, x_4 = 120, x_5 = 315, x_6 = 375$ $(n = 6)$
    - Cálculo das variâncias (amostral e populacional) e dos desvios-padrões (amostral e populacional) ...
    - Funções no Excel: var.p(), var.a(), desvpad.a(), desvpad.p()

Medidas de Informação

- **Quartis**:
  - A mediana divide a massa de dados em duas metades
  - Os quartis dividem em quatro partes com igual número (aproximadamente) de observações

    Exemplo: 2.2, 3,1. 1.0, -0.2, 5.2, 3.2, 7.5, -2.4
    Ordenados: -2.4, -0.2, 1.0, 2.2, 3.1, 3.2, 5.2, 7.5
    Primeiro quartil: (-0.2 + 1.0)/2 = 0.4
    Segundo quartil: (2.2 + 3.1)/2 = 2.65 (mesmo que a mediana)
    Terceiro quartil: (3.2 + 5.2)/2 = 4.2
  - Intervalo interquartil => terceiro quartil – primeiro quartil (menos sensível a observações extremas do que o desvio padrão)
  - Portanto, 25% das observações são menores do que o primeiro quartil enquanto 25% são maiores do que o terceiro quartil
- **Decis** - dividem a massa de dados em 10 grupos com igual número de observações
- **Percentis** – dividem a massa de dados em 100 grupos com igual número de observações

Medidas de Informação

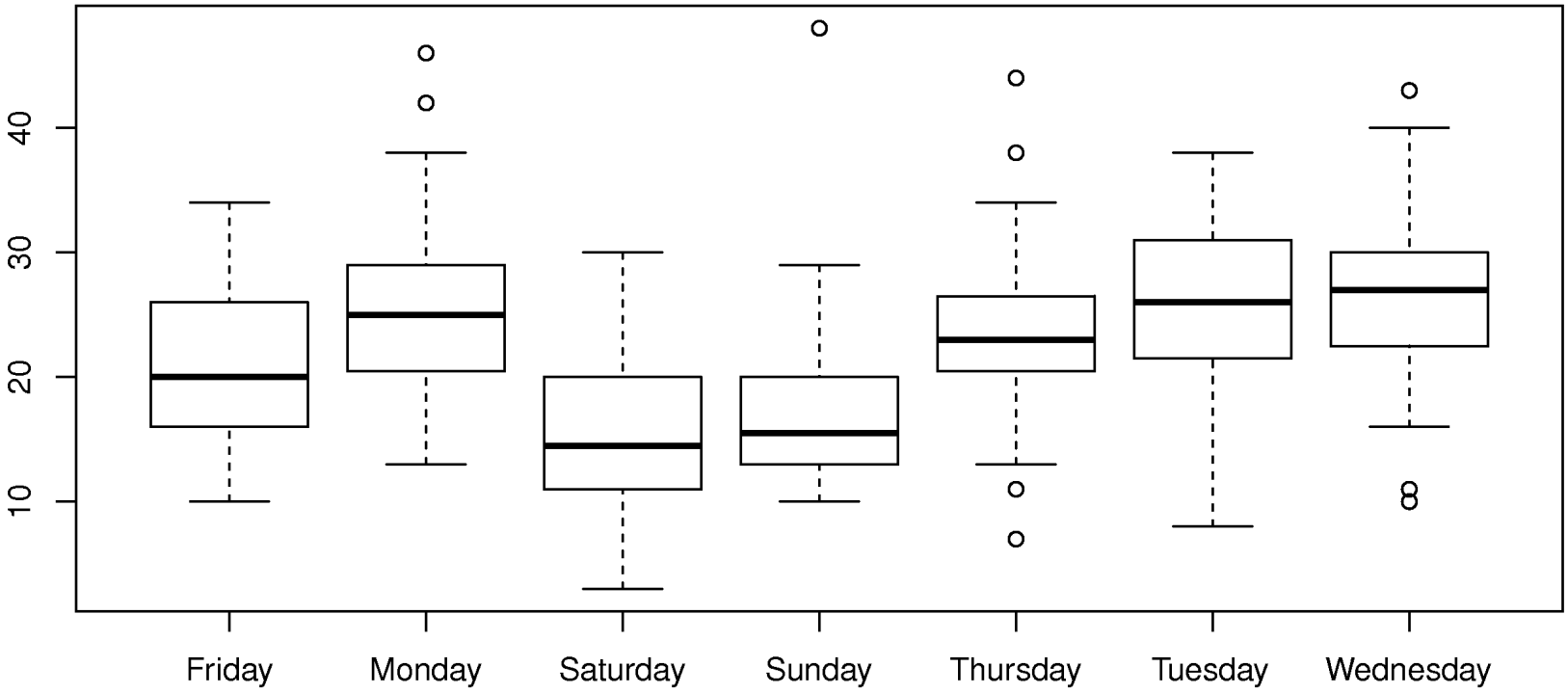
- Exemplos no Excel ou no R:
  - Planilha “IDH_Brasil_2010.xlsx” – contém informações municipais, referentes ao Atlas de Desenvolvimento com base no Censo Demográfico de 2010
  - Vamos analisar algumas das variáveis nessa base obtendo algumas das medidas de informação
  - Calcule as medidas descritivas:
    - Média, mediana, desvio-padrão e variância (amostral e populacional), quartis, percentis 1, 5, 10, 90, 95 e 99%, intervalo inter-quartil

    Para as variáveis:

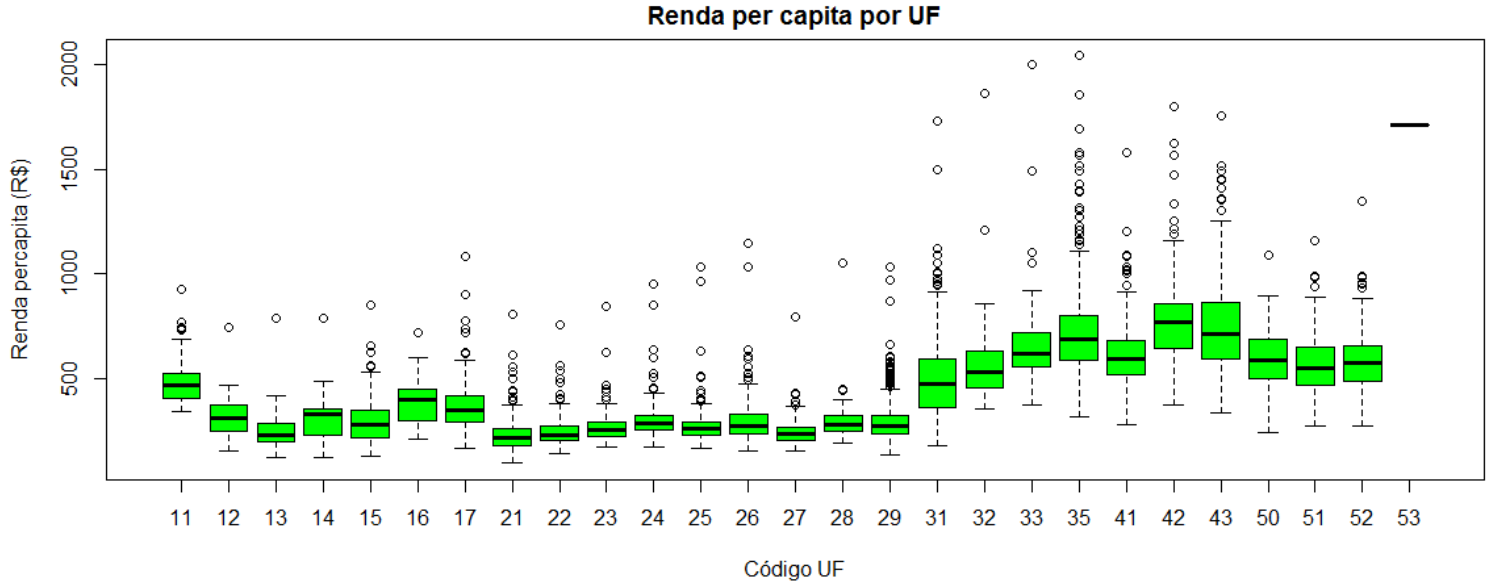
    - esperanca_vida_ao_nascer
    - renda_per_capita
    - IDHM

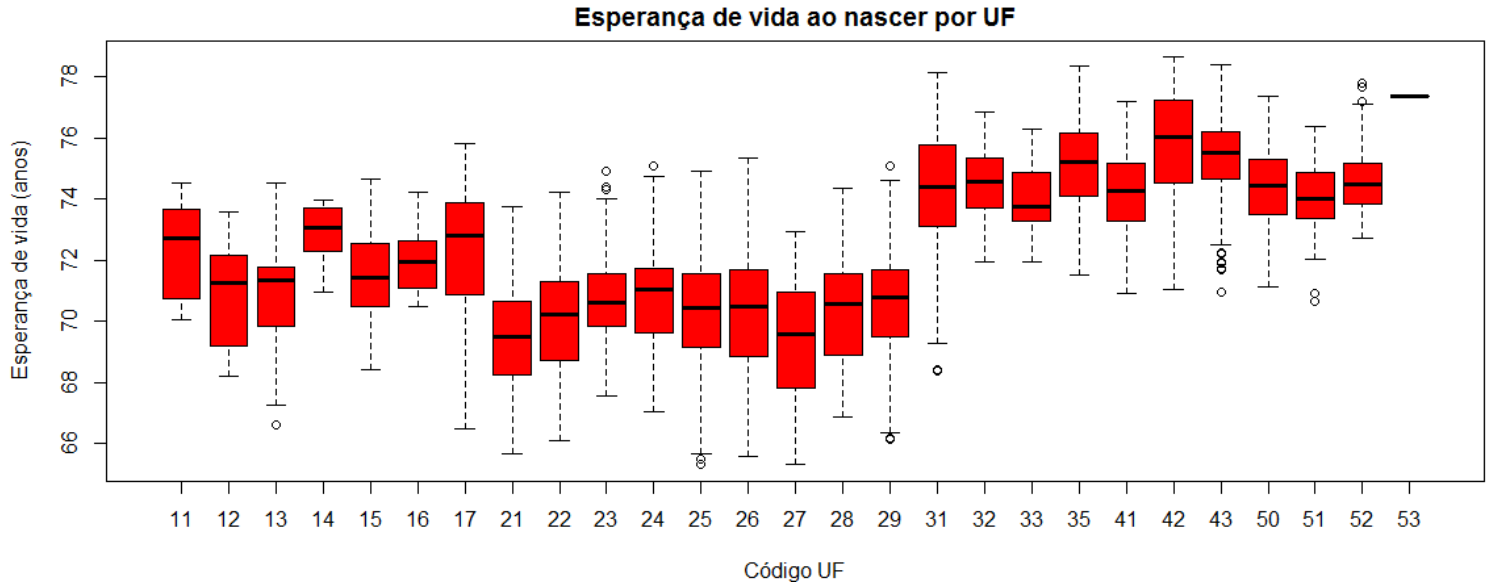
Medidas de Informação – Box Plots

- Apresentando várias medidas de forma integrada – **Box Plots**
  - Comando no R: `boxplot(website$Visits ~ website$DayOfWeek);`



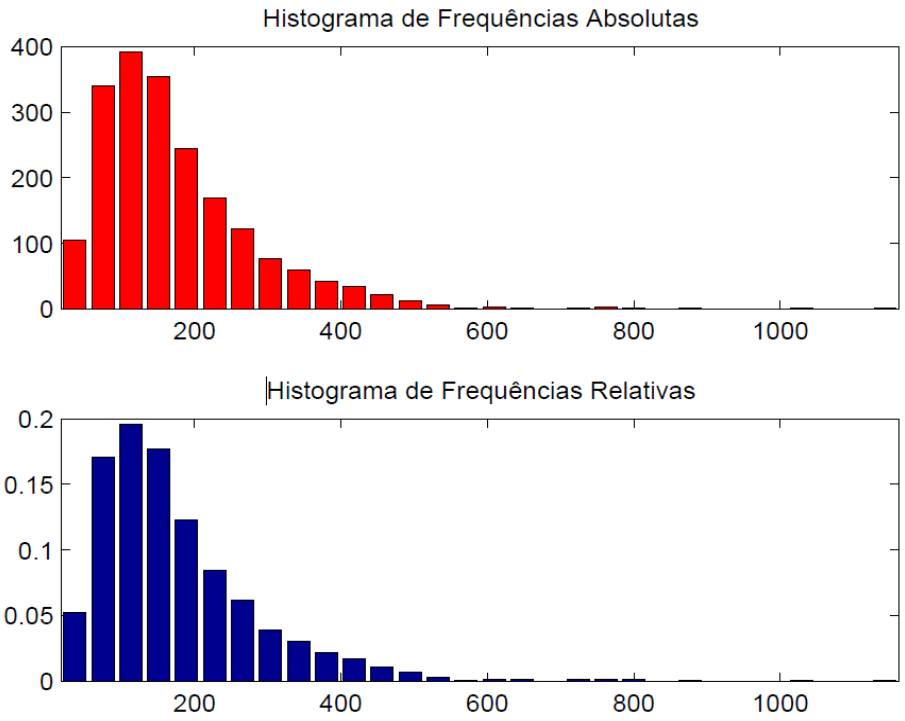
Medidas de Informação – Box Plots





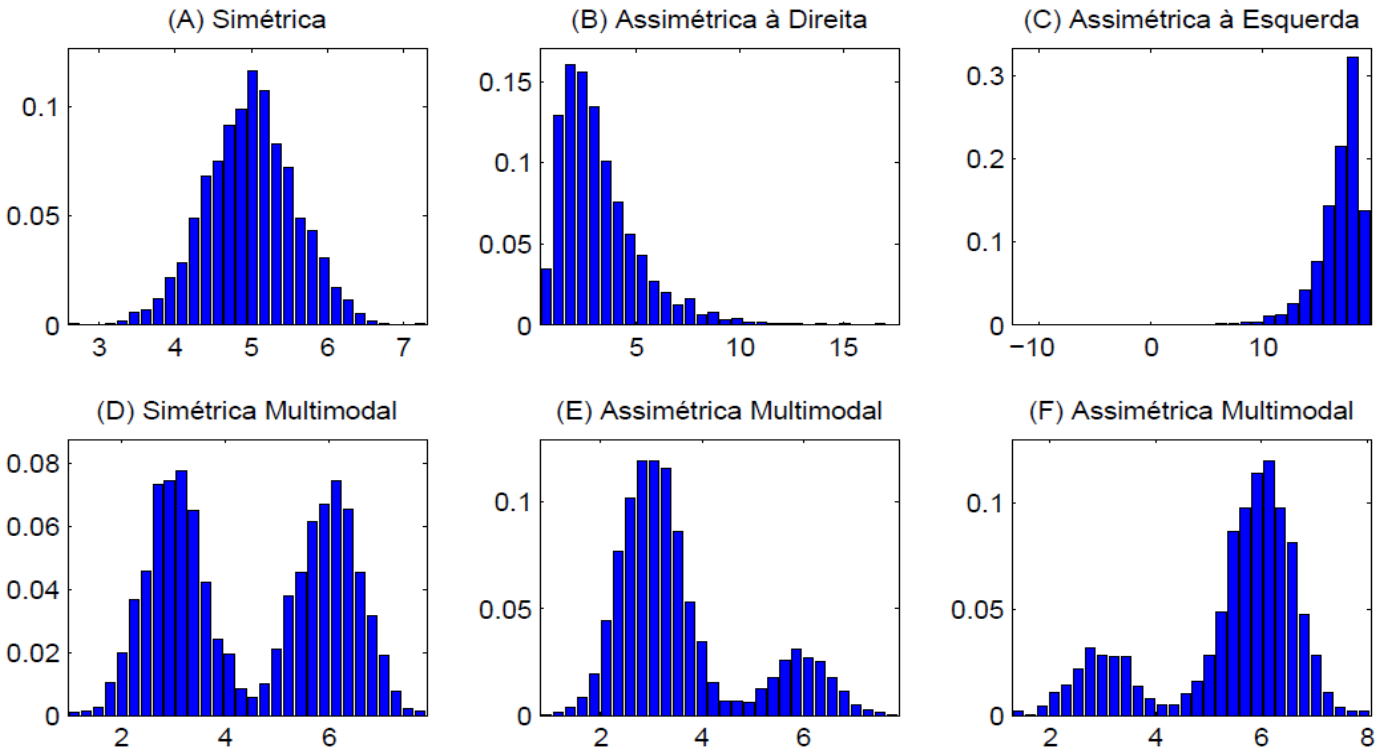
Análises Gráficas

- Histogramas de frequências absolutas e relativas
  - Divide o intervalo total dos dados em subintervalos iguais e conta o número de observações em cada subintervalo – frequências absolutas
  - Ao invés de usar o total de observações em cada subintervalo, utiliza o percentual de observações em cada subintervalo

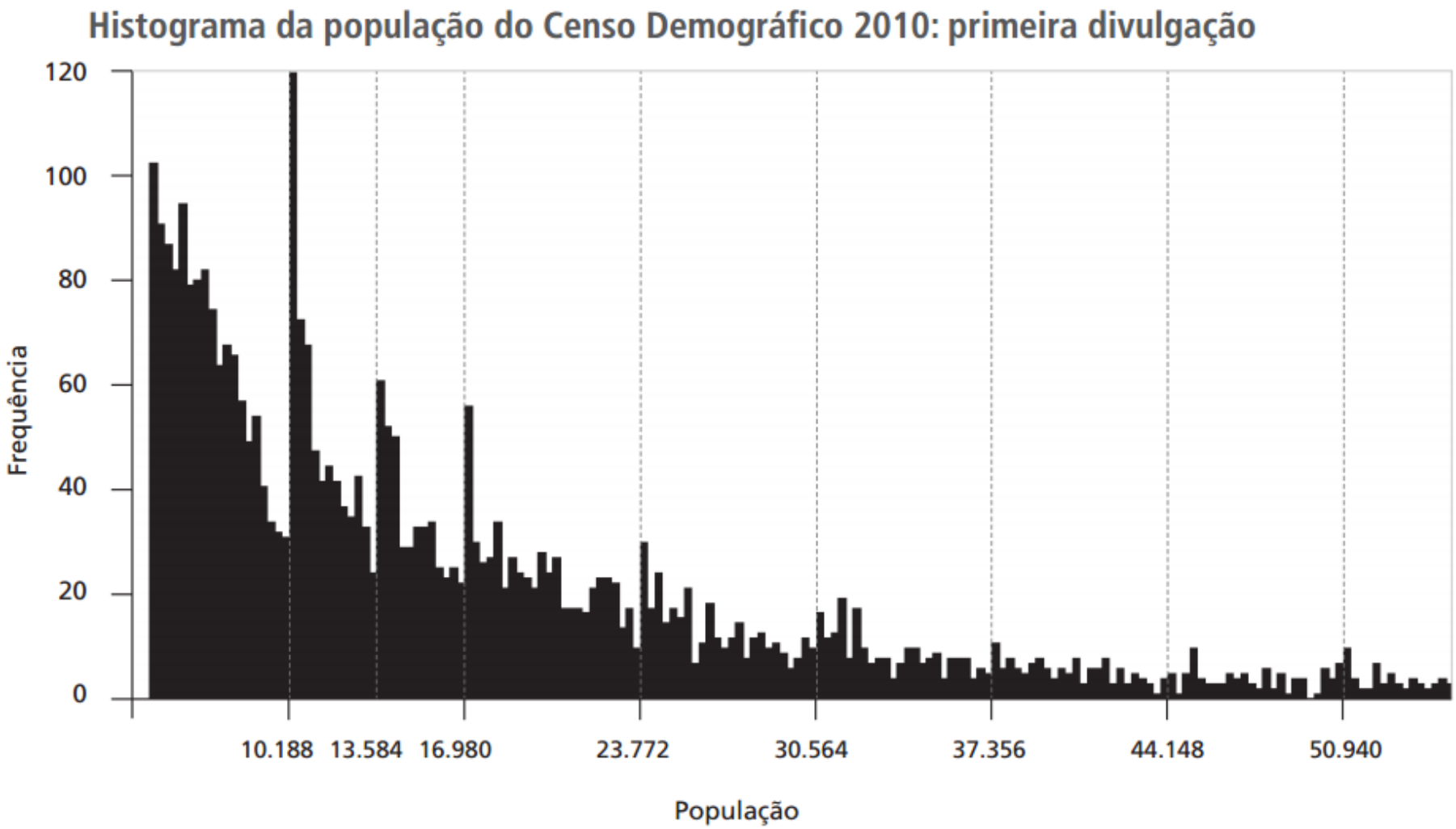


Análises Gráficas

- Identificando padrões gerais nos dados com histogramas
  - Distribuições simétricas e assimétricas
  - Modas



Análises Gráficas



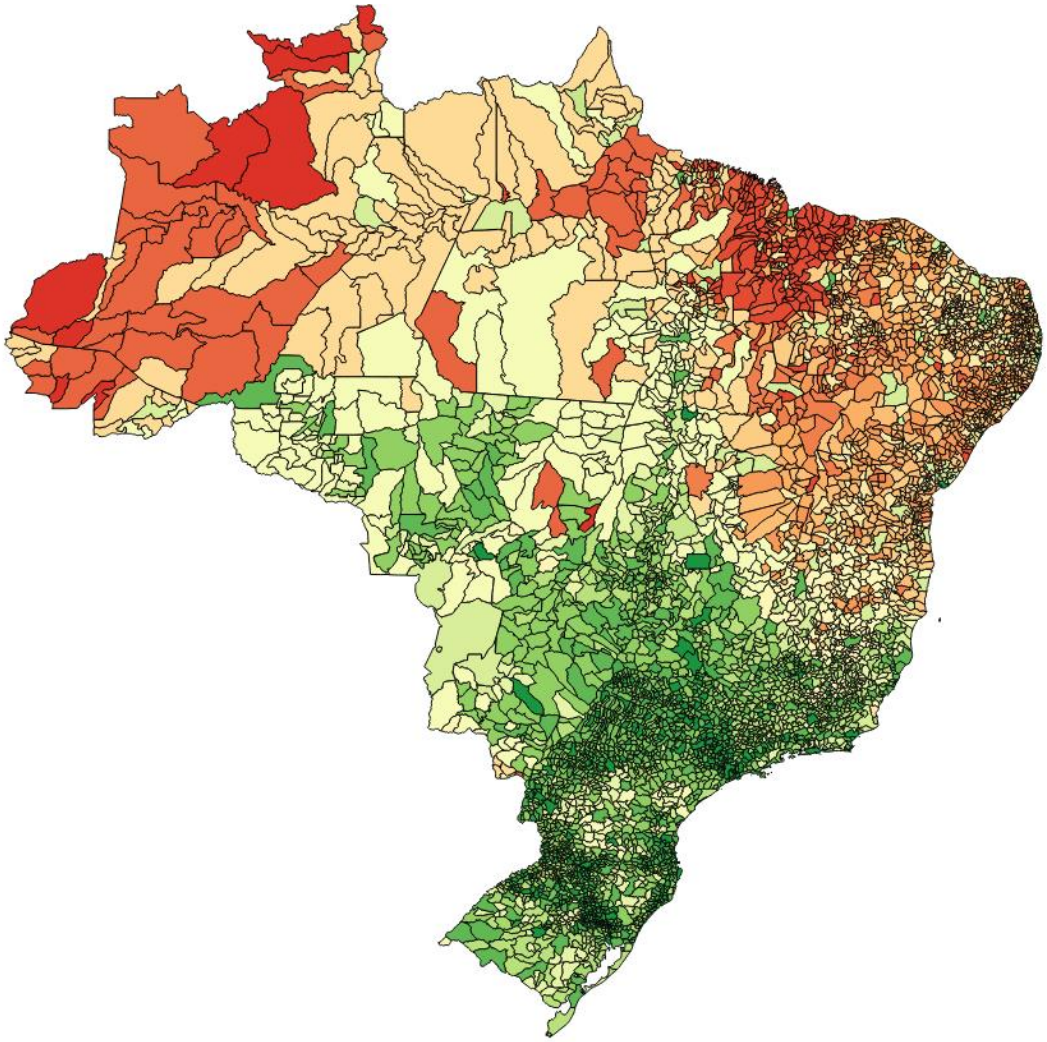
Fonte: IBGE (2011) e cálculos do autor.
Obs.: as mudanças de faixa do FPM estão indicadas pelas linhas pontilhadas. O tamanho das classes do histograma (bin) é igual a 283 habitantes.

Exemplo

Tipologias de Município Brasileiros através de Análise de Agrupamentos

Resultados

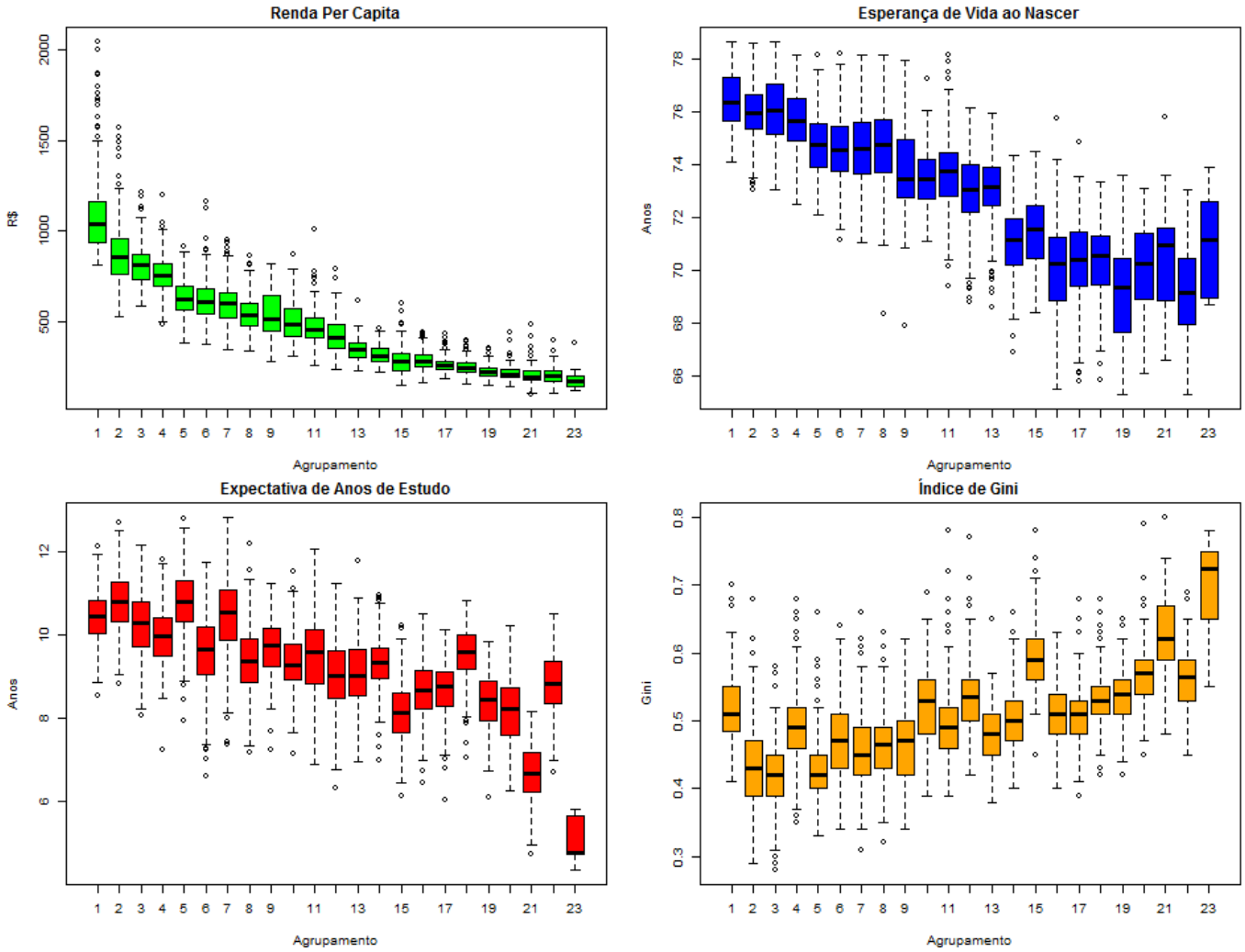
- A figura ao lado apresenta os 23 clusters (agrupamentos)
- Municípios em cor mais esverdeada possuem maior renda per capita
- Municípios em cor mais avermelhada possuem menor renda per capita



Observações:

Grupos de municípios mais pobres, com tendência de maior desigualdade

Grupo 23 se destaca pelos indicadores bem piores que os demais grupos

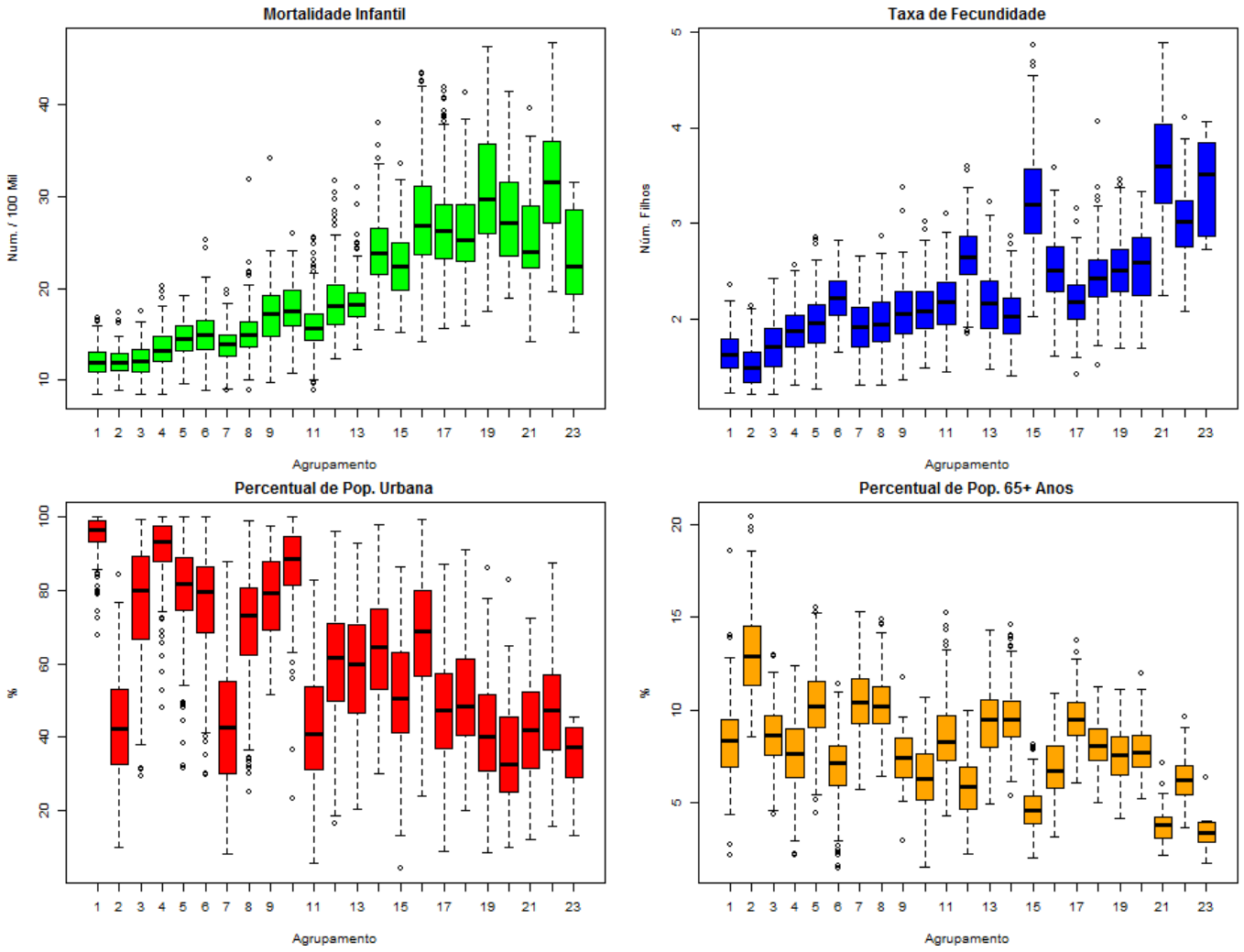


Observações:

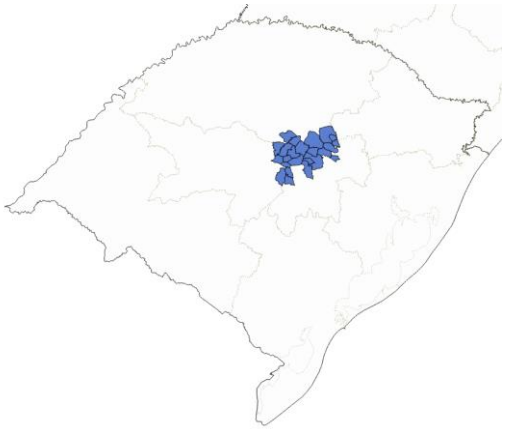
Grupo 2 com alto percentual de população rural e com alta renda

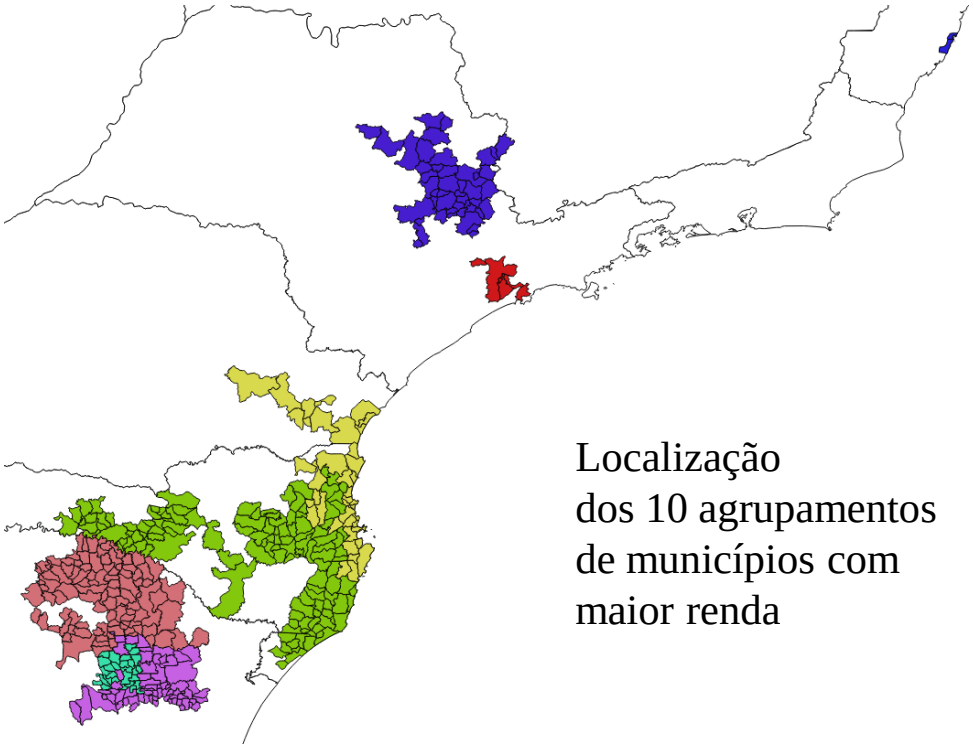
Grupos 7 e 11 com renda alta/mediana, e também com alto percentual de população rural

Grupo 15 com baixa renda e com alta fecundidade



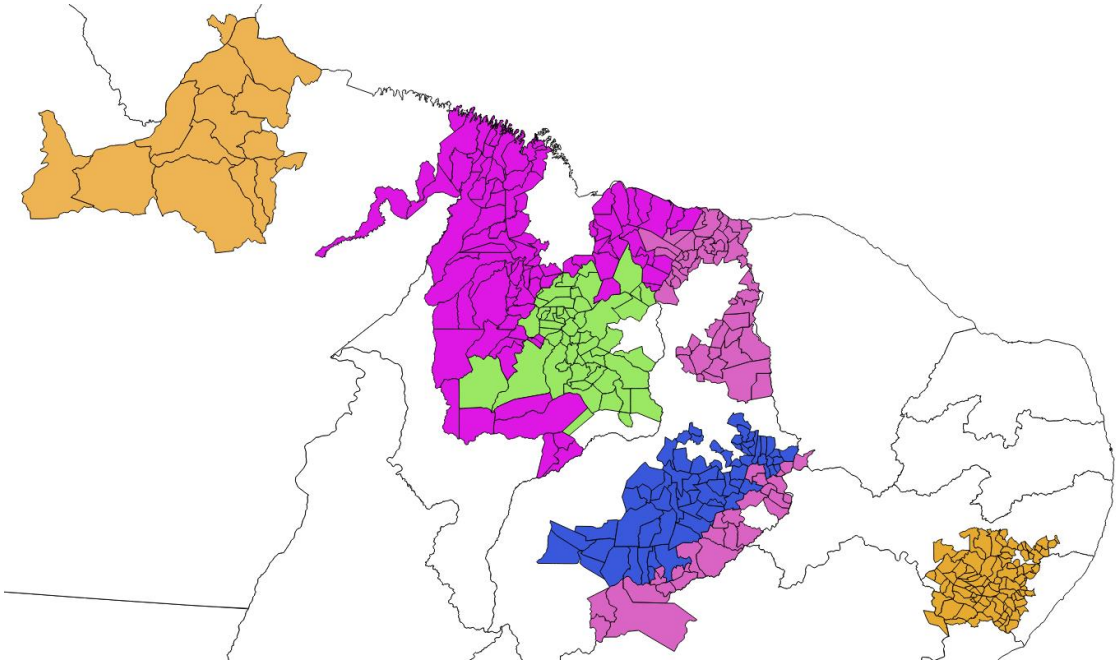
Agrupamento
com renda alta
e baixíssimo percentual
de população urbana

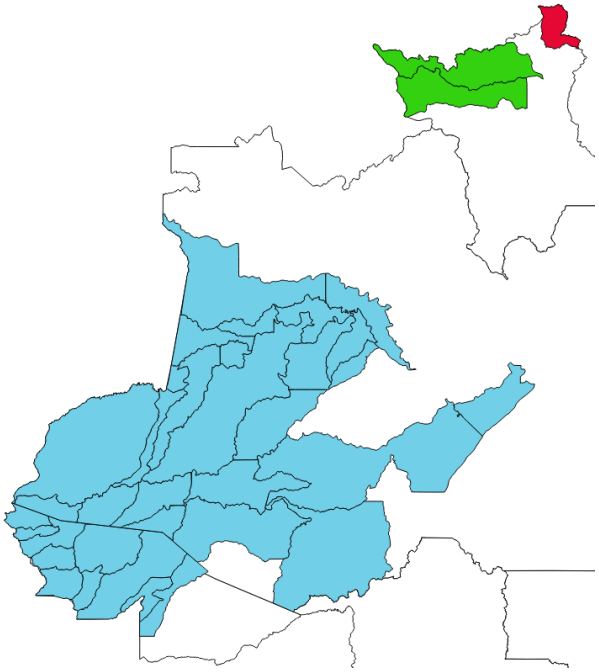




Localização
dos 10 agrupamentos
de municípios com
maior renda

Localização
dos 10 agrupamentos
de municípios com
menor renda





Coeficientes de Assimetria

- Medida para identificar assimetria dos dados
  - Conjunto de dados disponíveis: $x_1, x_2, x_3, \ldots, x_n$.
  - Coeficiente de assimetria (populacional):

$$CA_P = \frac{\frac{1}{n}\sum_{i=1}^{n}[x_i - \bar{x}]^3}{\left[\sqrt{\frac{1}{n}\sum_{i=1}^{n}[x_i - \bar{x}]^2}\right]^3}$$

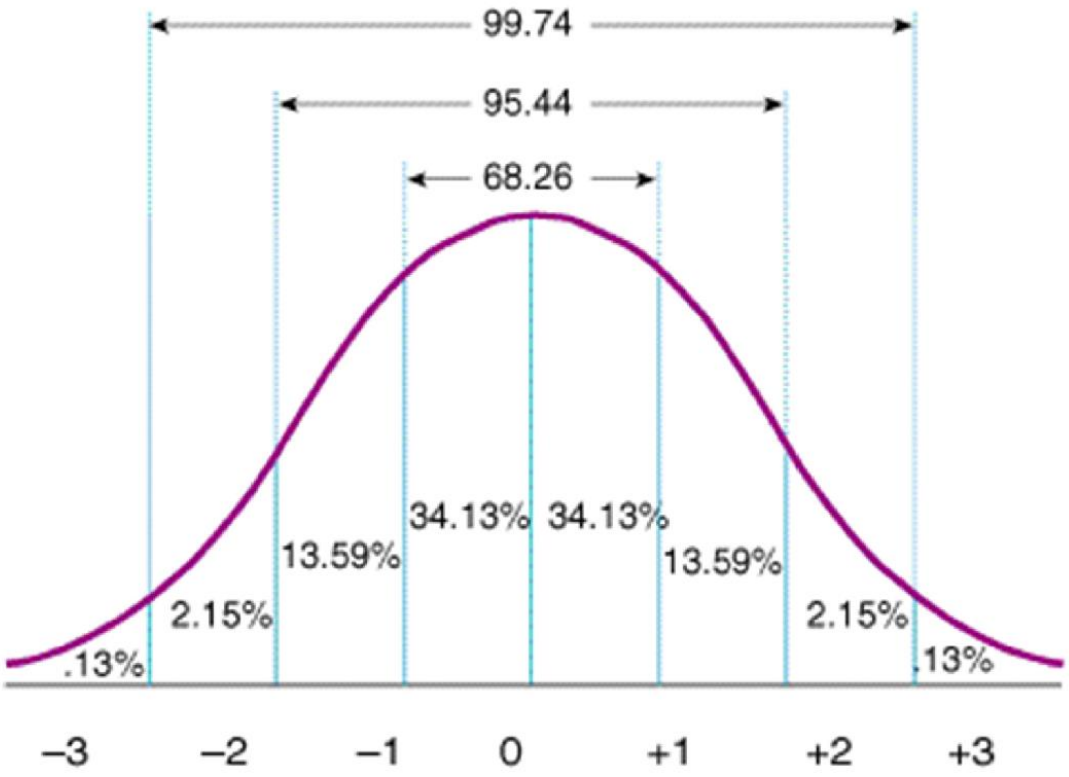
  - Coeficiente de assimetria (amostral):

$$CA_A = \frac{\sqrt{(n-1)n}}{n-2} \times \frac{\frac{1}{n}\sum_{i=1}^{n}[x_i - \bar{x}]^3}{\left[\sqrt{\frac{1}{n}\sum_{i=1}^{n}[x_i - \bar{x}]^2}\right]^3}$$

  - Termo em vermelho corresponde à correção para o viés de estimação
  - CA_A = 0 implica distribuição simétrica
  - CA_A > 0 (< 0) implica assimetria à direita (à esquerda)

A Distribuição Normal (Gaussiana)

- Diversos modelos nas mais diversas áreas assumem que os dados observados possuem distribuição similar à uma distribuição normal (curva de Gauss):



Curtose – 'caudas pesadas'

- Medida para identificar a frequência de eventos extremos
  - Conjunto de dados disponíveis: $x_1, x_2, x_3, \ldots, x_n$.
  - Curtose (populacional):

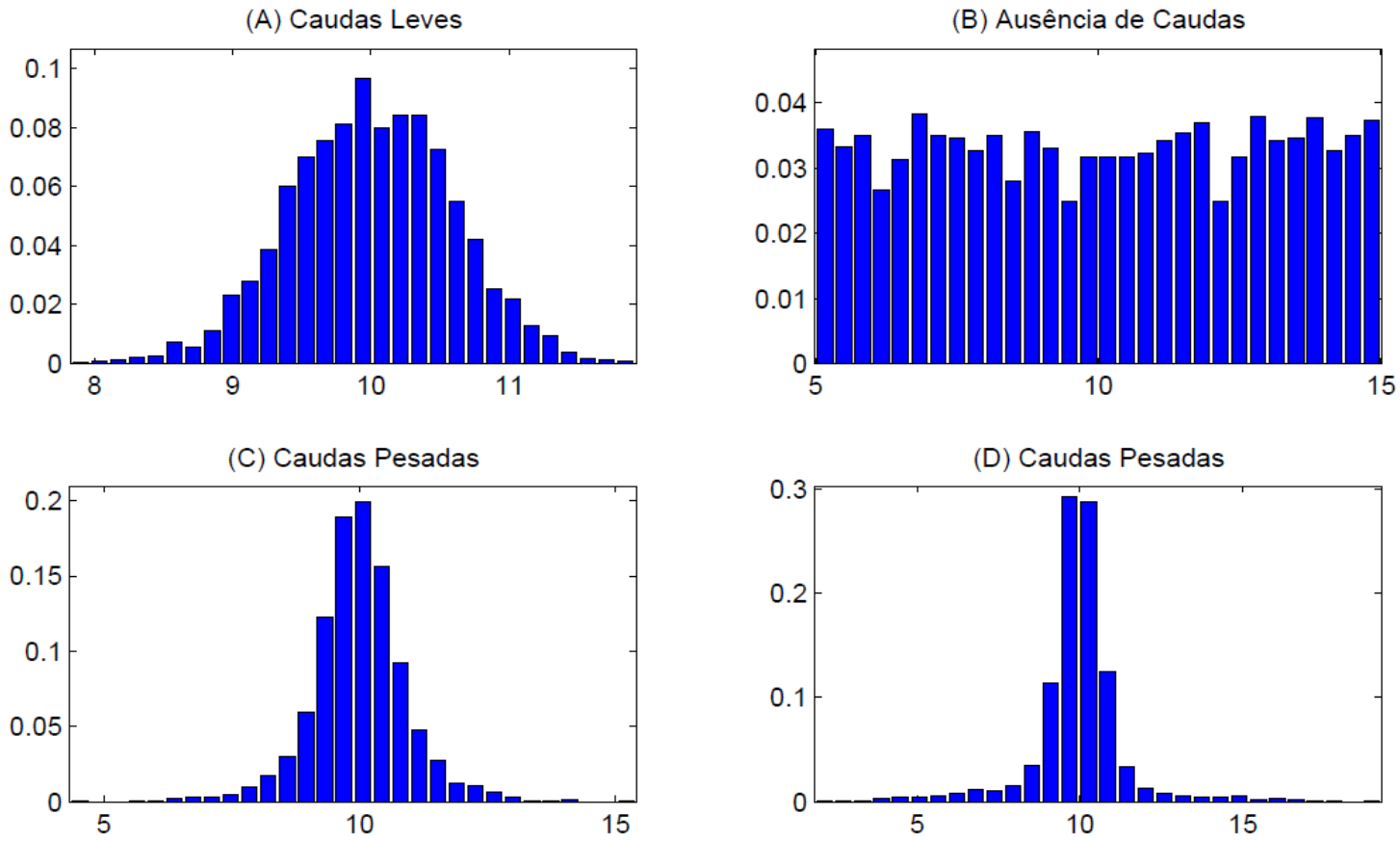
$$K_P = \frac{\frac{1}{n}\sum_{i=1}^{n}[x_i - \bar{x}]^4}{\left[\sqrt{\frac{1}{n}\sum_{i=1}^{n}[x_i - \bar{x}]^2}\right]^4}.$$

  - Curtose (amostral):

$$K_A = 3 + [(n+1)K_P - 3(n-1)] \times \frac{n-1}{(n-2)(n-3)}.$$

  - Para a distribuição normal, a curtose é igual a 3 sempre
  - Modelos de WACC (*weighted average cost of capital*) utilizam-se de estimativas para a remuneração do capital próprio do investidor
    - Modelos CAPM para estimar a remuneração pelo risco do negócio
    - Os modelos CAPM utilizam-se de dados financeiros (variações percentuais nos preços dos ativos)
    - As observações de variações percentuais em geral apresentam curtose bem maior do que 3

Curtose – ‘caudas pesadas’



Valores calculados para as curtoses: (A) $K_A = 3.0$; (B) $K_A = 1.8$; (C) $K_A = 5.9$; (D) $K_A = 10.8$

Introdução ao Software Estatístico R

Software R – Princípios Básicos

- O R é um *software* gratuito para análises estatísticas, econométricas e matemáticas.
- Foi desenvolvido baseado em uma linguagem anterior denominada **S.**
- O **R** deve ser instalado antes do **Rstudio** e pode ser obtido pelo sítio:

http://www.vps.fmvz.usp.br/CRAN/

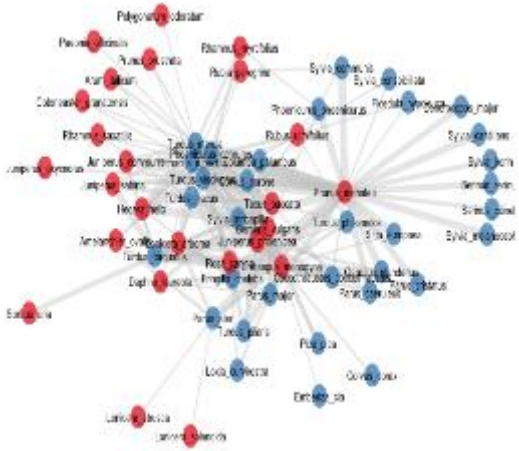
- O RStudio é uma IDE (Integrated Development Environment) para o uso do **R**.
- Por ser mais "amigável" é frequentemente utilizado.
- O **Rstudio** pode ser obtido por meio do *link*:

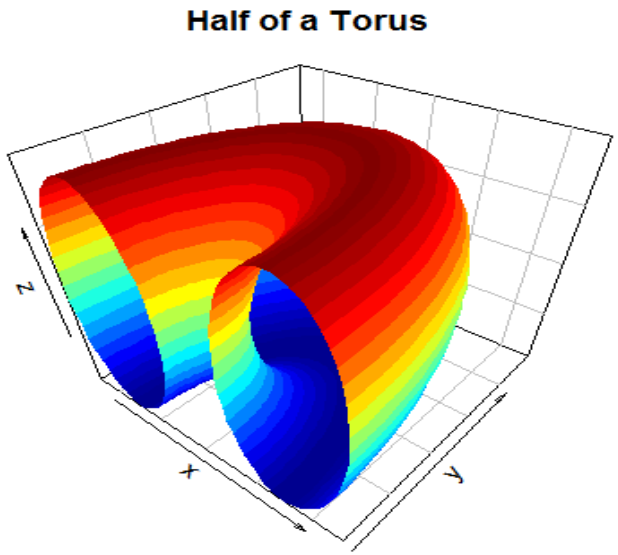
https://www.rstudio.com/products/rstudio/download/

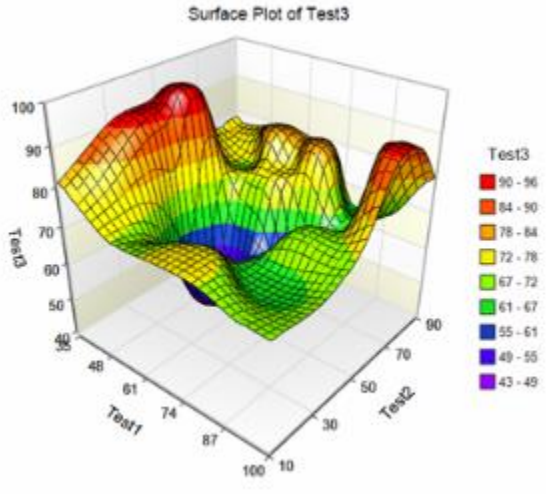
Por que o R?

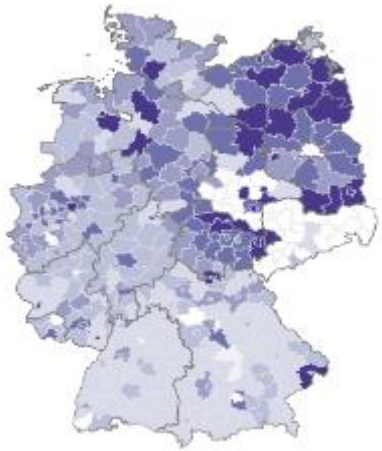
- R é gratuito. Como um projeto open-source, você pode usar R gratuitamente: não há necessidade de se preocupar com taxas de assinatura, gerenciamento de licenças ou limites do usuário.
- Mas tão importante, R é aberto: você pode inspecionar o código e mexer com ele (desde que respeitem os termos da *GNU General Public License versão 2 em que é distribuído*).
- Milhares de especialistas ao redor do mundo fizeram exatamente isso, e suas contribuições beneficiam as milhões de pessoas que usam R hoje.
- R é um idioma. No R, você faz a análise dos dados por funções e scripts escritos, não é apenas apontar e clicar. Isso pode parecer assustador, mas é uma língua fácil de aprender.
- R promove a experimentação e exploração, o que melhora a análise de dados e muitas vezes leva a descobertas que não seriam feitas de outra forma.
- Os documentos de script registram todo o seu trabalho, desde o acesso aos dados até a geração de relatórios e pode imediatamente re-executado a qualquer momento.
- Scripts também tornam mais fácil a automatização de uma sequência de tarefas que podem ser integradas a outros processos.

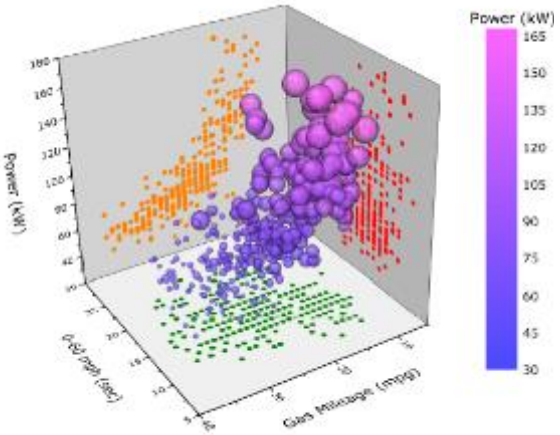
Por que o R?

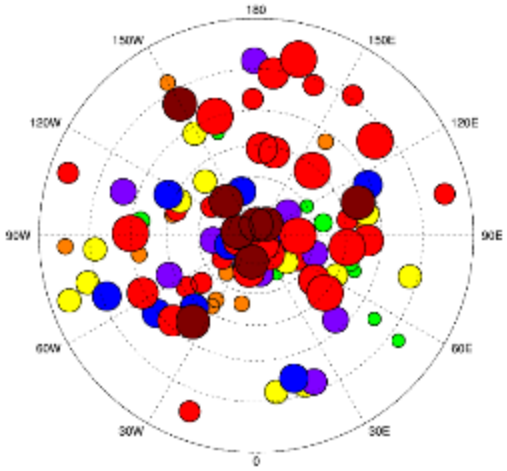



Half of a Torus
z
x
y


Surface Plot of Test3
Test3
Test1
Test2
Test3
90 - 96
84 - 90
78 - 84
72 - 78
67 - 72
61 - 67
55 - 61
49 - 55
43 - 49




Power (kW)
Old mpg (mpg)
Gas Mileage (mpg)

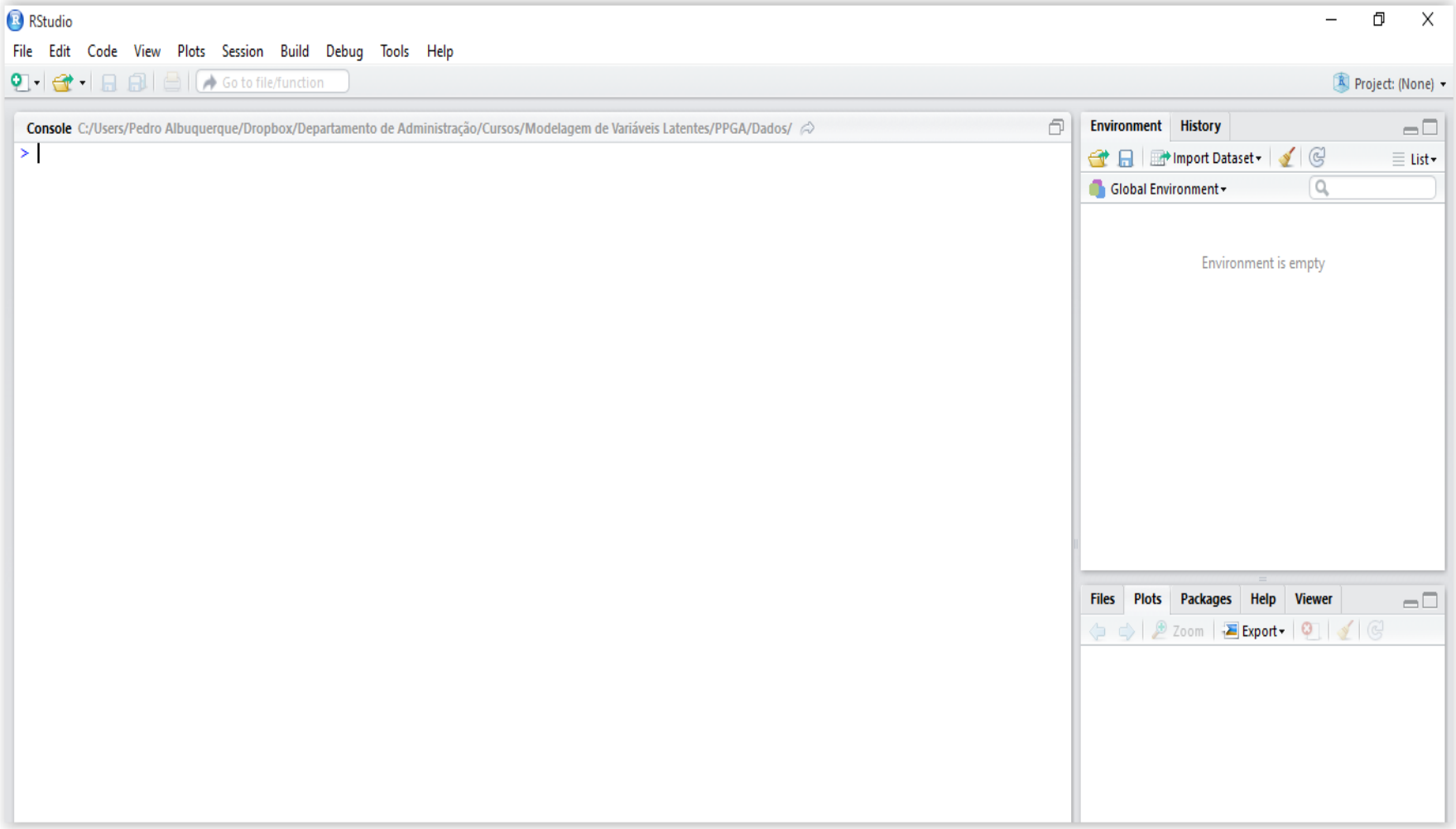


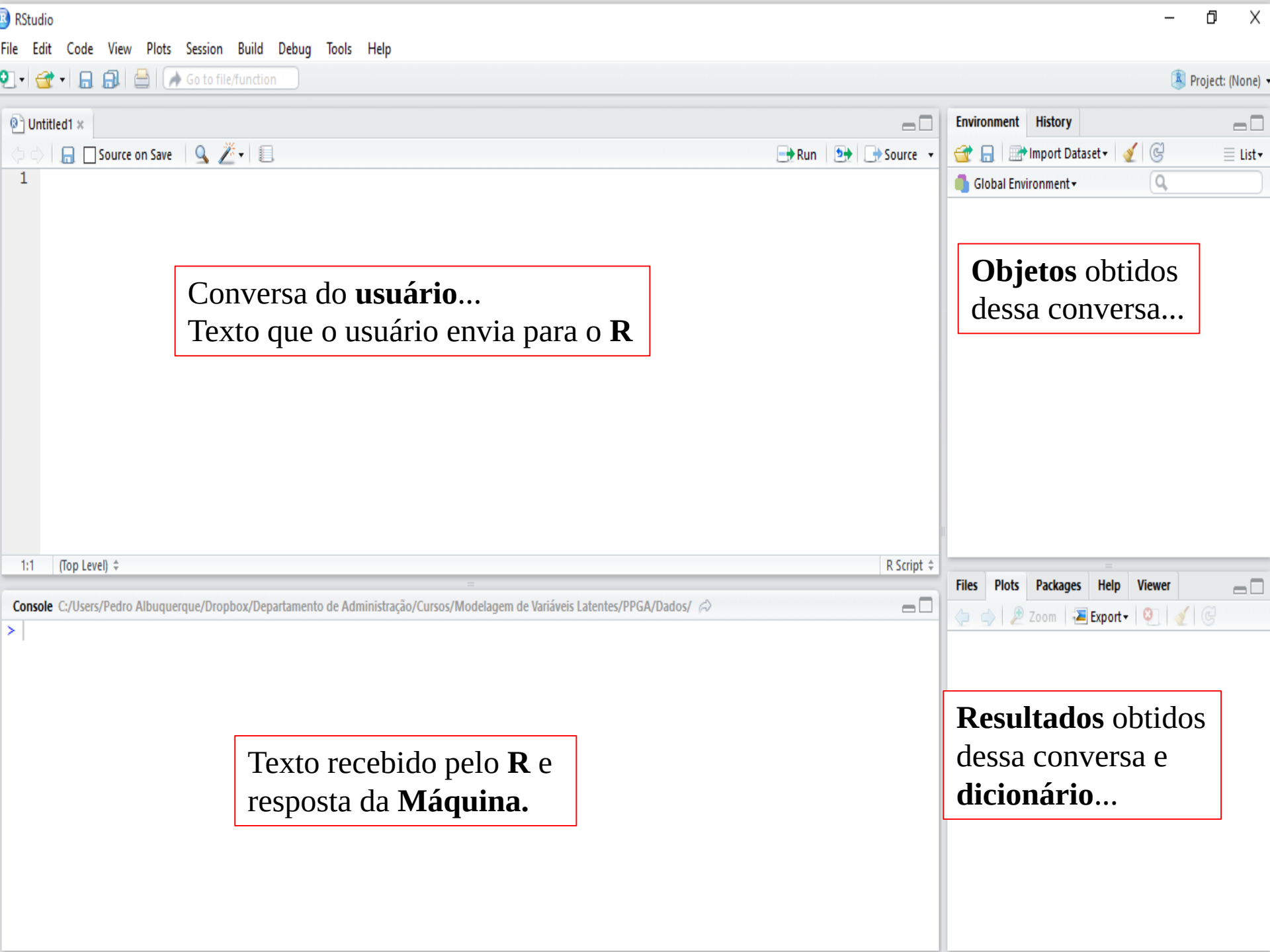
Por que o R?

- **Uma, vibrante e robusta comunidade.**
- Com milhares de colaboradores e mais de dois milhões de usuários ao redor do mundo.

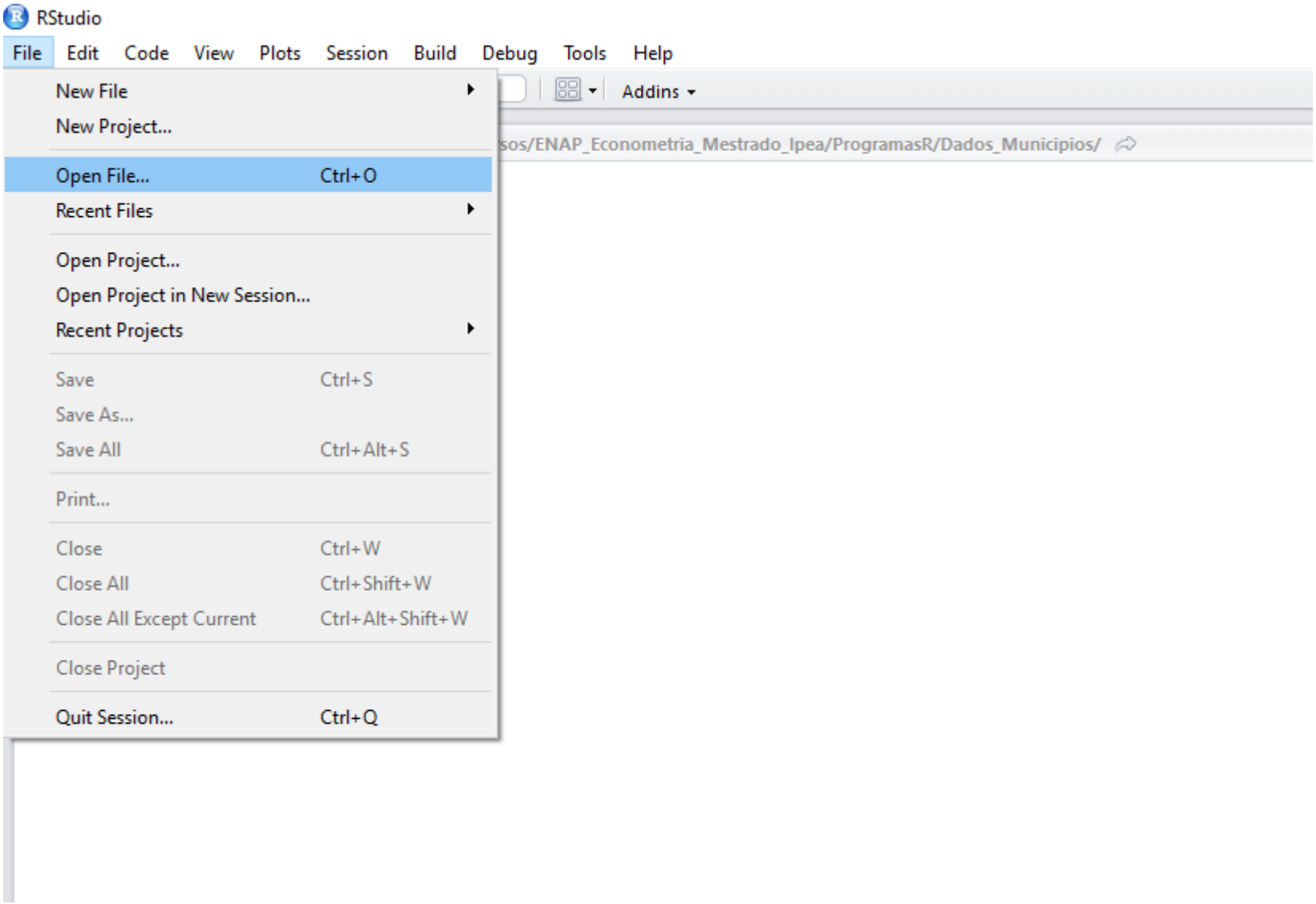
- **Possibilidades ilimitadas.** Com o R, você não está restrito a escolha de um conjunto pré-definido de rotinas.
- Você pode usar o código contribuído por outros membros da comunidade *open-source*, ou estender o R com suas próprias funções.

Software R – Princípios Básicos

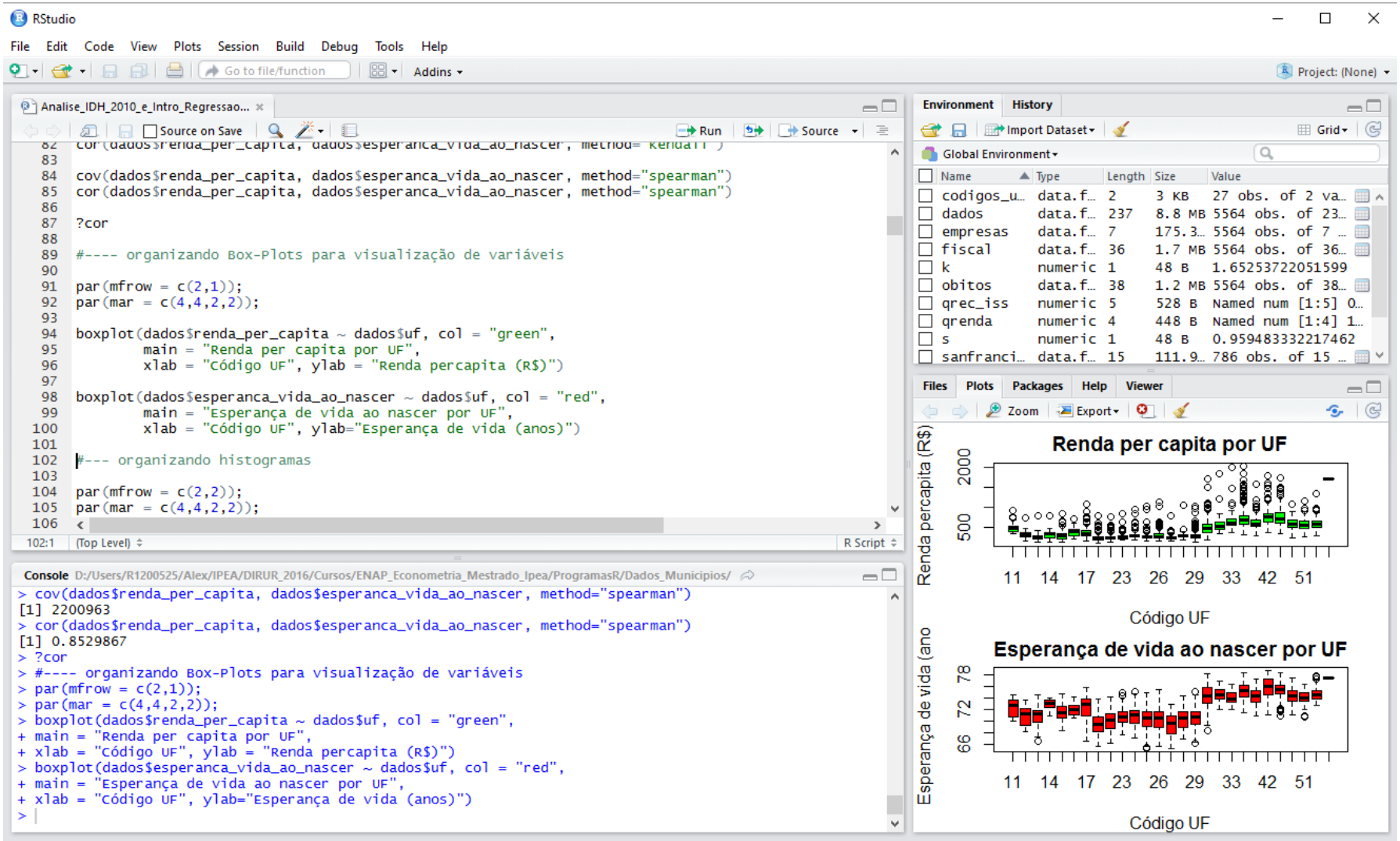



RStudio
File Edit Code View Plots Session Build Debug Tools Help
Go to file/function
Project: (None)
Untitled1
Source on Save
Run
Source
1
1:1
(Top Level)
R Script
Environment
History
Import Dataset
List
Global Environment
Files
Plots
Packages
Help
Viewer
Zoom
Export
Console C:/Users/Pedro Albuquerque/Dropbox/Departamento de Administração/Cursos/Modelagem de Variáveis Latentes/PPGA/Dados/
>
Conversa do **usuário**...
Texto que o usuário envia para o **R**
Objetos obtidos dessa conversa...
Texto recebido pelo **R** e resposta da **Máquina.**
Resultados obtidos dessa conversa e **dicionário**...

Software R – Princípios Básicos



Software R – Princípios Básicos



Exercícios em Excel ou em R – Não Precisa Entregar

Utilizando a planilha "IDH_Brasil_2010.xlsx", ...

- Plot histogramas para as seguintes variáveis na tabela de dados:
  - IDHM_educacao
  - IDHM_renda
  - IDHM_logenvidade
  - expec_anos_estudo
- Para essas variáveis acima, calcule o coeficiente de assimetria e a kurtose
  - Funções no Excel: distorção() e distorção.p() para coeficientes de assimetria amostral e populacional; curt() para kurtose amostral (não tem kurtose populacional)
  - Verifique se os coeficientes de assimetria estão coerentes com os histogramas para as quatro variáveis estudadas

Relação entre Variáveis

Relações entre Variáveis

- Ao invés de estudar as variáveis individualmente, nós queremos estudar como variáveis na amostra de dados se relacionam
  - Por exemplo, podemos estar interessados na relação entre o ativo imobilizado em serviço (AIS) e os ativos referentes a obrigações especiais, para concessionárias de energia elétrica
  - Sejam então (x_1, y_1), (x_2, y_2), ... , (x_n, y_n), os pares de valores referentes às *n* unidades observacionais
  - Covariância (populacional):

$$Cov_P(X,Y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

  - Covariância (amostral):

$$Cov_A(X,Y) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

  - Quando X = Y, a covariância é igual a ... ?
  - A covariância não é invariante a mudanças de escala (quando passamos de toneladas para kg por exemplo)

Relações entre Variáveis

- Correlação de Pearson entre duas variáveis:
  - Considere novamente $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$, os pares de valores referentes às *n* unidades observacionais
  - Correlação (populacional):

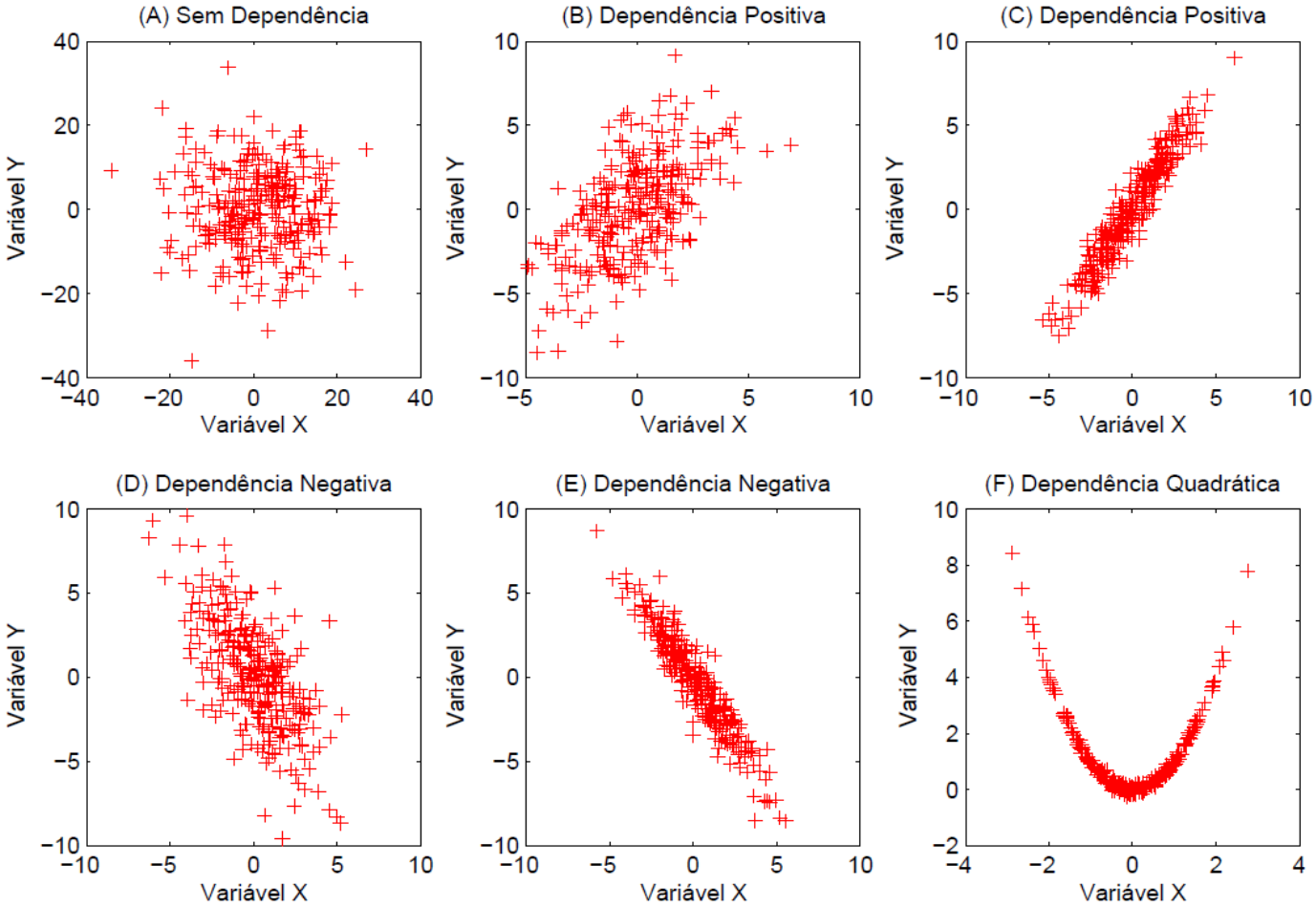
$$\rho = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 \times \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2}}.$$

  - Correlação (amostral):

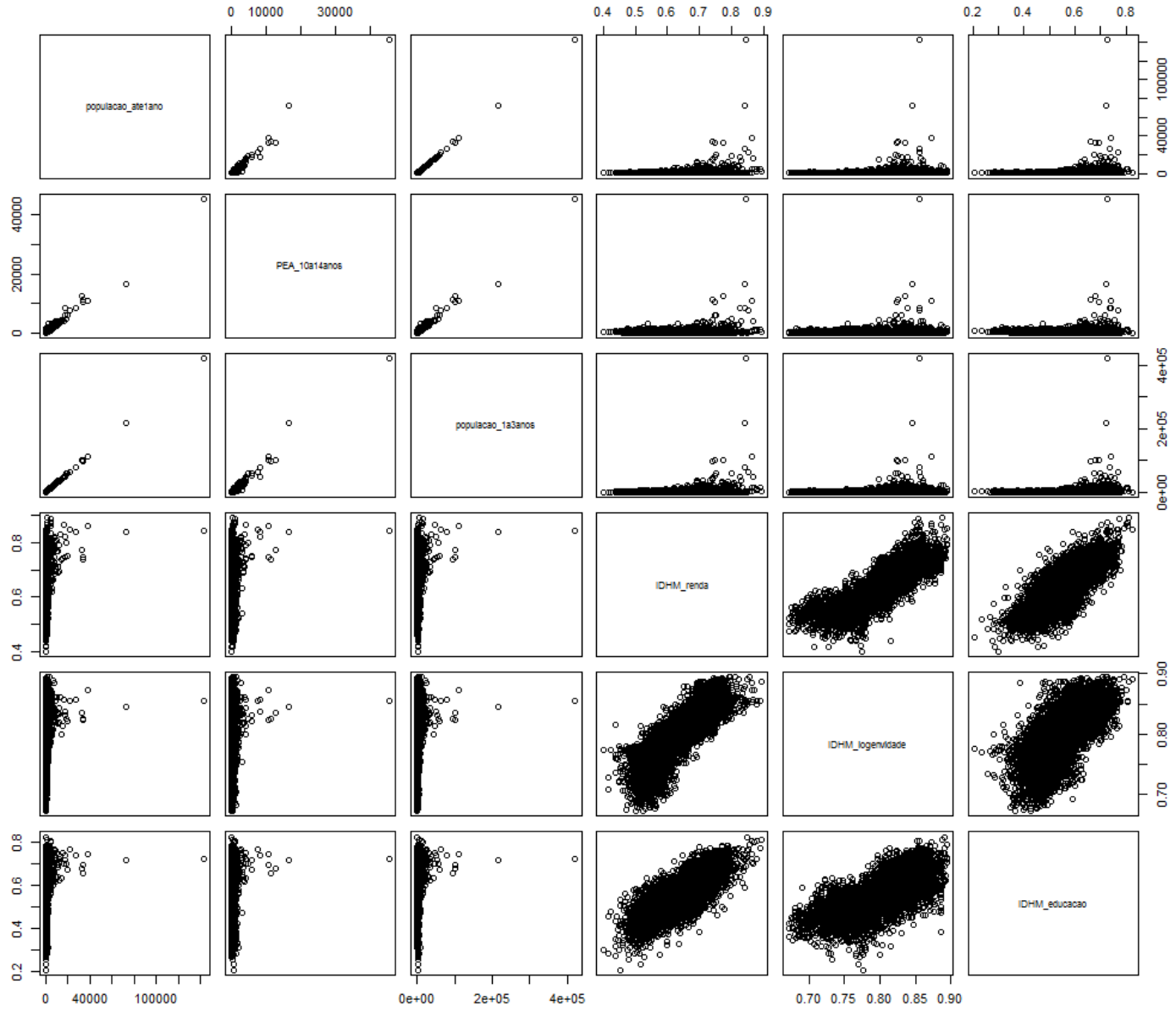
$$r = \frac{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 \times \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2}}.$$

  - Quando X = Y, a covariância é igual a ... ?
  - O coeficiente de correlação nos fornece uma indicação sobre **relações lineares** entre as duas variáveis sendo estudadas conjuntamente

Gráficos de Dispersão



- Valores dos coeficientes de correlação amostral são: (A) r = - 0.01; (B) r = 0.63; (C) r = 0.95; (D) r = - 0.63; (E) r = - 0.96; (F) r = - 0.01 nos gráficos



- Comando “pairs()” no R

Matriz de Variância-Covariância

Variâncias na diagonal principal e covariâncias fora da diagonal principal (matriz simétrica)

$$\Sigma = \begin{bmatrix} \mathrm{Var}(X) & \mathrm{Cov}(X,Y) & \mathrm{Cov}(X,Z) \\ \mathrm{Cov}(X,Y) & \mathrm{Var}(Y) & \mathrm{Cov}(Y,Z) \\ \mathrm{Cov}(X,Z) & \mathrm{Cov}(Y,Z) & \mathrm{Var}(Z) \end{bmatrix}$$

Matriz de Correlações

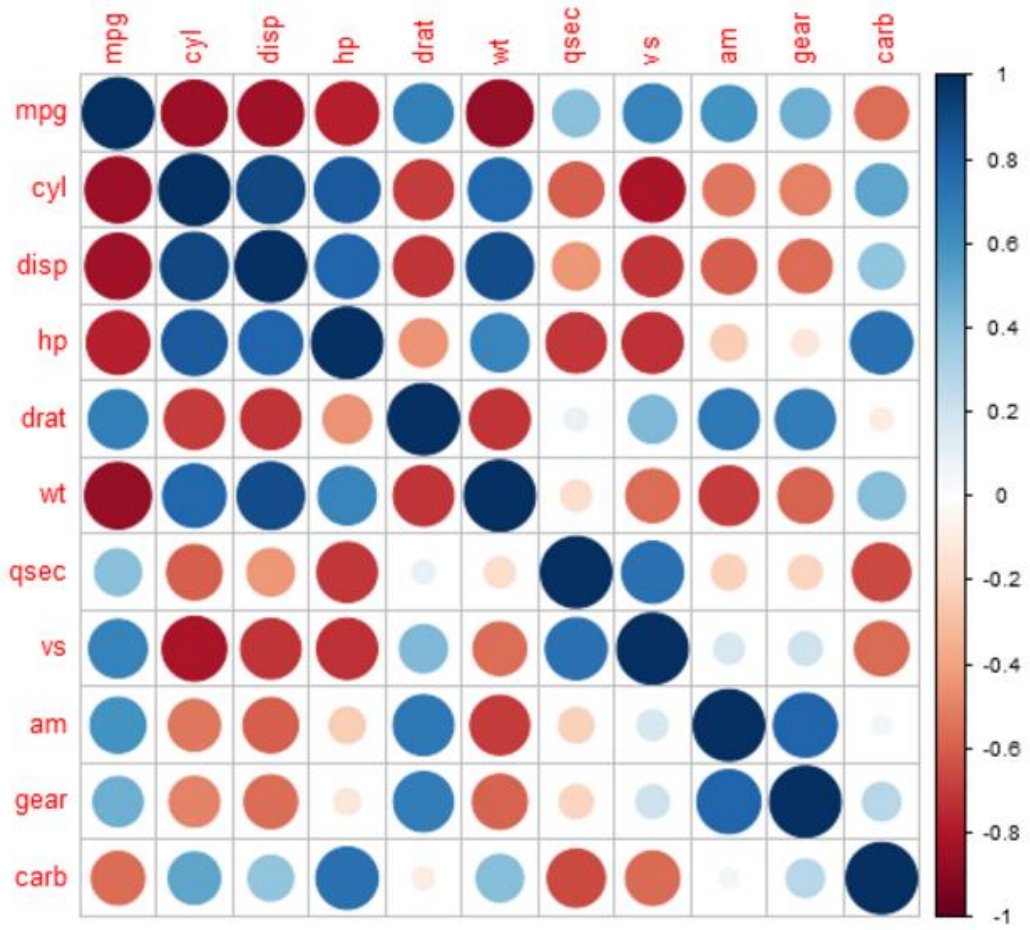
Correlação entre as variáveis fora da diagonal principal (matriz simétrica com diagonal principal com todos os elementos iguais a 1)

```
. pwcorr
```

	Happin~s	Exercise	Sleep	Jobsat~n	Pets
Happiness	1.0000				
Exercise	0.6056	1.0000			
Sleep	-0.1952	-0.4974	1.0000		
Jobsatisfa~n	0.8601	0.7312	0.0246	1.0000	
Pets	0.6590	0.7897	-0.4082	0.5847	1.0000

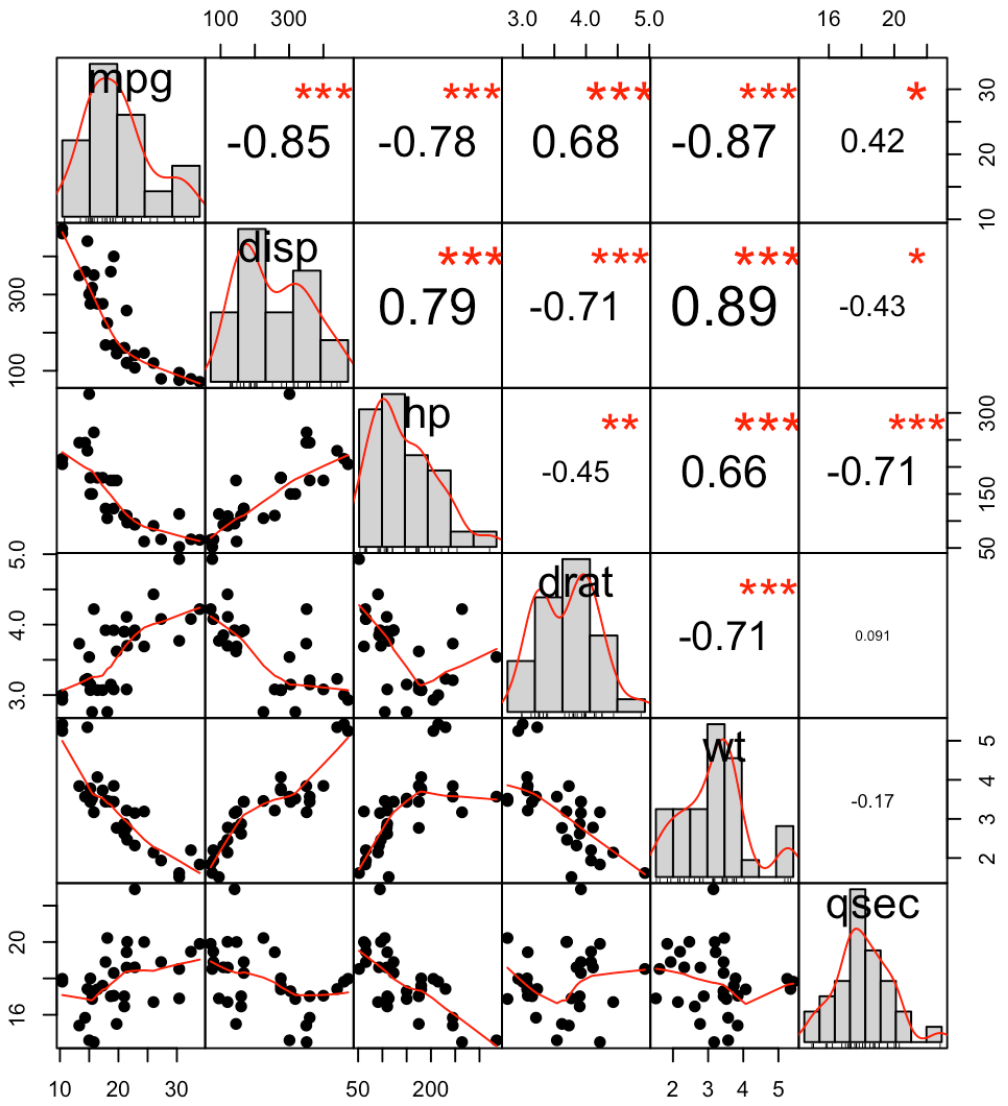
Matriz de Correlações

```
library(corrplot)
M <- cor(mtcars)
corrplot(M, method="circle")
```



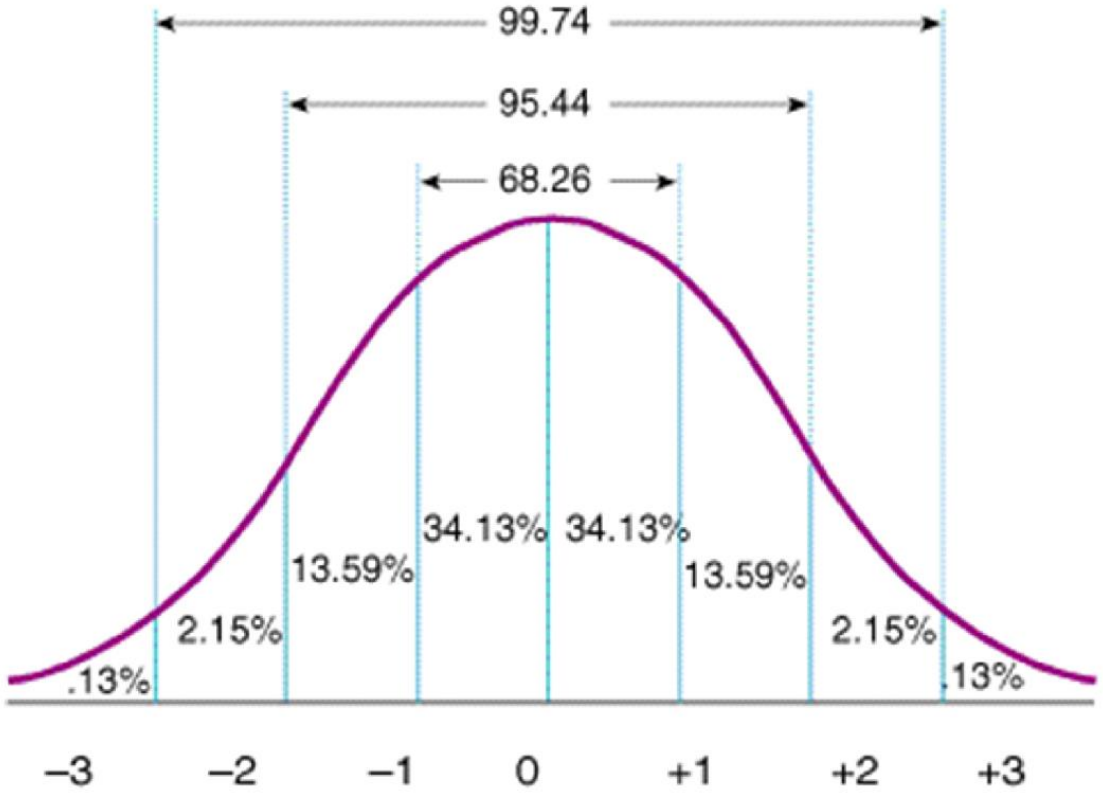
Matriz de Correlações

```
library(psych)
pairs.panels(dados_cor)
```



A Distribuição Normal (Gaussiana) Univariada

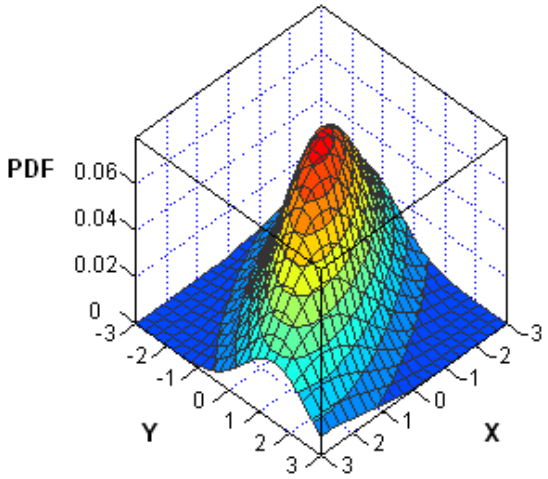
- Aplica-se a uma variável aleatória apenas

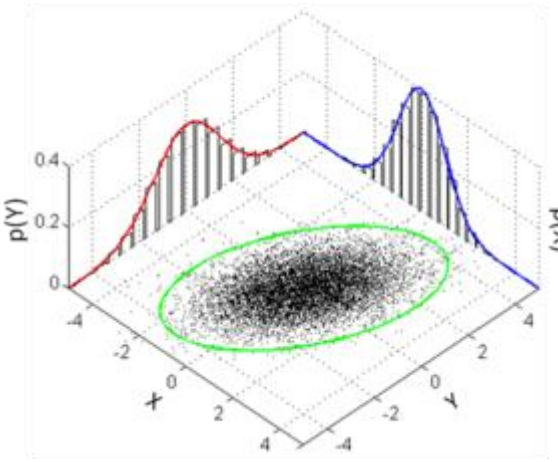


- Depende de dois parâmetros: a média μ e a variância σ^2

A Distribuição Normal Multivariada

- Aplica-se a um vetor de variáveis, que podem ser correlacionadas entre sim





- As distribuições normais multivariadas dependem de dois parâmetros:
  - Um vetor de médias de cada variável separada $(\mu_1, \mu_2, \dots, \mu_k)$
  - Uma matriz de covariâncias Σ

- Fato importante: se um vetor tem distribuição normal multivariada, com vetor de médias $(\mu_1, \mu_2, \dots, \mu_k)$ e matriz de covariâncias Σ, então cada elemento individualmente possui distribuição normal univariada. Primeiro elemento tem média μ_1 e variância σ_1^2

Exercícios em R – Para Entregar

Exercício 1 - Utilizando a tabela "IDH_Brasil_2010.csv", ...

- Considere as seguintes variáveis na tabela de dados:
  - IDHM_educacao
  - IDHM_renda
  - IDHM_logenvidade
  - expec_anos_estudo
- Para os pares entre as variáveis acima, calcule as covariâncias e os coeficientes de correção
  - Funções no Excel caso queira checar: covariação.s() e covariação.p() para covariâncias amostral e populacional. Correl() para coeficiente de correlação amostral.
- Faça um gráfico de dispersão entre o IDHM_educacao e o IDHM_renda
- Calcule as matrizes de correlação e covariâncias para as quatro variáveis acima
- **Em grupos de 2 ou 3 alunos – entregar o código em R e os resultados**
- **Prazo para entrega – 2 semanas**

Modelos de Regressão

Exemplo

Table 4. Final logistic regression model of the variable amputation as a function of social and clinical variables

Variable	β	S.E.	Wald χ^2	df	p-value	OR	95% CI
Lack of primary care assistance	1.193	0.584	4.176	1	0.041	3.30	1.05–10.36
Previous amputation	2.390	0.740	10.434	1	0.001	10.91	2.56–46.51
CKD	0.835	0.576	2.102	1	0.147	2.31	0.75–7.12
CAD	1.68	0.689	5.92	1	0.015	5.35	1.38–20.68
AA	2.77	1.07	6.67	1	0.010	15.90	1.95–129.63
Hemoglobin A1C	1.58	0.282	31.46	1	<0.001	4.87	2.80–8.47
Constant	−14.33	2.32	38.18	1	<0.001	--	---

Rcr (Cox and Snell R^2)=0.547; RN (Nagelkerke R^2)=0.749

B: Coefficient of the logistic regression equation to predict the dependent variable using the independent variable.

SE: Standard errors associated with the coefficients.

Wald: Wald chi-squared test to test the null hypothesis that the constant is equal to 0

df: Degree of freedom for the Wald chi-squared test.

Modelos de Regressão Linear

Modelos de Regressão

- Modelos de regressão para estudar a relação entre duas ou mais variáveis

$$y_i = g(x_{1i}, x_{2i}, \dots, x_{ki}) + \epsilon_i$$

  - Variável explicada, ou predita, ou dependente, ou resposta y_i
  - Variáveis preditoras, ou explicativas, ou covariáveis $x_{1i}, x_{2i}, \dots, x_{ki}$
  - O termo ϵ_i corresponde à parte do que observamos para a variável resposta, que não é explicada pelas variáveis preditoras
  - A função g(.) pode ter uma forma funcional conhecida, pré-especificada, ou pode ter uma forma funcional desconhecida

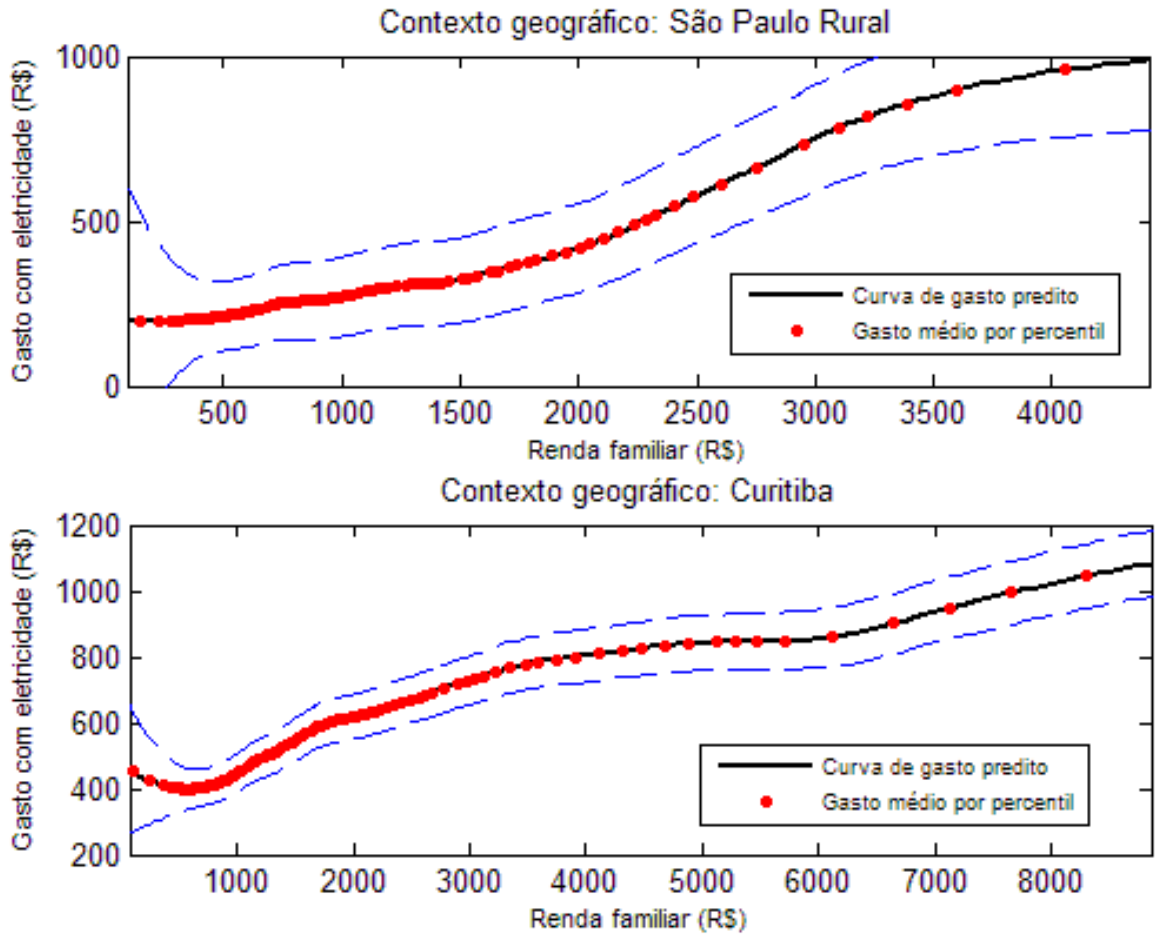
- Quanto à forma funcional para g(.),

  - Quando a função g(.) é pré-especificada, chamamos de **regressão paramétrica**
  - Quando a função g(.) é desconhecida e é estimada pelos dados, chamamos de regressão **não-paramétrica** ou **semi-paramétrica**

- Na prática, regressões paramétricas são mais utilizadas, principalmente a chamada **regressão linear**

Modelos de Regressão

- Relação entre gastos domiciliares com energia elétrica e renda dos domicílios – regressões semi-paramétricas



Modelos de Regressão

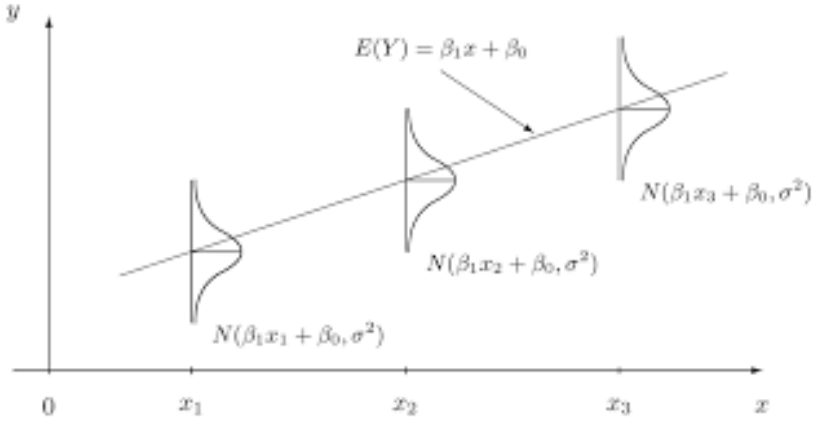
- Modelos de regressão linear
  - O tipo mais comum de modelo de regressão é modelo de regressão linear
  - Forma funcional é simplesmente uma expressão linear das variáveis explicativas
  - O termo ϵ_i corresponde à parte do que observamos para a variável resposta, que não é explicada pelas variáveis preditoras

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ \ldots + \beta_k x_{ki} + \epsilon_i$$

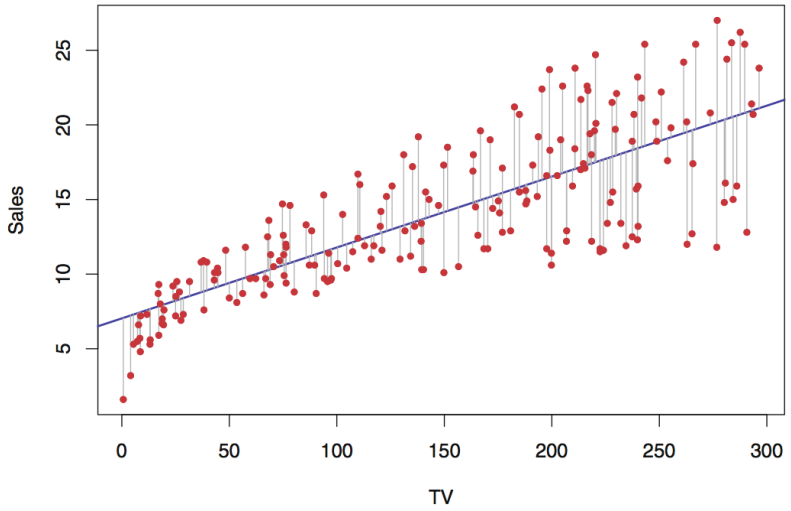
- Diversas hipóteses são descritas na literatura para o modelo de regressão linear na sua forma mais básica
  - **Os termos ϵ_i têm distribuição normal**
  - Os termos ϵ_i são não correlacionados entre eles
  - Os termos ϵ_i possuem variância constante (erros homoscedásticos)
  - Os termos ϵ_i são não correlacionados com as variáveis explicativas $x_{1i}, x_{2i}, \ldots, x_{ki}$
- Na prática, essas hipóteses básicas não se aplicam, e diversas técnicas foram criadas para tratar diferentes casos para os quais essas hipóteses não são satisfeitas

Modelos de Regressão

- Regressão linear (simples) com uma variável explicativa – erros ϵ_i com distribuições normais

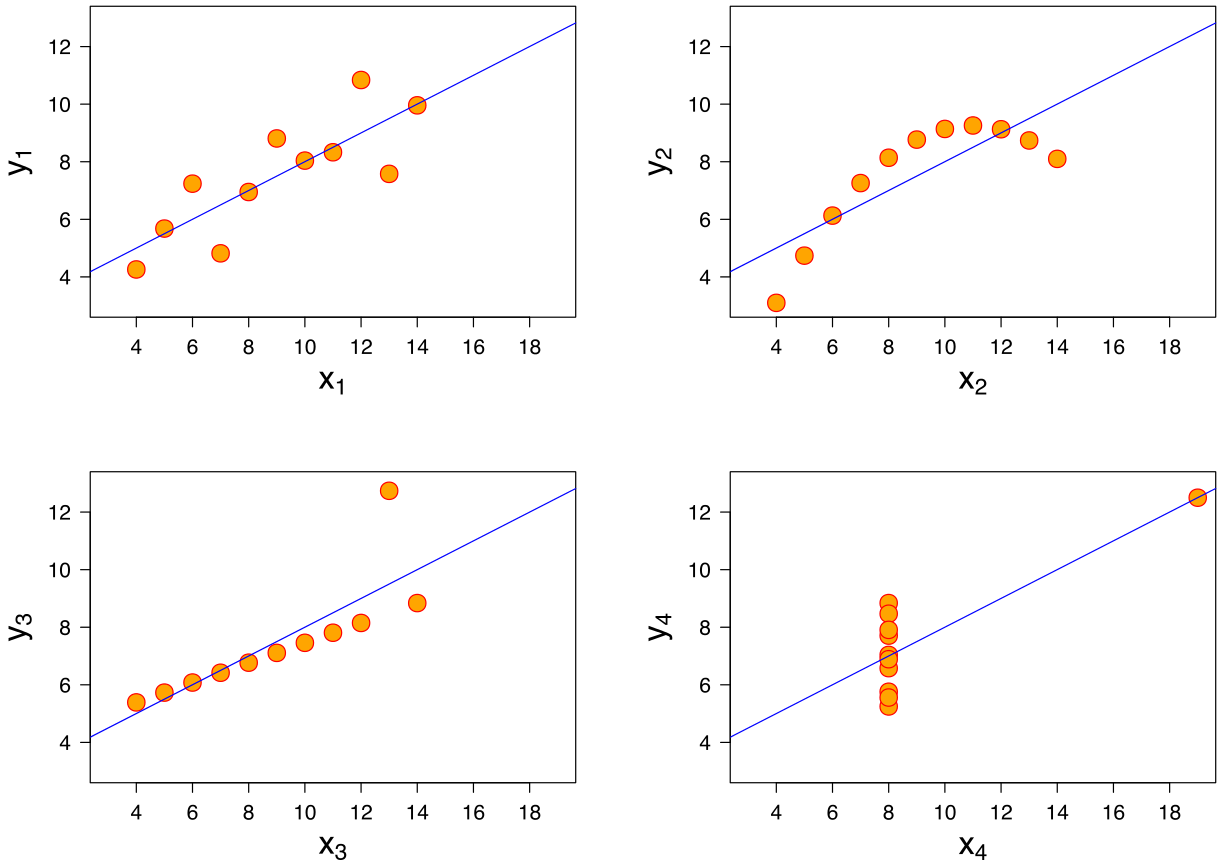


- Hipótese de homoscedasticidade é violada



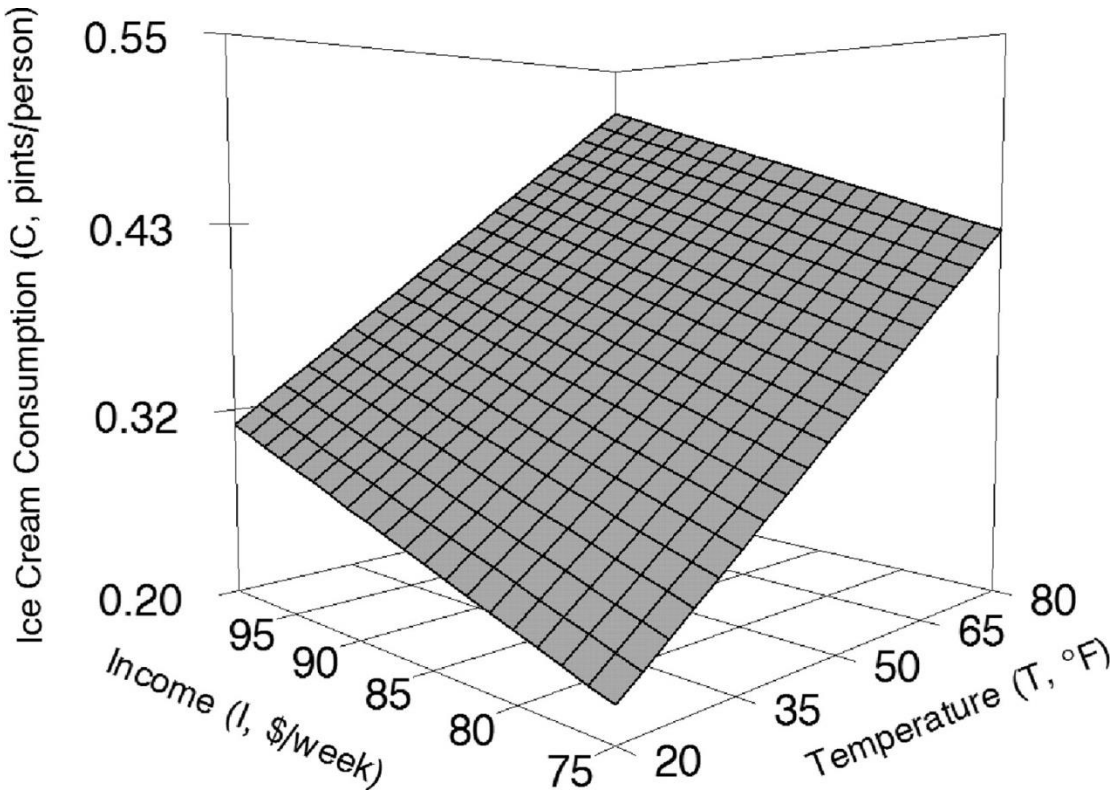
Modelos de Regressão

- Regressão linear (simples) com uma variável explicativa



Modelos de Regressão

- Regressão linear múltipla (mais de uma variável explicativa)



Modelos de Regressão

- Previsão da qualidade do vinho
- Equação de Ashenfelter
- Essa equação corretamente previu os “vinhos do século” para as safras de 1989 e 1990

Wine quality = 12.145 – 0.00117 x winter rainfall
+ 0.0614 x average growing season temperature
- 0.00386 x harvest rainfall

Modelos de Regressão

- Exemplo: regressão para estudar o consumo de energia elétrica por mês (em kWh) versus características dos domicílios
  - Variável explicada: consumo de energia em kWh por mês do domicílio na amostra
  - Variáveis explicativas: televisores, ferros elétricos, geladeiras, aparelhos de som, número de moradores, indicador de área urbana ou rural, renda mensal per capita do domicílio

Análise econométrica para estudar o padrão de consumo das famílias que possuem determinados itens eletroeletrônicos

Variável no modelo de regressão	Parâmetro estimado	Erro padrão	Significância estatística (p-valor)
Quantidade de televisores	45.408	0.879	<.0001
Quantidade de ferros elétricos	17.409	1.422	<.0001
Quantidade de geladeiras	22.141	2.063	<.0001
Quantidade de aparelhos de som	2.598	0.712	0.0003
Número de moradores	13.552	0.426	<.0001
Indicador de área urbana	17.162	2.270	<.0001
Renda mensal per capita	0.0178	0.001	<.0001

Modelos de Regressão

- Estimação dos coeficientes desconhecidos $\beta_0, \beta_1, \beta_2, \ldots, \beta_k$ para cada variável no modelo de regressão
  - Considere um conjunto específico de valores para $\beta_0, \beta_1, \beta_2, \ldots, \beta_k$
  - Com base nesses valores, podemos calcular o valor previsto para a variável resposta

$$\hat{y}_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_k x_{ki}$$

  - O erro de previsão é dado por

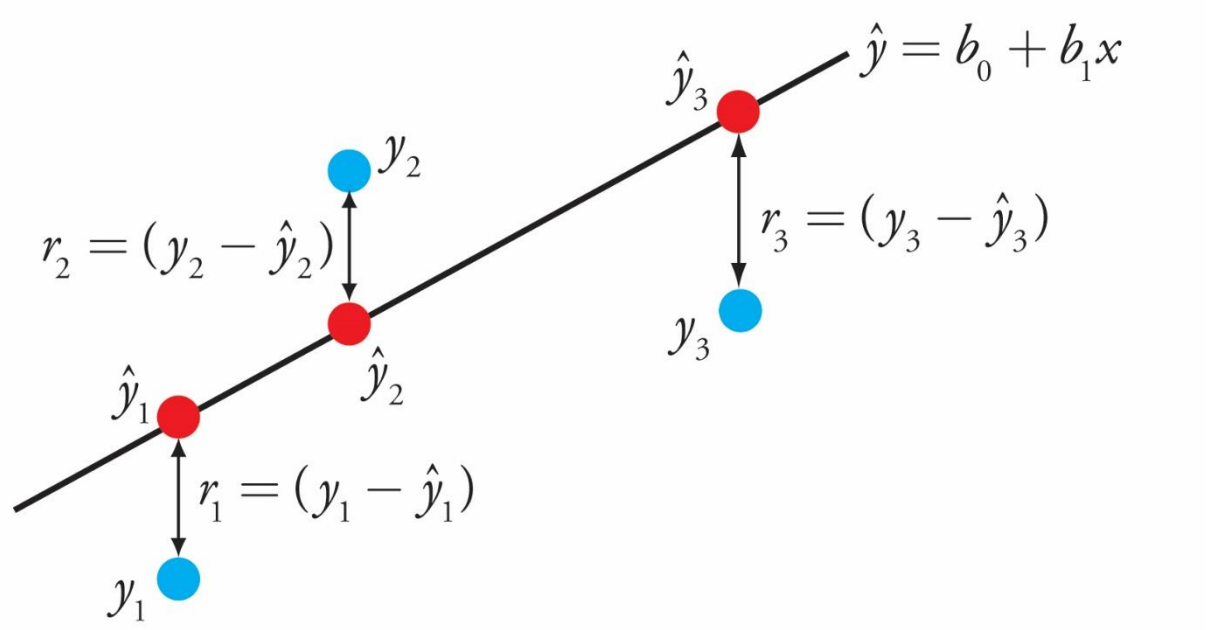
$$r_i = y_i - \hat{y}_i$$

  - A soma dos erros de previsão ao quadrado para todos as observações na amostra é dada por

$$SQE = \sum_{i=1}^{n} [y_i - \hat{y}_i]^2$$

  - Posemos então escolher um conjunto de valores dos coeficientes $\beta_0, \beta_1, \beta_2, \ldots, \beta_k$ de forma a minimizar a soma dos erros quadráticos SQE

Modelos de Regressão



- O método de estimação dos coeficientes $\beta_0, \beta_1, \beta_2, \ldots, \beta_k$ pela minimização da soma dos erros ao quadrado é conhecido **método de mínimos quadrados ordinário** (OLS)

Modelos de Regressão

- Estimação dos coeficientes desconhecidos $\beta_0, \beta_1, \beta_2, \ldots, \beta_k$ via método de mínimos quadrados ordinário possui fórmula fechada a partir da amostra observada

  - No caso de uma única variável explicativa, $y_i = \beta_0 + \beta_1 x_{1i} + \epsilon_i$, podemos estimar β_1 com a expressão

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \dot{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \dot{x})^2}$$

  - O coeficiente β_0 é estimado por

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \dot{x}$$

  - Para o caso de mais de uma variável explicativa, há fórmulas matriciais bem diretas para o cálculo dos coeficientes estimados

  - Na prática, os coeficientes da regressão linear podem ser estimados via qualquer software estatístico ou via planilhas eletrônicas do tipo Excel

Modelos de Regressão

- Coeficiente de determinação (R^2) da regressão

  - Para um modelo de regressão qualquer estimado, é interessante termos uma medida do ajuste dessa regressão

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ \ldots + \beta_k x_{ki} + \ \epsilon_i$$

  - A medida de ajuste está intrinsicamente ligada à importância do termo ϵ_i. Esse termo corresponde à parcela da variável explicada y_i que não é explicada pelas variáveis independentes

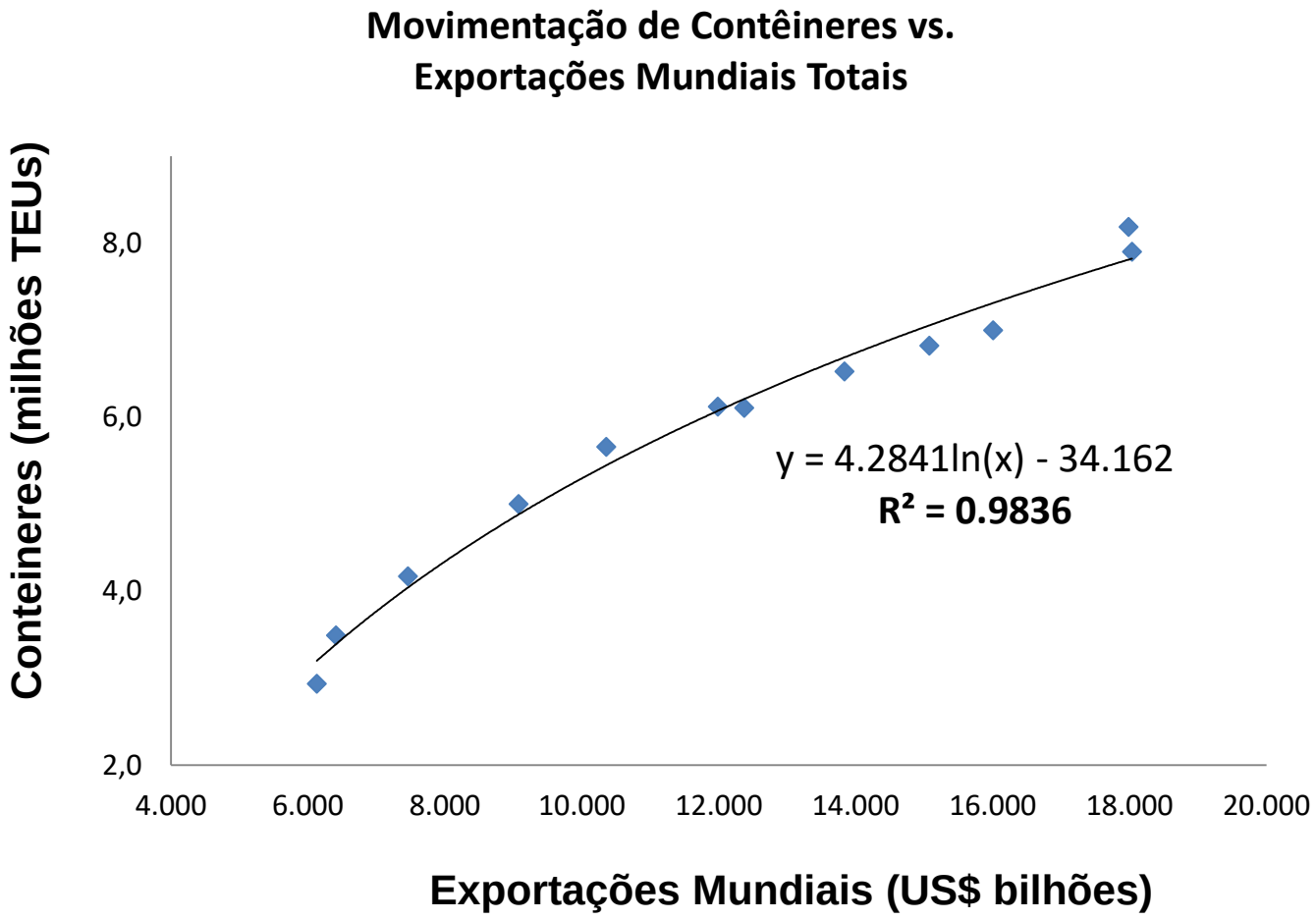
  - A medida mais comumente utilizada para verificar o ajuste de uma regressão é chamada coeficiente de determinação, que é calculada pela expressão:

$$R^2 = \left[1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}\right] = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

  - Pode-se mostrar que o coeficiente de determinação varia entre 0 e 1 (dado que o intercepto está incluído na regressão)

  - O coeficiente de determinação pode ser interpretado como o percentual da variação da variável predita que é explicado pela regressão

Modelos de Regressão



Modelos de Regressão

- Intepretação do coeficiente de determinação
    - Percentual da variação da variável dependente que pode ser explicado pela variação das variáveis independentes
- Cuidado: quando incluímos variáveis na equação, independente de essas fazerem sentido ou não, o R^2 sempre aumenta
- Alternativa para "avaliar" a inclusão da nova variável: **R^2 ajustado**

$$R^2_{ajustado} = 1 - \left[\frac{(1 - R^2)(n - 1)}{n - k - 1}\right]$$

- *n* é o número de observações na amostra
- *k* é o número de variáveis explicativas (sem considerar a constante)
- Quando incluímos variáveis 'desnecessárias' na regressão, o R^2 ajustado diminui

Modelos de Regressão

- Tipos de dados utilizados:
  - Dados *cross-section* – um instante específico no tempo
  - Dados de séries temporais – observações sequenciais ao longo do tempo (exemplo, séries trimestrais de PIB, séries mensais de índices de preço etc.)
  - Dados de painel – dados por unidades *cross-section*, observados em vários momentos do tempo
- Outros tipos de dados:
  - Dados espaciais
    - Dados por polígonos – exemplo, municípios ou setores censitários
    - Dados em pontos específicos – por exemplo, locais de assaltos
  - Microdados – exemplos, registros administrativos
- Dependendo do tipo de dados, há técnicas específicas
  - Tratamento específico de estrutura de correlações entre resíduos ϵ_i

Modelos de Regressão

- **Exercício para entregar em 2 semanas:**
  - Como de costume, os exercícios podem ser entregues em grupos de 2 ou três alunos, e o grupo deve submeter o código em R utilizado para responder ao exercício, juntamente com a discussão dos resultados
  - Utilize como base o código em R 'Analise_de_Regressao_Linear_Exercicios_Praticos_1'
  - Rode a regressão de acordo com o modelo abaixo:

```
mod1.ex <- lm(dados3$mort_infantil ~ dados3$renda_per_capita
                                   + dados3$indice_gini
                                   + dados3$salario_medio_mensal
                                   + dados3$perc_criancas_extrem_pobres
                                   + dados3$perc_criancas_pobres
                                   + dados3$perc_pessoas_dom_agua_estogo_inadequados
                                   + dados3$perc_pessoas_dom_paredes_inadequadas
                                   + dados3$perc_pop_dom_com_coleta_lixo)
```

Modelos de Regressão

- **Exercício para entregar em 2 semanas (continuação):**
  - Questão 1: No modelo anterior, quais as variáveis explicativas e qual a variável dependente?
  - Questão 2: Os coeficientes encontrados estão com os sinais de acordo com o esperado?
  - Questão 3: Qual o percentual da variabilidade da mortalidade infantil que é explicada pelas variáveis explicativas?
  - Questão 4: Utilizando o comando abaixo, crie a variável 'perc_pop_rural', indicando o percentual do município que vive em domicílios na zona rural. Adicione essa variável ao modelo de regressão. Com base no coeficiente estimado, "controlando-se" para as variáveis já presentes no modelo, qual o efeito da localização na zona rural sobre a taxa de mortalidade infantil?

```
dados3$perc_pop_rural <- dados3$populacao_rural / dados3$populacao_total
```

  - Questão 5: Com a inclusão da nova variável, o que aconteceu com o coeficiente de determinação e com o R^2 ajustado?
  - Questão 6: Os dados utilizados para essa regressão são dados do tipo *cross-section*, do tipo séries de tempo ou do tipo dados de painel?

Modelos de Regressão

- Inclusão de variáveis qualitativas como variáveis preditoras
- Exemplo, queremos ver como a renda per capita dos municípios é afetada pela região na qual o município se localiza
- Temos cinco regiões: NO, SU, SE, NE, CO
- Precisamos transformar a informação qualitativa em informações quantitativas
- Maneira comumente utilizada:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ \ldots + \beta_k x_{ki}$$
$$+\delta_1 D_{SU} + \delta_2 D_{NO} + \delta_3 D_{SE} + \delta_4 D_{NE} + \delta_5 D_{CO} + \epsilon_i$$

- Na equação acima, novos parâmetros a serem estimados: $\delta_1, \delta_2, \delta_3, \delta_4, \delta_5$
- As variáveis $D_{SU}, D_{SE}, D_{NO}, D_{NE}, D_{CO}$ são chamadas **variáveis *dummy***

Modelos de Regressão

- As variáveis *dummy* $D_{SU}, D_{SE}, D_{NO}, D_{NE}, D_{CO}$ são definidas como:
  - Caso o município da observação i esteja contido na região Sul, então o valor de D_{SU}=1, e os valores das demais variáveis dummy será zero
  - Caso o município da observação i esteja contido na região Nordeste, então o valor de D_{NE}=1, e os valores das demais variáveis dummy será zero
  - E assim por diante ...
- Problema: não podemos incluir todas as variáveis dummy na regressão ao mesmo tempo, adicionalmente ao intercepto β_0
  - (1) Se mantivermos o intercepto, temos que retirar uma das variáveis dummy
  - (2) Se mantivermos todas as variáveis dummy, precisamos retirar o intercepto
- Observação:
  - Essas exclusões alternativas são feitas para evitarmos problemas de **multicolinearidade** (perfeita)
  - Dependendo da alternativa (1) ou (2) acima, a interpretação dos parâmetros muda

Modelos de Regressão

- Especificações alternativas:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ \ldots + \beta_k x_{ki}$$
$$+\delta_1 D_{SU} + \delta_2 D_{NO} + \delta_3 D_{SE} + \delta_4 D_{NE} + \epsilon_i$$

- Ou:

$$y_i = \beta_1 x_{1i} + \beta_2 x_{2i} + \ \ldots + \beta_k x_{ki}$$
$$+\delta_1 D_{SU} + \delta_2 D_{NO} + \delta_3 D_{SE} + \delta_4 D_{NE} + \delta_5 D_{CO} + \epsilon_i$$

- Para fins de previsão, as duas especificações retornam resultados idênticos
- É possível incluir mais de uma variável qualitativa na regressão, sempre atentando para problemas de multicolinearidade
- A primeira especificação é mais utilizada

Modelos de Regressão

- Exemplo:

$$[Salário]_i = \beta_0 + \beta_1[Experiência]_i + \delta_1[DummyMulher]_i + \epsilon_i$$

- Para um indivíduo do sexo masculino, a equação se torna:

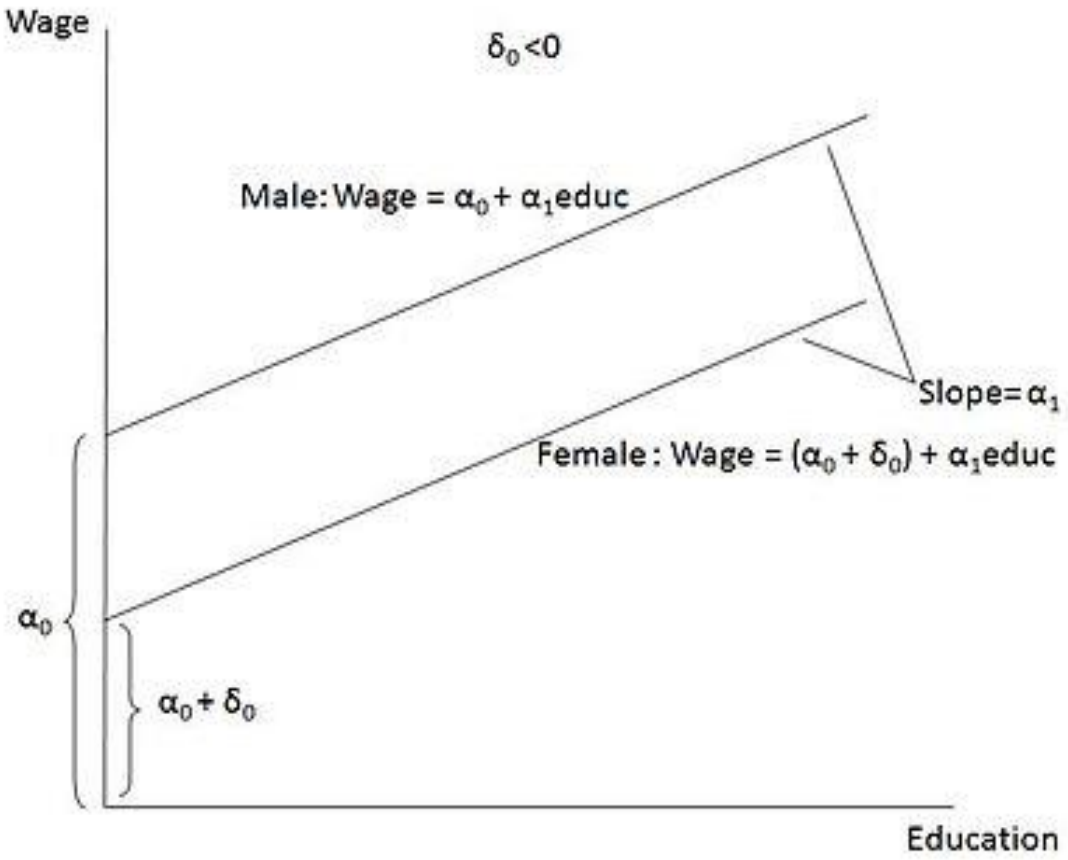
$$[Salário]_i = \beta_0 + \beta_1[Experiência]_i + \epsilon_i$$

- Para um indivíduo do sexo feminino, a equação se torna:

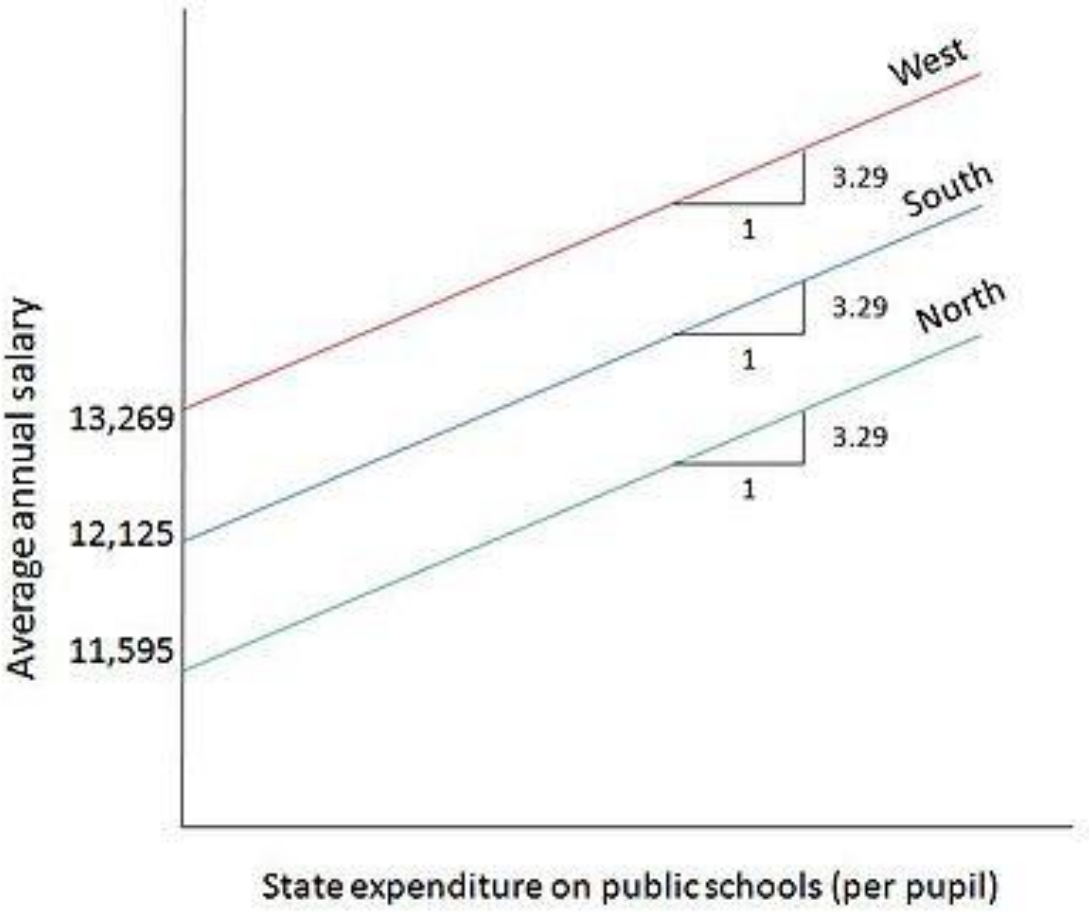
$$[Salário]_i = (\beta_0 + \delta_1) + \beta_1[Experiência]_i + \epsilon_i$$

- Note que a diferença está no **intercepto**: β_0 para os homens e $(\beta_0 + \delta_1)$ para as mulheres

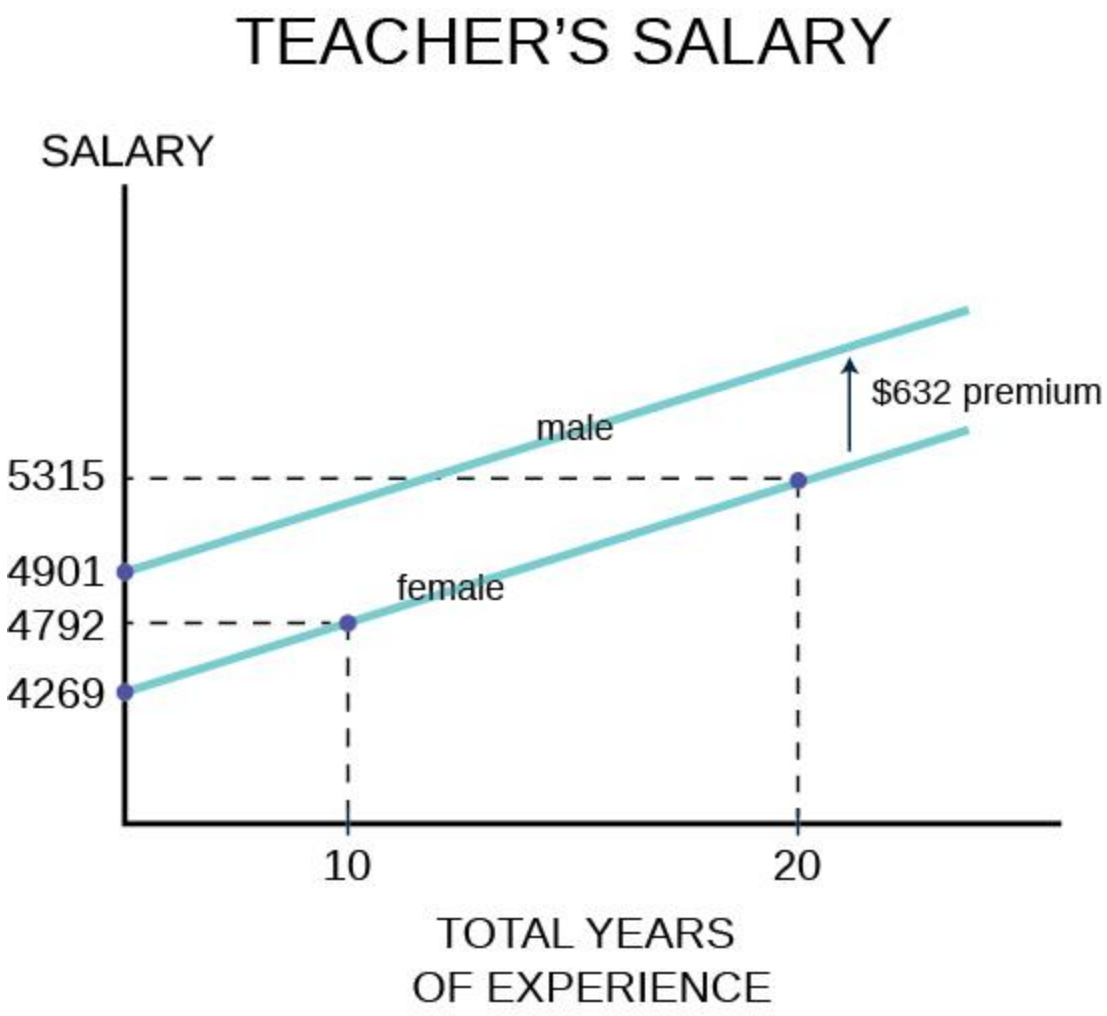
Modelos de Regressão



Modelos de Regressão

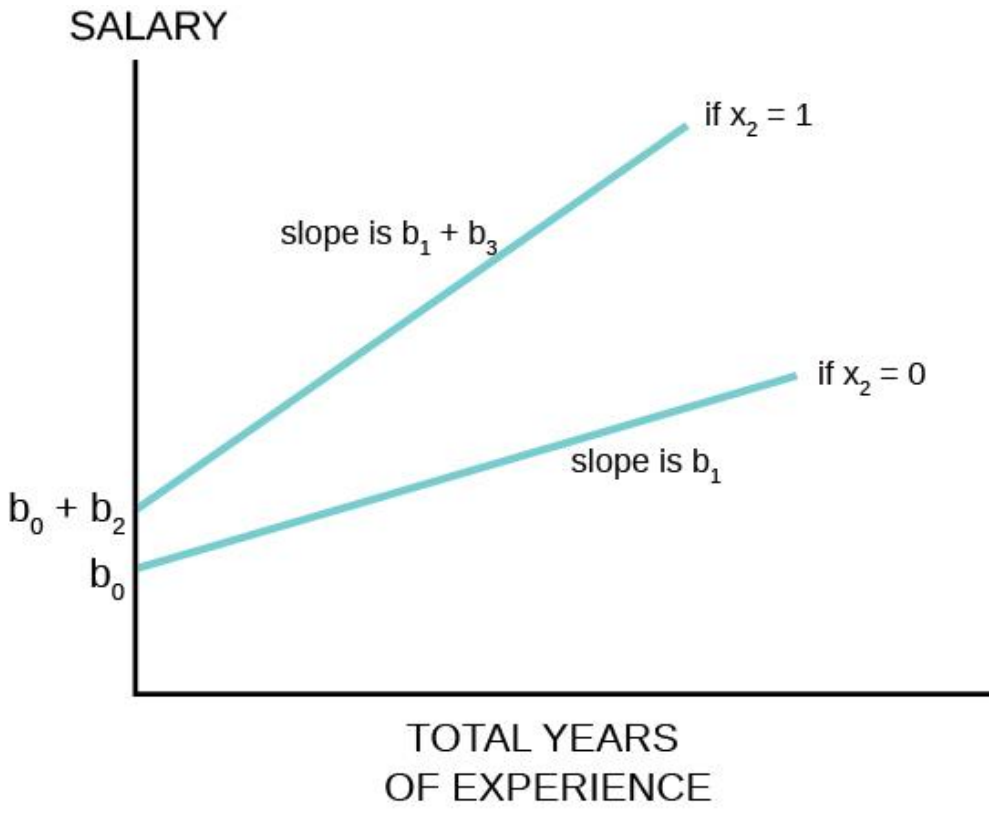


Modelos de Regressão



Modelos de Regressão

E se quisermos que o coeficiente dos anos de experiência variem entre homens e mulheres?



$$\hat{y} = b_0 + b_2x_2 + b_1x_1 + b_3x_2x_1$$

Modelos de Regressão

- Alteração no modelo com dummy para mulheres:

$$[Salário]_i = \beta_0 + \beta_1[Experiência]_i + \delta_1[DummyMulher]_i + \gamma_1[DummyMulher]_i \times [Experiência]_i + \epsilon_i$$

- Para um indivíduo do sexo masculino, a equação se torna:

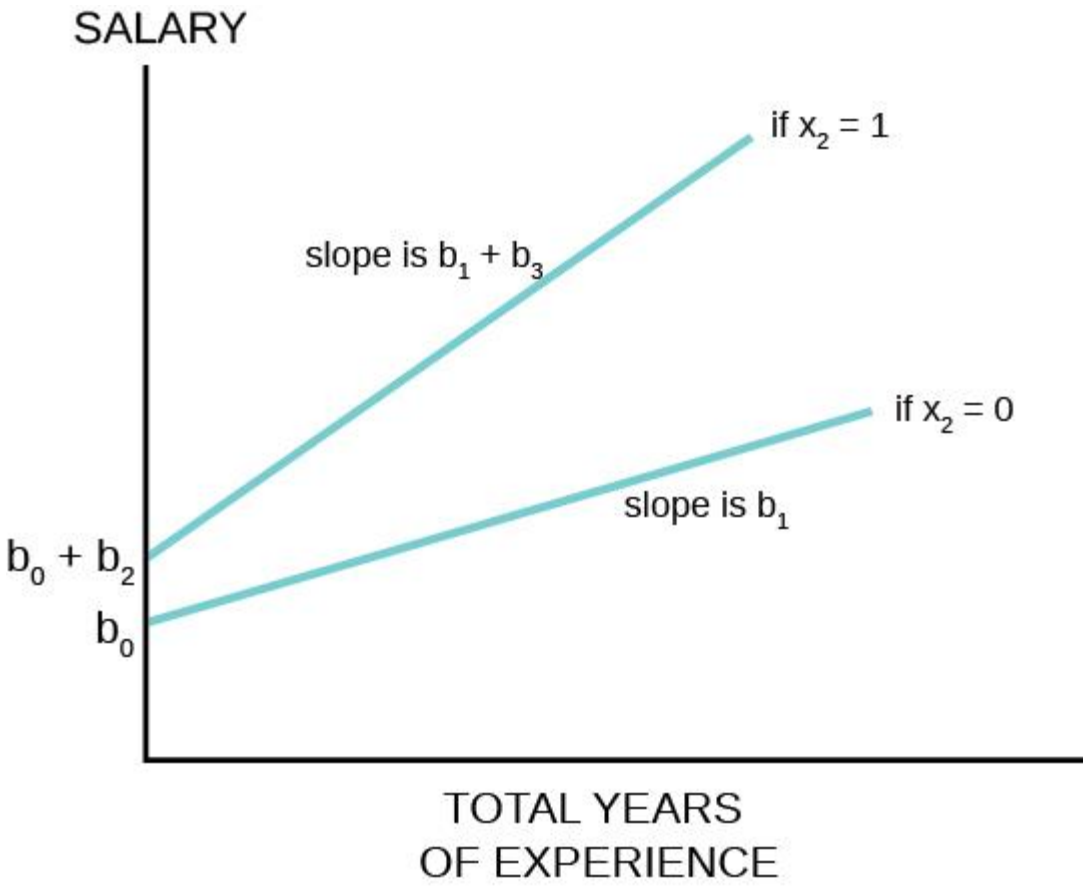
$$[Salário]_i = \beta_0 + \beta_1[Experiência]_i + \epsilon_i$$

- Para um indivíduo do sexo feminino, a equação se torna:

$$[Salário]_i = (\beta_0+\delta_1) + (\beta_1+\gamma_1)\,[Experiência]_i + \epsilon_i$$

- Diferença no **intercepto**: β_0 para os homens e $(\beta_0+\delta_1)$ para as mulheres
- Diferença no **coeficiente** da variável anos de experiência: β_1 versus $(\beta_1+\gamma_1)$

Modelos de Regressão



$$\hat{y} = b_0 + b_2x_2 + b_1x_1 + b_3x_2x_1$$

Modelos de Regressão

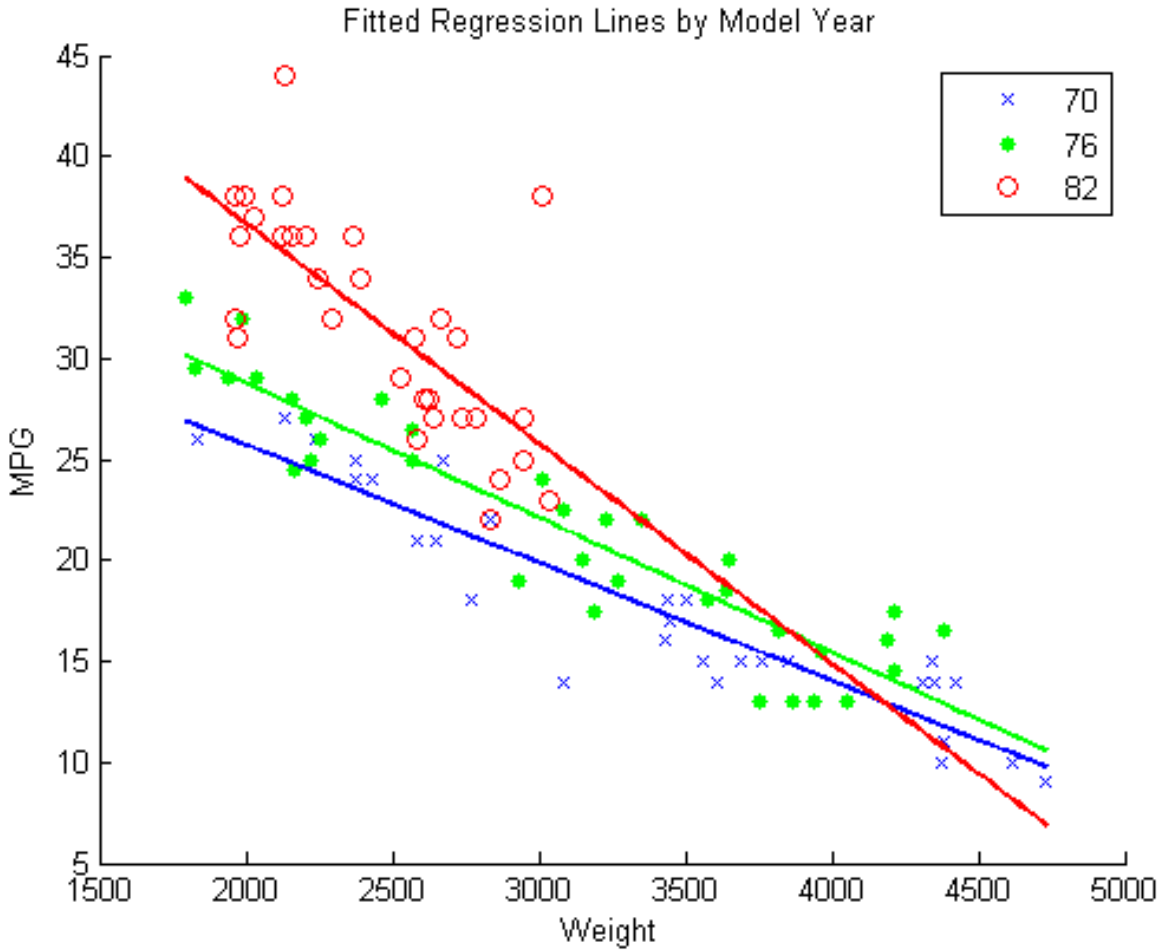
- Inclusão de variáveis qualitativas como variáveis preditoras, afetando também os coeficientes de variáveis quantitativas
- Alteração apenas no intercepto da regressão, sem afetar os coeficientes de outras variáveis:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ \ldots + \beta_k x_{ki}$$
$$+\boldsymbol{\delta_1 D_{SU}} + \boldsymbol{\delta_2 D_{NO}} + \boldsymbol{\delta_3 D_{SE}} + \boldsymbol{\delta_4 D_{NE}} + \epsilon_i$$

- Lembrando que excluímos a dummy do Centro-Oeste (região de "referência")
- Alteração afetando os coeficientes das outras variáveis

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ \ldots + \beta_k x_{ki}$$
$$+\delta_1 D_{SU} + \delta_2 D_{NO} + \delta_3 D_{SE} + \delta_4 D_{NE}$$
$$+\boldsymbol{\gamma_1[D_{SU} \times x_{1i}]} + \boldsymbol{\gamma_2[D_{NO} \times x_{1i}]} + \boldsymbol{\gamma_3[D_{SE} \times x_{1i}]} + \boldsymbol{\gamma_4[D_{NE} \times x_{1i}]} \ + \epsilon_i$$

Modelos de Regressão



Modelos de Regressão

- **Exercício para entregar em 2 semanas (continuação):**
  - Rode agora a regressão com efeitos das Regiões sobre a mortalidade infantil:

```
mod2.ex <- lm(dados3$mort_infantil ~ dados3$renda_per_capita
        + dados3$indice_gini
        + dados3$salario_medio_mensal
        + dados3$perc_criancas_extrem_pobres
        + dados3$perc_criancas_pobres
        + dados3$perc_pessoas_dom_agua_estogo_inadequados
        + dados3$perc_pessoas_dom_paredes_inadequadas
        + dados3$perc_pop_dom_com_coleta_lixo
        + dados3$perc_pop_rural
        + as.factor(dados3$Regiao))
```

  - Questão 7: Com base nos resultados dessa nova equação, qual o efeito das regiões Norte, Sul, Nordeste e Sudeste, mesmo depois de “controlarmos” para as variáveis incluídas no modelo?

Modelos de Regressão

- **Exercício para entregar em 2 semanas (continuação):**
  - Rode agora a regressão com efeitos das Regiões sobre a mortalidade infantil:

```
mod3.ex <- lm(dados3$mort_infantil ~ dados3$renda_per_capita
        + dados3$indice_gini
        + dados3$salario_medio_mensal
        + dados3$perc_criancas_extrem_pobres
        + dados3$perc_criancas_pobres
        + dados3$perc_pessoas_dom_agua_estogo_inadequados
        + dados3$perc_pessoas_dom_paredes_inadequadas
        + dados3$perc_pop_dom_com_coleta_lixo
        + dados3$perc_pop_rural
        + as.factor(dados3$Regiao)
        + as.factor(dados3$Regiao)*dados3$renda_per_capita)
```

  - Questão 8: Com base nos resultados dessa nova equação, como o efeito da renda per capita, sobre mortalidade infantil, se altera de acordo com a macrorregião do município?
  - Questão 9: Houve uma melhora no R2 ajustado quando adicionamos os efeitos das macrorregiões sobre o coeficiente da renda per capita (mod3 versus mod2)?

Modelos de Regressão

- **Exercício para entregar em 2 semanas (continuação):**
  - Vamos incluir agora uma interação entre a macrorregião e a variável "perc_pessoas_dom_agua_estogo_inadequados":

```
mod3.ex <- lm(dados3$mort_infantil ~ dados3$renda_per_capita
        + dados3$indice_gini
        + dados3$salario_medio_mensal
        + dados3$perc_criancas_extrem_pobres
        + dados3$perc_criancas_pobres
        + dados3$perc_pessoas_dom_agua_estogo_inadequados
        + dados3$perc_pessoas_dom_paredes_inadequadas
        + dados3$perc_pop_dom_com_coleta_lixo
        + dados3$perc_pop_rural
        + as.factor(dados3$Regiao)
        + as.factor(dados3$Regiao)*dados3$renda_per_capita
        + as.factor(dados3$Regiao)*dados3$perc_pessoas_dom_agua_estogo_inadequados)
```

  - Questão 10: Vamos assumir que a variável "perc_pessoas_dom_agua_estogo_inadequados" seja uma variável direta de política pública. De acordo com os resultados da regressão acima, em qual região políticas de melhoria do acesso a água e esgoto seriam mais eficazes para reduzir a mortalidade infantil?

Modelos de Regressão

- Vimos acima como flexibilizar a especificação da regressão linear, para incluir o efeito de variáveis qualitativas
- Podemos também incluir termos polinomiais para flexibilizar a forma funcional da regressão
- Podemos incluir, por exemplo, termo quadrático da variável x_{1i} para capturar alguma não-linearidade na equação:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ \ldots + \beta_k x_{ki} + \boldsymbol{\beta_{(k+1)} x_{1i}^2} + \epsilon_i$$

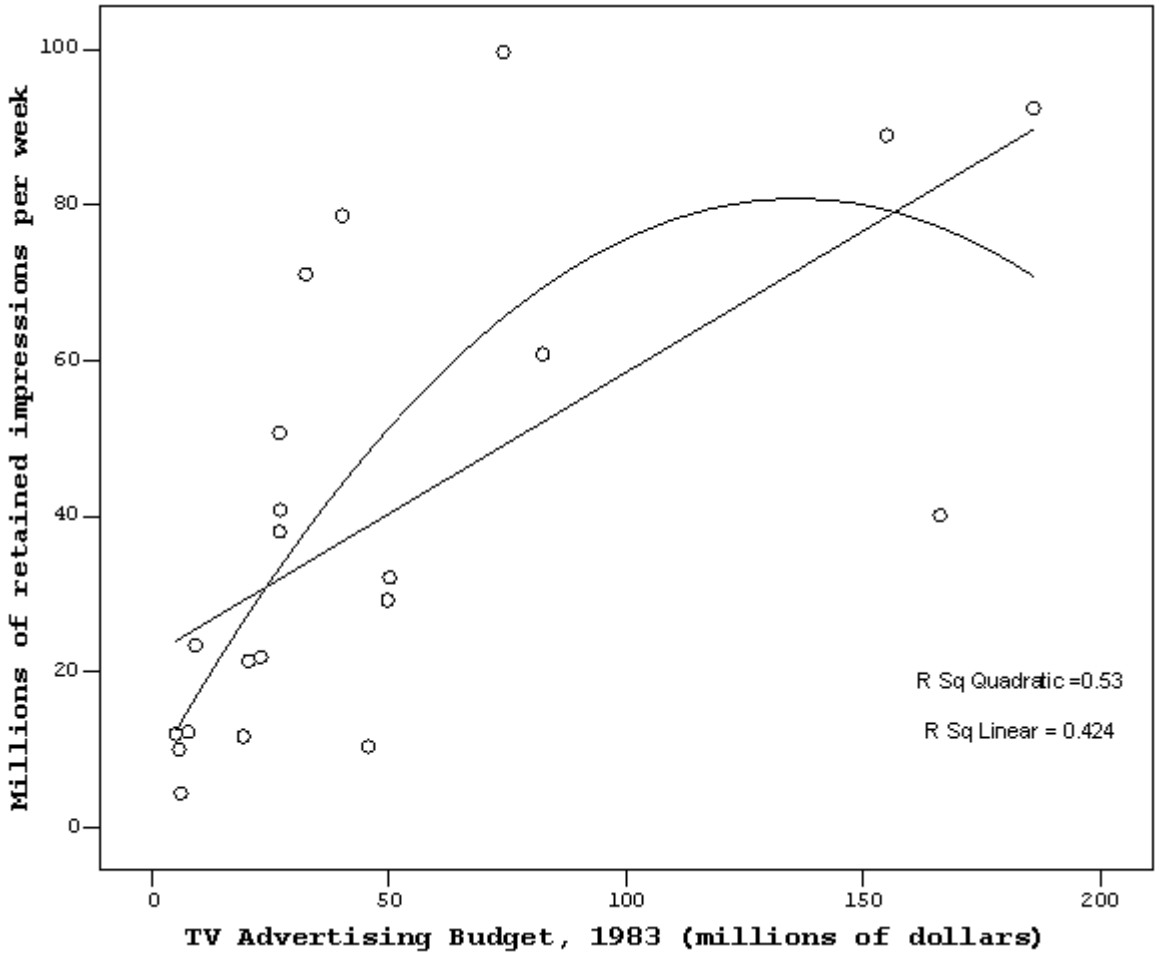
- O mesmo pode ser feito para outras variáveis na equação

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ \ldots + \beta_k x_{ki} + \boldsymbol{\beta_{(k+1)} x_{1i}^2} + \boldsymbol{\beta_{(k+2)} x_{2i}^2} + \epsilon_i$$

- Podemos incluir também termos de ordem maior que 2:

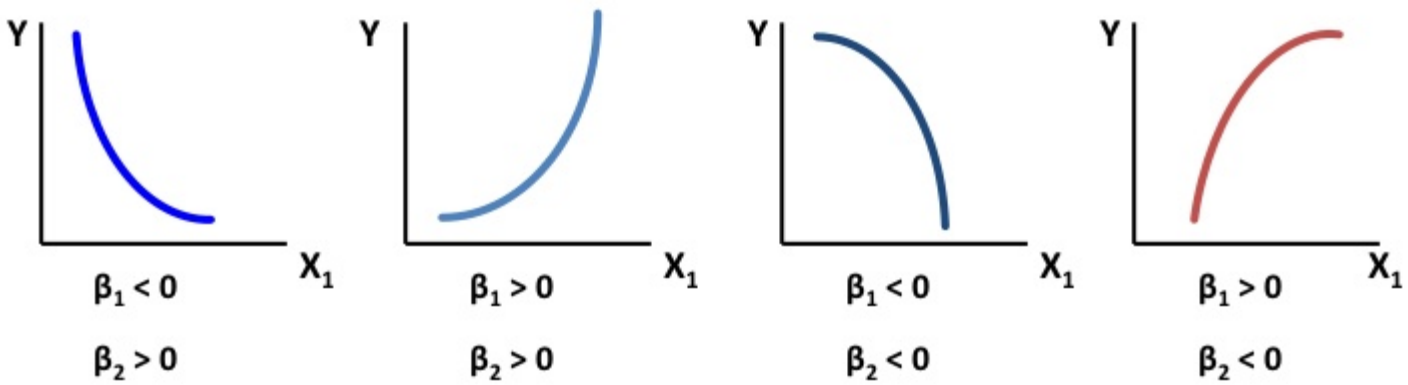
$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ \ldots + \beta_k x_{ki} + \boldsymbol{\beta_{(k+1)} x_{1i}^2} + \boldsymbol{\beta_{(k+2)} x_{1i}^3} + \epsilon_i$$

Modelos de Regressão



Modelos de Regressão

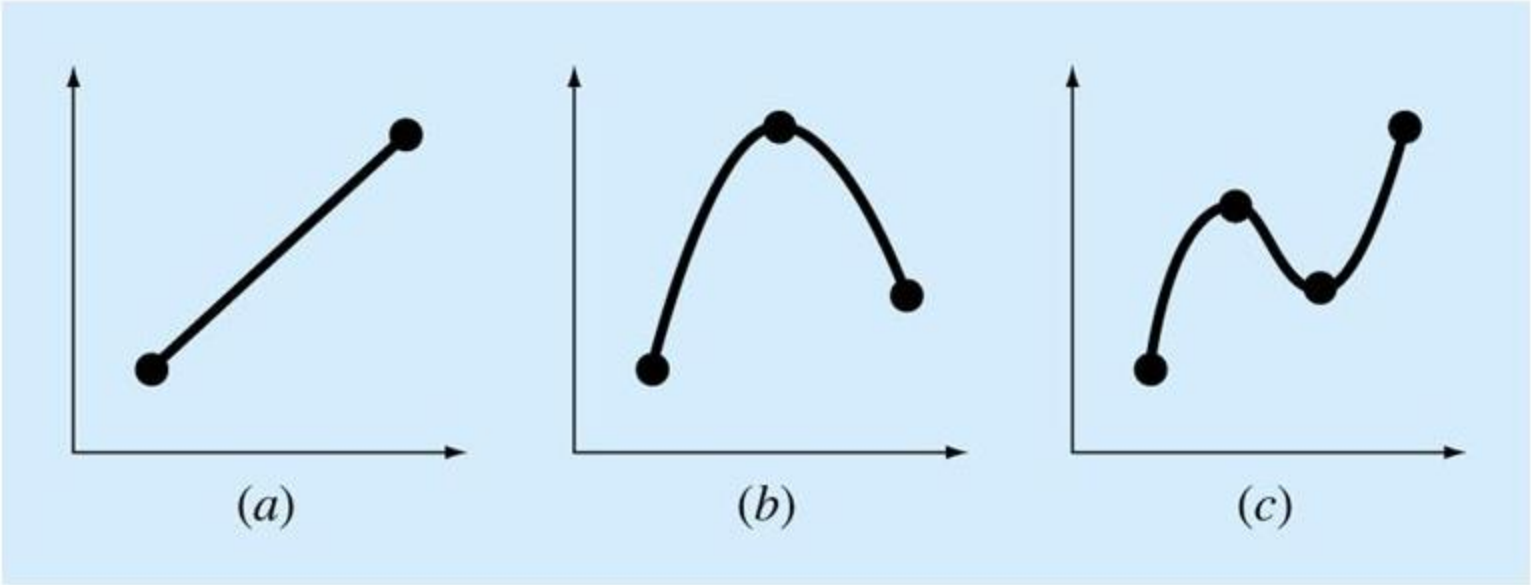
$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{1i}^2 + \varepsilon_i$$



β_1 = the coefficient of the linear term
β_2 = the coefficient of the squared term

Modelos de Regressão

$$f(x) = a_0 + a_1 x + a_2 x^2 + \cdots + a_n x^n$$



Modelos de Regressão

- Inclusão de um termo quadrático para a renda per capita, utilizando o R:

```
mod1a.ex <- lm(dados3$mort_infantil ~ dados3$renda_per_capita
        + I(renda_per_capita^2)
        + dados3$indice_gini
        + dados3$salario_medio_mensal
        + dados3$perc_criancas_extrem_pobres
        + dados3$perc_criancas_pobres
        + dados3$perc_pessoas_dom_agua_estogo_inadequados
        + dados3$perc_pessoas_dom_paredes_inadequadas
        + dados3$perc_pop_dom_com_coleta_lixo, data = dados)

summary(mod1a.ex)
```

Modelos de Regressão

```
Call:
lm(formula = dados3$mort_infantil ~ dados3$renda_per_capita +
    I(renda_per_capita^2) + dados3$indice_gini + dados3$salario_medio_mensal +
    dados3$perc_criancas_extrem_pobres + dados3$perc_criancas_pobres +
    dados3$perc_pessoas_dom_agua_estogo_inadequados + dados3$perc_pessoas_dom_paredes_inadequadas +
    dados3$perc_pop_dom_com_coleta_lixo, data = dados)

Residuals:
     Min       1Q   Median       3Q      Max
-17.7662  -2.4401  -0.3568   1.8741  20.0708

Coefficients:
                                              Estimate Std. Error t value Pr(>|t|)
(Intercept)                                  2.612e+01  9.728e-01  26.853  < 2e-16 ***
dados3$renda_per_capita                     -2.344e-02  1.864e-03 -12.576  < 2e-16 ***
I(renda_per_capita^2)                        1.158e-05  9.274e-07  12.489  < 2e-16 ***
dados3$indice_gini                          -6.568e+00  1.377e+00  -4.771 1.88e-06 ***
dados3$salario_medio_mensal                 -1.753e-01  9.385e-02  -1.868  0.06182 .
dados3$perc_criancas_extrem_pobres           3.168e-02  1.201e-02   2.638  0.00837 **
dados3$perc_criancas_pobres                  1.171e-01  1.381e-02   8.475  < 2e-16 ***
dados3$perc_pessoas_dom_agua_estogo_inadequados  4.129e-02  5.985e-03   6.899 5.81e-12 ***
dados3$perc_pessoas_dom_paredes_inadequadas  3.644e-02  7.833e-03   4.651 3.37e-06 ***
dados3$perc_pop_dom_com_coleta_lixo         -2.641e-03  6.441e-03  -0.410  0.68183
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.983 on 5554 degrees of freedom
Multiple R-squared:  0.6891,    Adjusted R-squared:  0.6886
F-statistic:  1368 on 9 and 5554 DF,  p-value: < 2.2e-16
```

Intervalos de Confiança
e
Testes de Hipóteses

Simulações de Monte Carlo

- Avaliação do Programa Saúde da Família (PSF)

- Coletamos informações de municípios entre 2010 e 2013, incluindo

    - Dados sócio econômicos dos municípios
    - Cobertura do programa saúda de família em 2010 (percentual da população do município atendida)
    - Redução (por número de habitantes) do número de óbitos relativos a doenças que podem ser evitadas com o PSF; redução entre 2010 e 2013

- Queremos estudar o impacto do programa saúde da família na redução do número de óbitos. Queremos estudar se o impacto foi positivo, no sentido de que maior cobertura em 2010 incorreu em maior redução entre 2010 e 2013

- Rodamos uma regressão:

    - Variável dependente = redução do número de óbitos

    - Variáveis independentes:
        - Cobertura do PSF em 2010 (variável de interesse)
        - Informações socioeconômicas dos municípios (covariáveis)

Simulações de Monte Carlo

- Resultado da regressão com dados “fictícios”

```
mod1.psf <- lm(dados$reducao_obitos ~ dados$cobertura_psf
                                    + dados$renda_per_capita
                                    + dados$indice_gini
                                    + dados$perc_criancas_extrem_pobres
                                    + dados$perc_criancas_pobres
                                    + dados$perc_pessoas_dom_agua_estogo_inadequados
                                    + dados$perc_pessoas_dom_paredes_inadequadas
                                    + dados$perc_pop_dom_com_coleta_lixo)
summary(mod1.psf)
```

- Resultados:

```
Coefficients:
                                              Estimate Std. Error t value Pr(>|t|)
(Intercept)                                  3.261e-03  3.503e-03   0.931    0.352
dados$cobertura_psf                         -1.316e-03  9.437e-04  -1.394    0.163
dados$renda_per_capita                       2.991e-04  2.404e-06 124.432  < 2e-16 ***
dados$indice_gini                           -2.411e-02  5.262e-03  -4.582 4.70e-06 ***
dados$perc_criancas_extrem_pobres            4.697e-04  5.137e-05   9.143  < 2e-16 ***
dados$perc_criancas_pobres                  -3.136e-04  4.832e-05  -6.489 9.40e-11 ***
dados$perc_pessoas_dom_agua_estogo_inadequados -1.681e-04  2.543e-05  -6.611 4.18e-11 ***
dados$perc_pessoas_dom_paredes_inadequadas   -3.019e-04  3.348e-05  -9.017  < 2e-16 ***
dados$perc_pop_dom_com_coleta_lixo           1.785e-04  2.752e-05   6.487 9.50e-11 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.01707 on 5555 degrees of freedom
Multiple R-squared:  0.9533,          Adjusted R-squared:  0.9532
F-statistic: 1.417e+04 on 8 and 5555 DF,  p-value: < 2.2e-16
```

Simulações de Monte Carlo

- Vamos entender na regressão acima os chamados dados “fictícios”
    - Programa: “Analise_de_Regressao_Linear_Testes_de_Hipotese.R”
- Com base nos dados de municípios do IDHM, geramos uma variável nova, que indica artificialmente o nível de cobertura do PSF em 2010 (variável simulada)
- Geramos artificialmente também uma redução da mortalidade, a partir de dados de cobertura do PSF, dados sócioeconômicos, e uma variável de erro da regressão
- O coeficiente de impacto do PSF sobre mortalidade foi artificialmente especificado igual a zero
- Mesmo assim, a regressão resultou em um valor diferente de zero (coeficiente da regressão diferente de zero)
- Como então “separar” quando uma regressão tem coeficiente zero de fato ou tem um coeficiente que podemos considerar diferente de zero?
- Para resolver esse problema, montou-se todo o arcabouço conhecido como ***testes de hipótese***
- Para isso, é importante entender o conceito de ***distribuição do estimador*** para o coeficiente de regressão
- Vamos ver agora as chamadas ***simulações de Monte Carlo***

Simulações de Monte Carlo

- Simulações de Monte Carlo consistem em se simular o processo de geração da amostra aleatória observada por nós
- Em geral, temos apenas uma amostra aleatória
- As simulações de Monte Carlo nos ajudam a entender que tipo de erro estamos incorrendo por estar usando uma amostra (coletada de forma aleatória)
- Importante entender o chamado "Processo Gerador de Dados" (*Data Generating Process* – DGP)
- O DGP especifica como são gerados aleatoriamente os dados que observamos na (única) amostra que temos disponível
- Lembremos então do nosso modelo de regressão linear:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ \ldots + \beta_k x_{ki} + \epsilon_i$$

- A regressão tem o seu erro ou resíduo aleatório ϵ_i, que pode ser causado por um conjunto de fatores. Esse erro pode ser normal ou não (distribuição normal ou não)
- Vamos então simular um conjunto de diferentes amostras a partir da equação acima, para o problema de avaliação do PSF

Simulações de Monte Carlo

- A partir das Simulações de Monte Carlo, podemos estudar:
  - Qual a forma da distribuição dos betas estimados
  - Qual a dispersão dos betas estimados
  - O que acontece com o formato da distribuição dos betas quando o número de observações na amostra aumenta
  - O que acontece com a dispersão dos betas estimados quando o número de observações na amostra aumenta
- Em geral, observamos que, quando o número de observações na amostra aumenta:
  - O histograma dos betas estimados torna-se cada vez mais "normal" (mesmo quando os erros da regressão não são normais)
  - A dispersão dos betas estimados cai
  - O primeiro fato acima deve-se ao chamado Teorema Central do Limite
  - O segundo se deve à chamada Lei dos Grandes Números
  - A dispersão (desvio-padrão) dos betas cai com a raiz quadrada do número de observações na amostra; quando o número observações na amostra é multiplicado por quatro, a dispersão reduz-se à metade
  - A dispersão "estimada" (estimada com apenas uma amostra – acredite, é possível!) dos betas que obteríamos com simulações de Monte Carlo é dada pela coluna chamada de "Erro Padrão" ou "Std. Error"

Simulações de Monte Carlo

- Exemplo: regressão para estudar o consumo de energia elétrica por mês (em kWh) versus características dos domicílios
  - Variável explicada: consumo de energia em kWh por mês do domicílio na amostra
  - Variáveis explicativas: televisores, ferros elétricos, geladeiras, aparelhos de som, número de moradores, indicador de área urbana ou rural, renda mensal per capita do domicílio

Análise econométrica para estudar o padrão de consumo das famílias que possuem determinados itens eletroeletrônicos

Variável no modelo de regressão	Parâmetro estimado	Erro padrão	Significância estatística (p-valor)
Quantidade de televisores	45.408	0.879	<.0001
Quantidade de ferros elétricos	17.409	1.422	<.0001
Quantidade de geladeiras	22.141	2.063	<.0001
Quantidade de aparelhos de som	2.598	0.712	0.0003
Número de moradores	13.552	0.426	<.0001
Indicador de área urbana	17.162	2.270	<.0001
Renda mensal per capita	0.0178	0.001	<.0001

Intervalos de Confiança

- Pesquisas eleitorais: candidato A possui 15% das intenções de voto, com intervalo de confiança de 2%
  - Algumas pesquisas que esse é um intervalo de confiança com nível de cobertura de 95%
  - O que significa isso?
- Em geral, intervalos de confiança podem ser dados pela expressão:

$$[\text{Estimativa pontual}] \pm [\text{quantil da normal padronizada ou } t - Student] \times [\textbf{Erro Padrão}]$$

- A introdução aos intervalos de confiança foi coberta no curso de estatística básica
- O que são os quantis da distribuição normal?
- Por que utilizar os quantis da distribuição *t* ao invés dos da distribuição normal padronizada?
  - Nos livros antigos, utilizam-se quantis da *t*-Student para n <= 30 e da normal padronizada para n > 30
  - Aconselhável sempre se utilizar os quantis da distribuição *t*-Student (para n muito alto, os quantis se equivalem), pois a distribuição *t*-Student converge para uma distribuição normal quando a amostra aumenta
  - Todos os softwares estatísticos já dão os intervalos de confiança e testes de hipótese com base na *t*-Student

Intervalos de Confiança na Prática

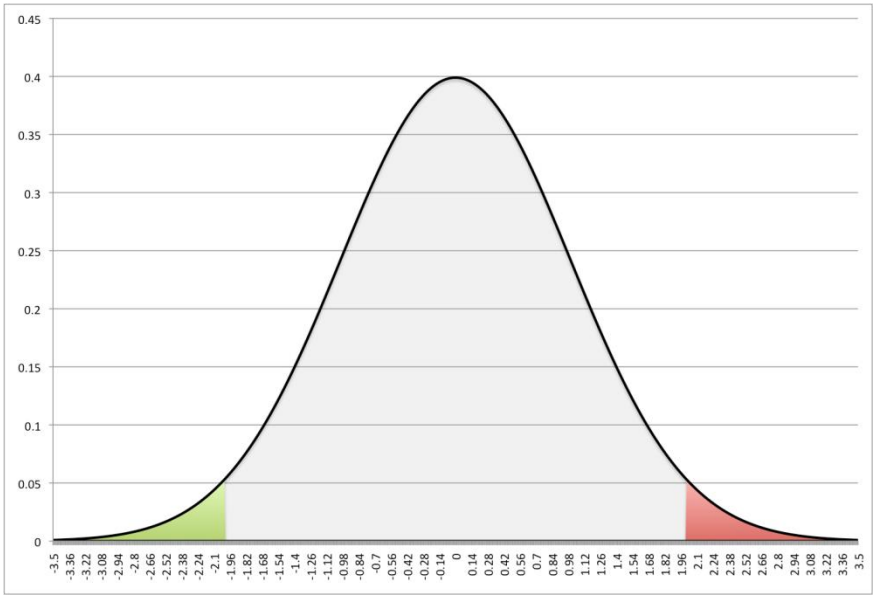
- Os valores 1.96, 1.645 e 2.58 são extraídos dos quantis de uma **distribuição normal padronizada** (distribuição normal com média 0 e variância igual a 1)

No R:

1.645 = -qnorm(0.95, 0, 1)
1.96 = -qnorm(0.975, 0, 1)
2.58 = -qnorm(0.995, 0, 1)

Ou simplesmente:

1.645 = -qnorm(0.95)
1.96 = -qnorm(0.975)
2.58 = -qnorm(0.995)



- Quando a amostra não possui muitas observações (por exemplo, < 30 observações), pode-se refinar a aproximação utilizando-se a distribuição *t-Student*, ao invés da distribuição normal.

- A distribuição *t-Student* também é simétrica, com média 0, mas possui caudas mais pesadas do que a distribuição normal.

Intervalos de Confiança na Prática

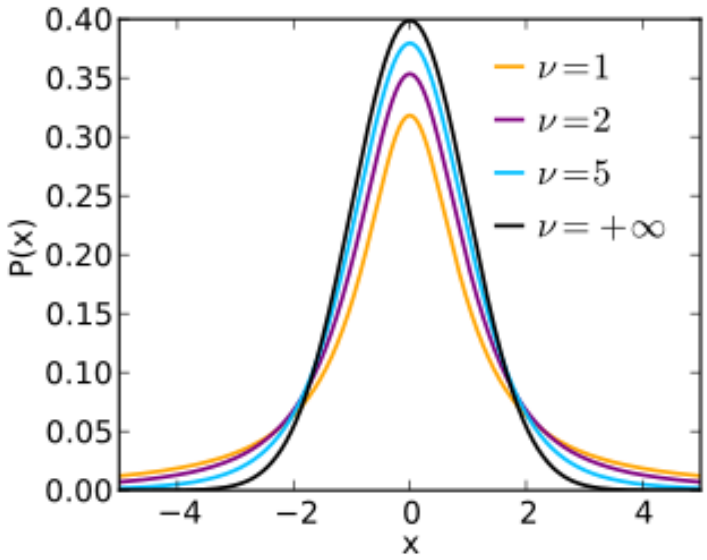
- A distribuição *t-Student* possui um parâmetro chamado de **número de graus de liberdade**, representado pela letra grega ν

- Quanto maior o valor de ν, mais leves são as caudas da distribuição *t-Student* e mais ela se aproxima de uma distribuição normal padronizada (média 0 e variância 1) - quando $\nu \rightarrow \infty$, converge para uma normal padronizada

- Na prática, para amostras com poucas observações (menos de 30 ou 40), aconselha-se utilizar os quantis da distribuição *t-Student*, com número de graus de liberdade igual a $\nu = n - 1 - k$, onde *n* é o número de observações na amostra, e *k* é o número de variáveis explicativas

- Para um intervalo de confiança com probabilidade de cobertura *p*, os quantis são dados no R por:

```
p <- 0.95;
q1 <- -qnorm((1-p)/2);          #---- normal
q2 <- -qt((1-p)/2, df=19);      #---- n = 20
q3 <- -qt((1-p)/2, df=29);      #---- n = 30
q4 <- -qt((1-p)/2, df=10000);   # converge para normal
```



Intervalos de Confiança na Prática

- Considere, por exemplo, uma amostra de *n* = 18 observações, com k = 3 variáveis explicativas na regressão, e vamos utilizar uma tabela da distribuição *t-Student*

- Nesse caso, $\nu = n - 1 - 3$ = 14, e os valores quantis são: 2.144, 2.977 e 1.761

```
q1 <- -qt((1-0.95)/2, df=14)
q2 <- -qt((1-0.99)/2, df=14)
q3 <- -qt((1-0.90)/2, df=14)
```

- Os intervalos de confiança nesses casos são:

$$\text{CI}_{95\%} = [\bar{X} - 2.144 \times [\text{Erro Padrão}], \quad \bar{X} + 2.144 \times [\text{Erro Padrão}]]$$

$$\text{CI}_{99\%} = [\bar{X} - 2.977 \times [\text{Erro Padrão}], \quad \bar{X} + 2.977 \times [\text{Erro Padrão}]]$$

$$\text{CI}_{90\%} = [\bar{X} - 1.761 \times [\text{Erro Padrão}], \quad \bar{X} + 1.761 \times [\text{Erro Padrão}]]$$

- Na prática, não precisamos nos calcular esses intervalos, pois eles já são dados com comandos adicionais por todos os softwares econométricos

Modelos de Regressão e Intervalos de Confiança

- Para modelos de regressão, podemos utilizar a função “confint”
    - No exercício anterior, considere o modelo que estimamos,

```
mod1.ex <- lm(dados3$mort_infantil ~ dados3$renda_per_capita
                                   + dados3$indice_gini
                                   + dados3$salario_medio_mensal
                                   + dados3$perc_criancas_extrem_pobres
                                   + dados3$perc_criancas_pobres
                                   + dados3$perc_pessoas_dom_agua_estogo_inadequados
                                   + dados3$perc_pessoas_dom_paredes_inadequadas
                                   + dados3$perc_pop_dom_com_coleta_lixo)
```

- Para obter os intervalos de confiança dos parâmetros estimados, podemos utilizar

```
confint(mod1.ex)                #--- probabilidade de cobertura de 95%
confint(mod1.ex, level = 0.9)   #--- probabilidade de cobertura de 90%
confint(mod1.ex, level = 0.8)   #--- probabilidade de cobertura de 80%
```

Intervalos de Confiança na Prática

- Ressaltamos que o intervalo de confiança também é uma variável aleatória. O intervalo varia com a amostra utilizada. Amostras diferentes geram intervalos de confiança diferente.
- Considere novamente as simulações de Monte Carlo que discutimos anteriormente
- Podemos estudar, por exemplo, o impacto do número de observações na amostra sobre a amplitude do intervalo de confiança
- Em geral, a amplitude cai à razão de raiz do número de observações na amostra; quando o número de observações é multiplicado por quatro, a amplitude em média cai à metade
- Mas qual a interpretação do chamado intervalo de confiança?
    - O intervalo de confiança contém ou não o parâmetro verdadeiro que queremos estimar.
    - O parâmetro verdadeiro é desconhecido (por isso estamos fazendo o estudo estatístico.
    - Em geral, se o estudo estatístico fosse repetido 1000 vezes, e fossem calculados 1000 intervalos de confiança (um para cada estudo) com nível de cobertura de 95%, em média, 950 desses intervalos irão conter o parâmetro verdadeiro.

Modelos de Regressão

- **Exercício 3 - para entregar em 2 semanas:**
  - Como de costume, os exercícios podem ser entregues em grupos de 2 ou três alunos, e o grupo deve submeter o código em R utilizado para responder ao exercício, juntamente com a discussão dos resultados
  - Utilize como base o código em R 'Analise_de_Regressao_Linear_Exercicios_Praticos_2'
  - Rode a regressão de acordo com o modelo abaixo:

```
mod1.ex <- lm(dados3$mort_infantil ~ dados3$renda_per_capita
                                   + dados3$indice_gini
                                   + dados3$salario_medio_mensal
                                   + dados3$perc_criancas_extrem_pobres
                                   + dados3$perc_criancas_pobres
                                   + dados3$perc_pessoas_dom_agua_estogo_inadequados
                                   + dados3$perc_pessoas_dom_paredes_inadequadas
                                   + dados3$perc_pop_dom_com_coleta_lixo)
```

Modelos de Regressão

- **Exercício 3 (continuação):**
  - Questão 1: Encontre os intervalos de confiança para os coeficientes estimados no modelo “mod1.ex”
  - Questão 2: Repita a questão 1 para o modelo “mod2.ex”, conforme abaixo

```
mod2.ex <- lm(dados3$mort_infantil ~ dados3$renda_per_capita
        + dados3$indice_gini
        + dados3$salario_medio_mensal
        + dados3$perc_criancas_extrem_pobres
        + dados3$perc_criancas_pobres
        + dados3$perc_pessoas_dom_agua_estogo_inadequados
        + dados3$perc_pessoas_dom_paredes_inadequadas
        + dados3$perc_pop_dom_com_coleta_lixo
        + dados3$perc_pop_rural
        + as.factor(dados3$Regiao))
```

Testes de Hipóteses para Modelos de Regressão Linear

Testes de Hipóteses

- Testes de hipótese são utilizados para testar se 'suspeitas' sobre os parâmetros do modelo de regressão são verdadeiras ou não

- O teste de hipótese mais comum em modelos de regressão é a respeito do valor de um determinado coeficiente específico

- Por exemplo, considere o modelo de regressão abaixo

$$[Salário]_i = \beta_0 + \beta_1[Experiência]_i + \delta_1[DummyMulher]_i + \gamma_1[DummyMulher]_i \times [Experiência]_i + \epsilon_i$$

- Conforme vimos anteriormente, a variável $[DummyMulher]_i$ indica se o trabalhador é do sexo feminino ou não

- O coeficiente δ_1 indica a diferença de salários entre homens e mulheres, 'controlando-se' para as demais variáveis

- Gostaríamos de testar se existe ou não discriminação no mercado de trabalho. Para isso, podemos testar se o coeficiente δ_1 é nulo ou diferente de zero

Testes de Hipóteses

- Para testar a presença ou não de discriminação, podemos então proceder com um teste de hipóteses

- Os testes de hipóteses possuem quatro elementos básicos:

  - **Hipóteses nula e alternativa**

  - **Estatística teste**

  - **Distribuição da estatística teste**

  - **Regressão de rejeição da hipótese nula**

- A hipótese nula no exemplo em questão é dada por: $H_0: \gamma_1 = 0$

- A hipótese alternativa é dada por: $H_A: \gamma_1 \neq 0$

Testes de Hipóteses

- A estatística teste utilizada nesse caso, tem expressão

$$t_{stat} = \frac{[\text{Estimativa do coeficiente}]}{[\text{Erro padrão do coeficiente}]} = \frac{\hat{\gamma}_1}{s.e._{\hat{\gamma}_1}}$$

- Se quiséssemos testar um valor mais geral (não zero) para o coeficiente, teríamos:

$$H_0: \gamma_1 = a$$
$$H_A: \gamma_1 \neq a$$

- A estatística teste seria dada então por:

$$t_{stat} = \frac{[\text{Estimativa do coeficiente}] - a}{[\text{Erro padrão do coeficiente}]} = \frac{\hat{\gamma}_1 - a}{s.e._{\hat{\gamma}_1}}$$

- Em geral, os sumários da regressão linear sempre trazem o erro padrão, e a estatística teste, para testar a hipótese nula de que cada coeficiente individualmente é igual a zero (ou seja, a = 0)

Testes de Hipóteses

```
Call:
lm(formula = dados3$mort_infantil ~ dados3$renda_per_capita +
    dados3$indice_gini + dados3$salario_medio_mensal + dados3$perc_criancas_extrem_pobres +
    dados3$perc_criancas_pobres + dados3$perc_pessoas_dom_agua_estogo_inadequados +
    dados3$perc_pessoas_dom_paredes_inadequadas + dados3$perc_pop_dom_com_coleta_lixo)

Residuals:
     Min       1Q   Median       3Q      Max
-16.5530  -2.4952  -0.3666   1.9344  20.8067

Coefficients:
                                              Estimate Std. Error t value Pr(>|t|)
(Intercept)                                  1.936e+01  8.196e-01  23.627  < 2e-16 ***
dados3$renda_per_capita                     -1.278e-03  5.784e-04  -2.209  0.02721 *
dados3$indice_gini                          -1.430e+01  1.247e+00 -11.470  < 2e-16 ***
dados3$salario_medio_mensal                 -1.775e-01  9.515e-02  -1.866  0.06212 .
dados3$perc_criancas_extrem_pobres           3.854e-02  1.216e-02   3.169  0.00154 **
dados3$perc_criancas_pobres                  2.159e-01  1.148e-02  18.812  < 2e-16 ***
dados3$perc_pessoas_dom_agua_estogo_inadequados  5.055e-02  6.021e-03   8.397  < 2e-16 ***
dados3$perc_pessoas_dom_paredes_inadequadas  4.297e-02  7.924e-03   5.423 6.12e-08 ***
dados3$perc_pop_dom_com_coleta_lixo         -7.045e-03  6.520e-03  -1.080  0.27999
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.038 on 5555 degrees of freedom
Multiple R-squared:  0.6804,    Adjusted R-squared:  0.6799
F-statistic:  1478 on 8 and 5555 DF,  p-value: < 2.2e-16
```

[Erro padrão do coeficiente]

Testes de Hipóteses

```
Call:
lm(formula = dados3$mort_infantil ~ dados3$renda_per_capita +
    dados3$indice_gini + dados3$salario_medio_mensal + dados3$perc_criancas_extrem_pobres +
    dados3$perc_criancas_pobres + dados3$perc_pessoas_dom_agua_estogo_inadequados +
    dados3$perc_pessoas_dom_paredes_inadequadas + dados3$perc_pop_dom_com_coleta_lixo)

Residuals:
     Min       1Q   Median       3Q      Max
-16.5530  -2.4952  -0.3666   1.9344  20.8067

Coefficients:
                                              Estimate Std. Error t value Pr(>|t|)
(Intercept)                                  1.936e+01  8.196e-01  23.627  < 2e-16 ***
dados3$renda_per_capita                     -1.278e-03  5.784e-04  -2.209  0.02721 *
dados3$indice_gini                          -1.430e+01  1.247e+00 -11.470  < 2e-16 ***
dados3$salario_medio_mensal                 -1.775e-01  9.515e-02  -1.866  0.06212 .
dados3$perc_criancas_extrem_pobres           3.854e-02  1.216e-02   3.169  0.00154 **
dados3$perc_criancas_pobres                  2.159e-01  1.148e-02  18.812  < 2e-16 ***
dados3$perc_pessoas_dom_agua_estogo_inadequados  5.055e-02  6.021e-03   8.397  < 2e-16 ***
dados3$perc_pessoas_dom_paredes_inadequadas   4.297e-02  7.924e-03   5.423 6.12e-08 ***
dados3$perc_pop_dom_com_coleta_lixo         -7.045e-03  6.520e-03  -1.080  0.27999
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.038 on 5555 degrees of freedom
Multiple R-squared:  0.6804,    Adjusted R-squared:  0.6799
F-statistic:  1478 on 8 and 5555 DF,  p-value: < 2.2e-16
```

t_{stat}

Testes de Hipóteses

- Quando o número de observações na amostra é muito alto, a estatística teste tem distribuição aproximadamente normal padronizada (média zero e desvio-padrão igual a 1)

- Portanto, podemos escrever

$$t = \frac{\hat{\gamma}_1 - a}{s.e._{\hat{\gamma}_1}} \approx \mathrm{N}(0,1)$$

- Em geral, podemos melhorar a aproximação da estatística teste, utilizando-se uma distribuição *t-Student*, com *(n-k-1)* graus de liberdade, onde *n* é o número de observações na amostra, *k* é o número de variáveis preditoras

- Quando o valor de *n* é muito alto, a aproximação normal e a aproximação via *t-Student* apresentam resultados praticamente idênticos

- Vimos que a distribuição *t-Student* converge para uma distribuição normal padronizada, quando o número de graus de liberdade aumenta

Testes de Hipóteses

- Finalmente, temos que ter uma regra de rejeição para a hipótese nula
- Nessa regra, temos que estabelecer a probabilidade de erro tipo I
- Essa probabilidade, corresponde à chance de rejeitarmos a hipótese nula, quando de fato ela é verdadeira
- Em geral, estabelecemos a probabilidade de erro tipo I (também denominada α do teste) igual a 1%, 5% ou 10%
- A partir daí, temos que encontrar os valores críticos do teste de hipótese, com base na distribuição para a estatística teste (por exemplo, normal padronizada ou *t-Student*)
- Esses valores críticos vão depender também da característica do teste de hipótese: bicaudal, unicaudal à direita, ou unicaudal à esquerda

Testes de Hipóteses

- Teste bicaudal, temos as hipóteses nula e alternativa da forma:

$$H_0: \gamma_1 = a$$
$$H_A: \gamma_1 \neq a$$

- Teste unicaudal à direita, temos:

$$H_0: \gamma_1 \leq a$$
$$H_A: \gamma_1 > a$$

- Finalmente, teste unicaudal à esquerda, temos:

$$H_0: \gamma_1 \geq a$$
$$H_A: \gamma_1 < a$$

Testes de Hipóteses

Valores críticos para o teste **bicaudal**:

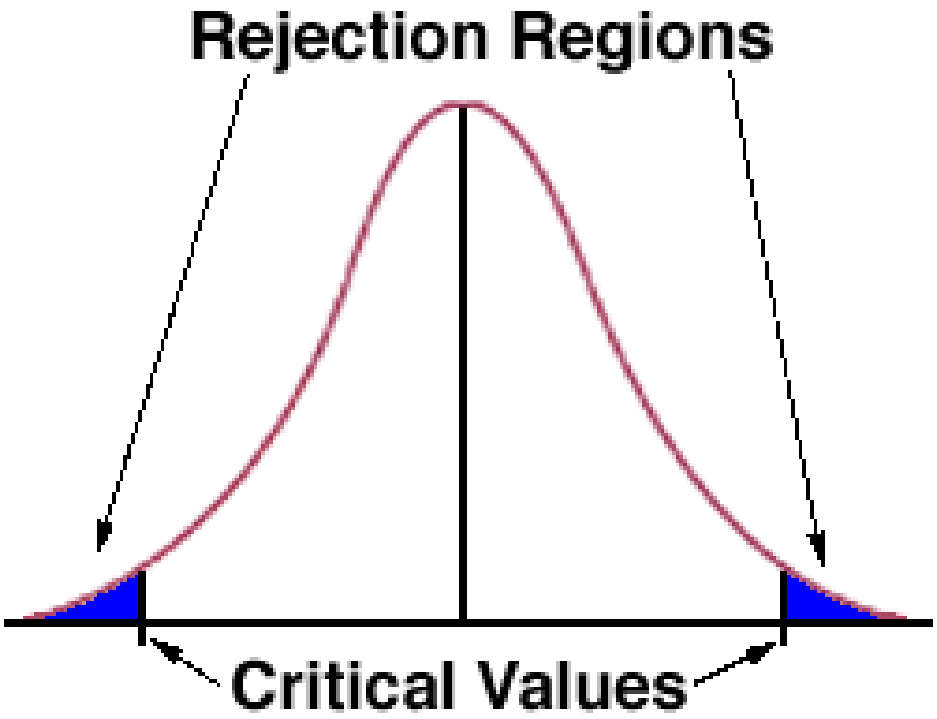
```
alpha <- 0.05;
q1 <- -qnorm(alpha/2);
q2 <- -qt(alpha/2, df=19);
q3 <- -qt(alpha/2, df=29);
q4 <- -qt(alpha/2, df=10000);
```

Obs. Note o sinal '-' nas fórmulas acima

Regra de rejeição da hipótese nula:

$|t_{stat}| > \text{valor crítico}$

Para um alpha = 5%, uma regra de bolsa é valor crítico aproximadamente igual a 2



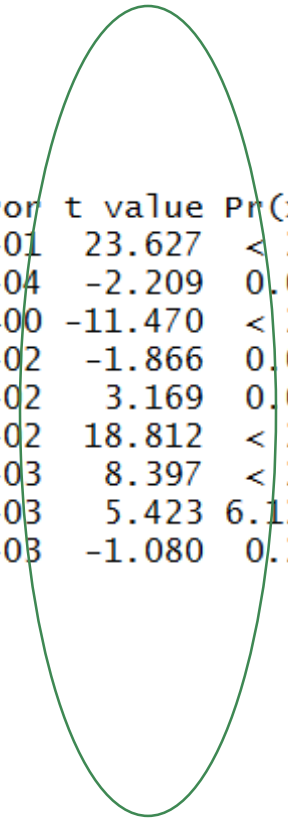
Testes de Hipóteses

```
Call:
lm(formula = dados3$mort_infantil ~ dados3$renda_per_capita +
    dados3$indice_gini + dados3$salario_medio_mensal + dados3$perc_criancas_extrem_pobres +
    dados3$perc_criancas_pobres + dados3$perc_pessoas_dom_agua_estogo_inadequados +
    dados3$perc_pessoas_dom_paredes_inadequadas + dados3$perc_pop_dom_com_coleta_lixo)

Residuals:
     Min       1Q   Median       3Q      Max
-16.5530  -2.4952  -0.3666   1.9344  20.8067

Coefficients:
                                              Estimate Std. Error t value Pr(>|t|)
(Intercept)                                  1.936e+01  8.196e-01  23.627  < 2e-16 ***
dados3$renda_per_capita                     -1.278e-03  5.784e-04  -2.209  0.02721 *
dados3$indice_gini                          -1.430e+01  1.247e+00 -11.470  < 2e-16 ***
dados3$salario_medio_mensal                 -1.775e-01  9.515e-02  -1.866  0.06212 .
dados3$perc_criancas_extrem_pobres           3.854e-02  1.216e-02   3.169  0.00154 **
dados3$perc_criancas_pobres                  2.159e-01  1.148e-02  18.812  < 2e-16 ***
dados3$perc_pessoas_dom_agua_estogo_inadequados  5.055e-02  6.021e-03   8.397  < 2e-16 ***
dados3$perc_pessoas_dom_paredes_inadequadas  4.297e-02  7.924e-03   5.423 6.12e-08 ***
dados3$perc_pop_dom_com_coleta_lixo         -7.045e-03  6.520e-03  -1.080  0.27999
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.038 on 5555 degrees of freedom
Multiple R-squared:  0.6804,    Adjusted R-squared:  0.6799
F-statistic:  1478 on 8 and 5555 DF,  p-value: < 2.2e-16
```



Testes de Hipóteses

Valores críticos para o teste **unicaudal à direita**:

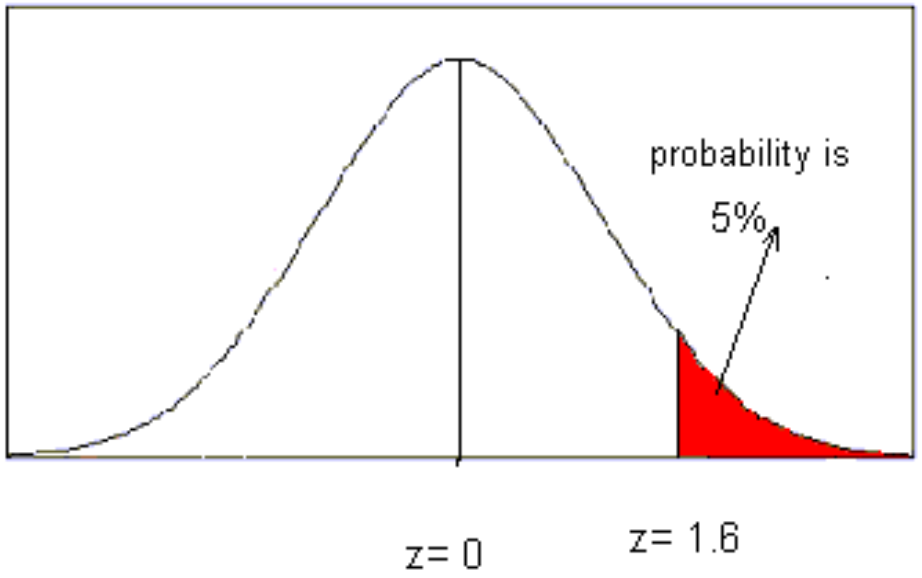
```
alpha <- 0.05;
q1 <- -qnorm(alpha);
q2 <- -qt(alpha, df=19);
q3 <- -qt(alpha, df=29);
q4 <- -qt(alpha, df=10000);
```

Obs. Note o sinal '-' nas fórmulas acima

Regra de rejeição da hipótese nula:

$t_{stat} >$ valor crítico

Obs. Não se aplica o valor absoluto nesse caso



Testes de Hipóteses

Valores críticos para o teste **unicaudal à esquerda**:

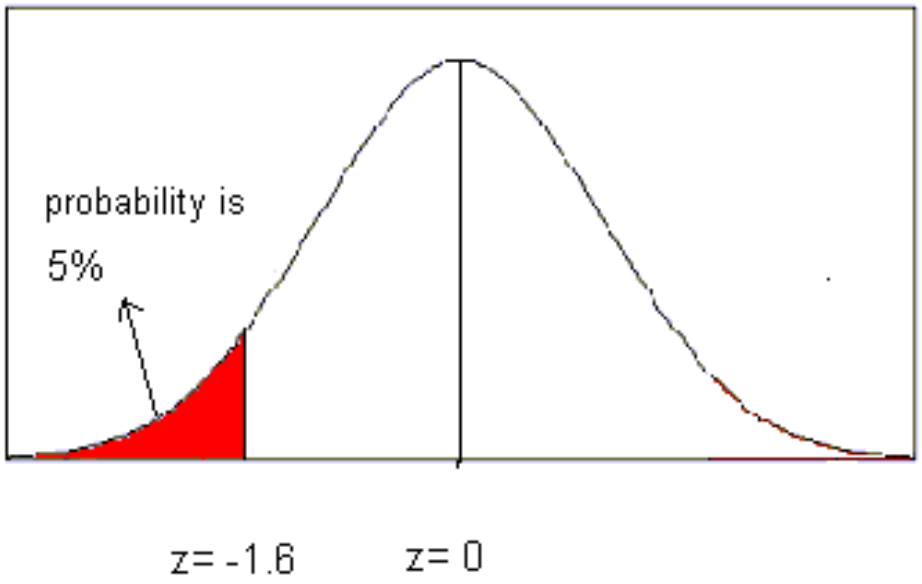
```
alpha <- 0.05;
q1 <- qnorm(alpha);
q2 <- qt(alpha, df=19);
q3 <- qt(alpha, df=29);
q4 <- qt(alpha, df=10000);
```

Obs. Note que não há mais o sinal '-' nas fórmulas acima

Regra de rejeição da hipótese nula:

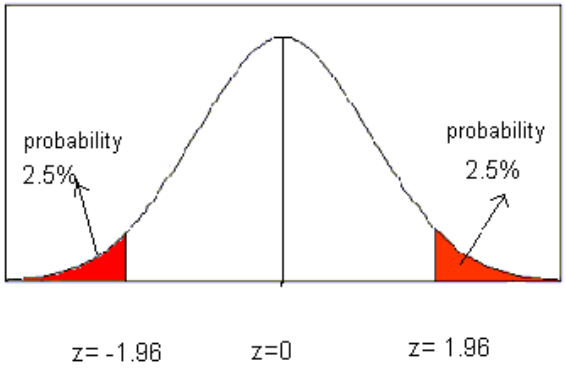
$t_{stat} <$ valor crítico

Obs. Não se aplica o valor absoluto nesse caso



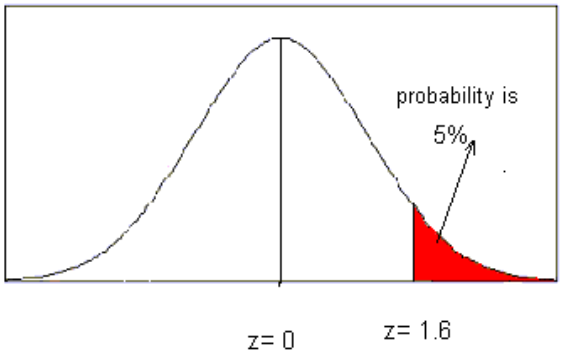
- Teste bicaudal :

$$H_0: \gamma_1 = a$$
$$H_A: \gamma_1 \neq a$$



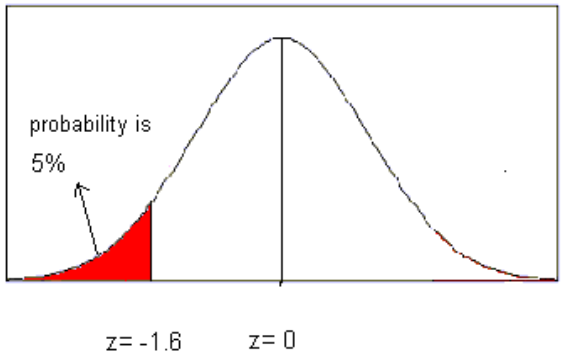
- Teste unicaudal à direita:

$$H_0: \gamma_1 \leq a$$
$$H_A: \gamma_1 > a$$



- Teste unicaudal à esquerda:

$$H_0: \gamma_1 \geq a$$
$$H_A: \gamma_1 < a$$



$$y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \beta_3 x_{3,i} + \ \ldots + \beta_k x_{k,i} + \epsilon_i$$

Regressão da taxa de internações por condições sensíveis sobre variáveis representativas da evolução da atenção básica por região

Internações por condições sensíveis	Região Norte	Região Nordeste	Região Sudeste	Região Sul	Região Centro-Oeste
Variáveis explicativas	Coeficientes e erros-padrão robustos				
Cobertura das ESFs	0,0484 (0,0682)	0,0143 (0,0331)	-0,1103*** (0,0308)	-0,0837** (0,0406)	-0,0677 (0,0644)
Cobertura dos ACS	0,1749** (0,0768)	0,1321*** (0,0426)	-0,1011*** (0,0355)	-0,0153 (0,0406)	-0,0111 (0,0606)
Cobertura dos cadastramentos	0,0039 (0,121)	-0,0305 (0,0442)	-0,1596*** (0,0447)	-0,0015 (0,0501)	-0,1792 (0,1121)
Observações	3592	14322	13239	9444	3671
Grupos (municípios)	449	1792	1664	1187	466

Testes de Hipóteses e P-Valores

- Na discussão acima, vimos que, para rejeitar a hipótese nula, precisamos comparar o valor da estatística teste com o valor crítico
- Essa comparação depende de se o teste é do tipo bicaudal, unicaudal à direita, ou unicaudal à esquerda
- Uma outra forma de se verificar se rejeitamos ou não a hipótese nula é através da utilização do que chamamos de *p-valor*
- Para entender o conceito de p-valor, considere um teste bicaudal para o coeficiente γ_1 da regressão para testar a discriminação de salários entre homens e mulheres
- As hipóteses nulas e alternativas são:

$$H_0: \gamma_1 = a$$
$$H_A: \gamma_1 \neq a$$

Testes de Hipóteses e P-Valores

- Para rejeitar a hipótese nula com nível de significância de 1%, no teste bi-caudal, precisamos que a estatística teste satisfaça (de acordo com a normal padronizada):

$$|t_{stat}| = \left|\frac{\hat{\gamma}_1}{s.e._{\hat{\gamma}_1}}\right| > 2.58$$

- Para rejeitar a hipótese nula com nível de significância de 5%, no teste bi-caudal, precisamos que a estatística teste satisfaça (de acordo com a normal padronizada):

$$|t_{stat}| = \left|\frac{\hat{\gamma}_1}{s.e._{\hat{\gamma}_1}}\right| > 1{,}96$$

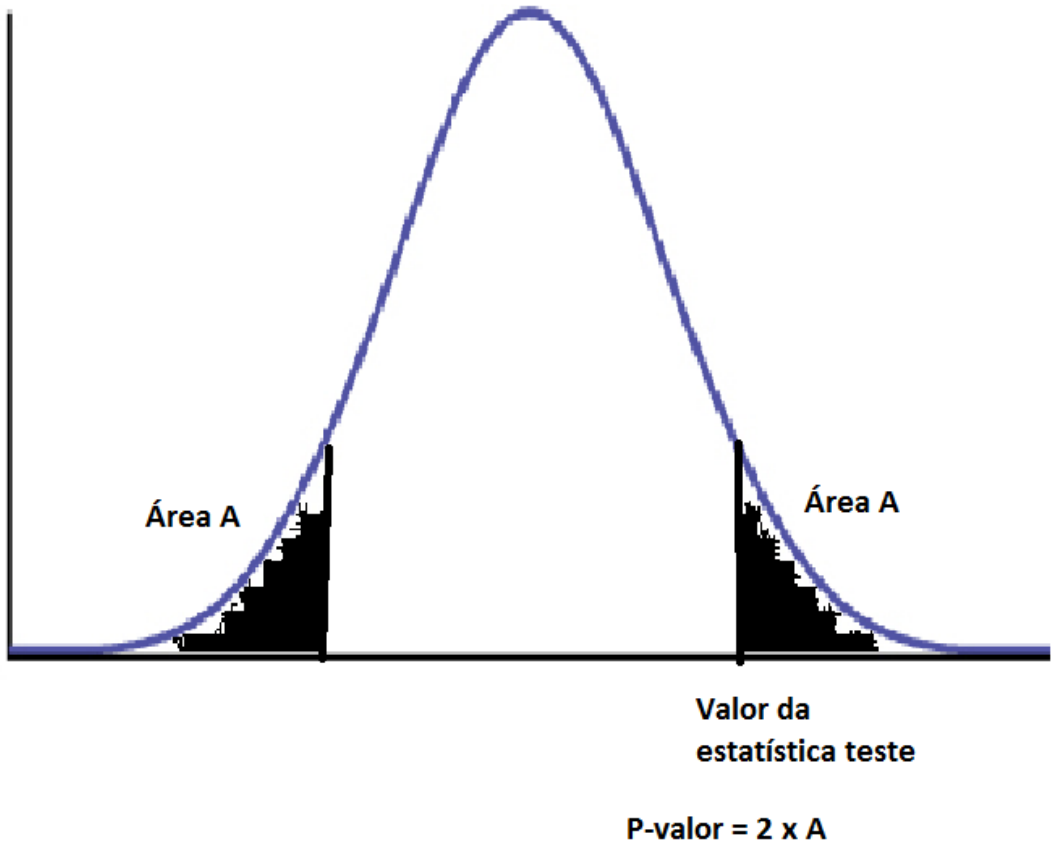
- Para rejeitar a hipótese nula com nível de significância de 10%, no teste bi-caudal, precisamos que a estatística teste satisfaça (de acordo com a normal padronizada):

$$|t_{stat}| = \left|\frac{\hat{\gamma}_1}{s.e._{\hat{\gamma}_1}}\right| > 1{,}645$$

- No caso de usarmos uma distribuição t-Student, os valores são um pouco maiores, dependendo do número de graus de liberdade

Testes de Hipóteses e P-Valores

No caso de testes bicaudais, o *p-valor* corresponde à soma das áreas nos extremos da distribuição, a partir do valor da estatística teste

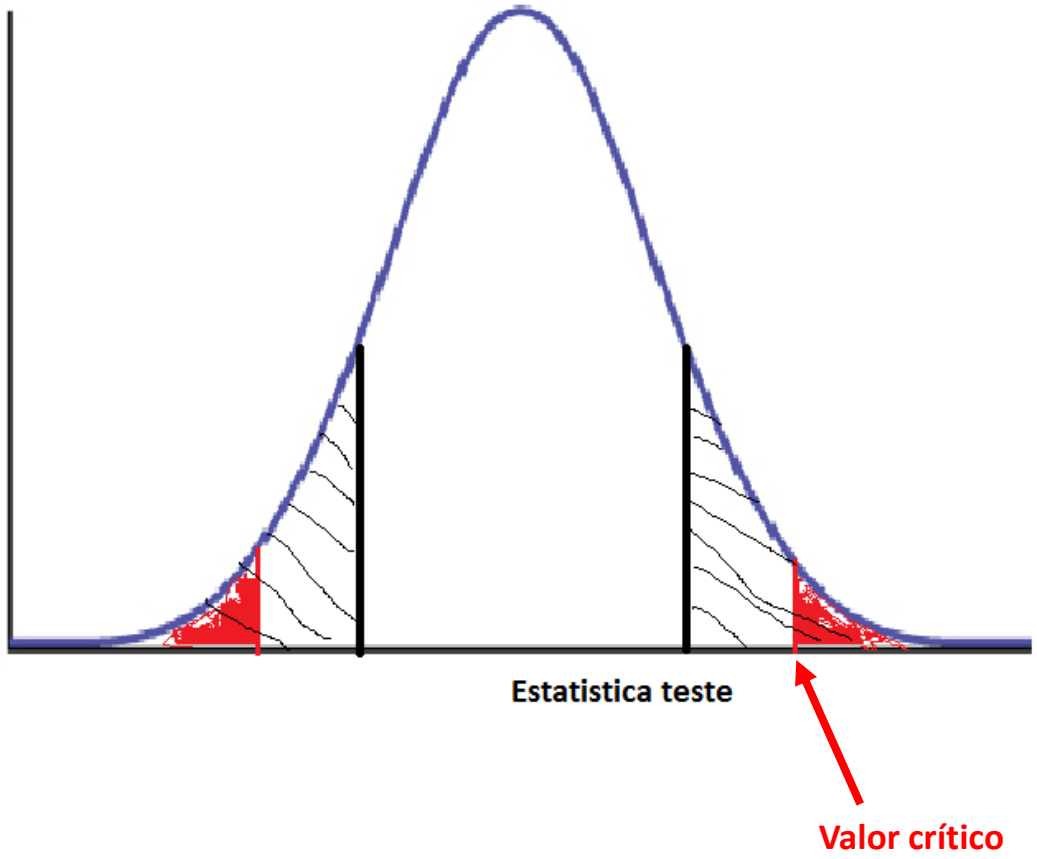


Testes de Hipóteses e P-Valores

Situação 1 – não rejeitamos a hipótese nula

Estatística teste não é maior do que o valor crítico

Ao mesmo tempo, o *p-valor* (soma das áreas hachuradas) é maior do que o nível do teste α (igual à soma das áreas em vermelho)

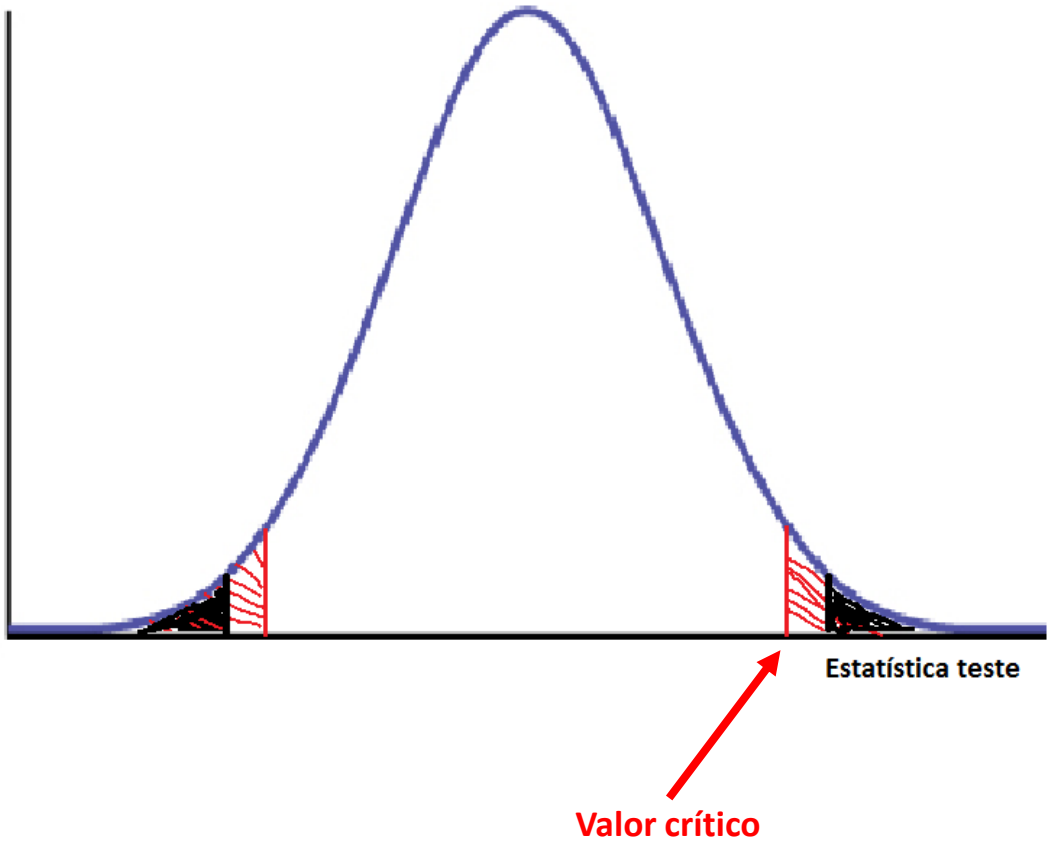


Testes de Hipóteses e P-Valores

Situação 2 – rejeitamos a hipótese nula

Estatística teste é maior do que o valor crítico

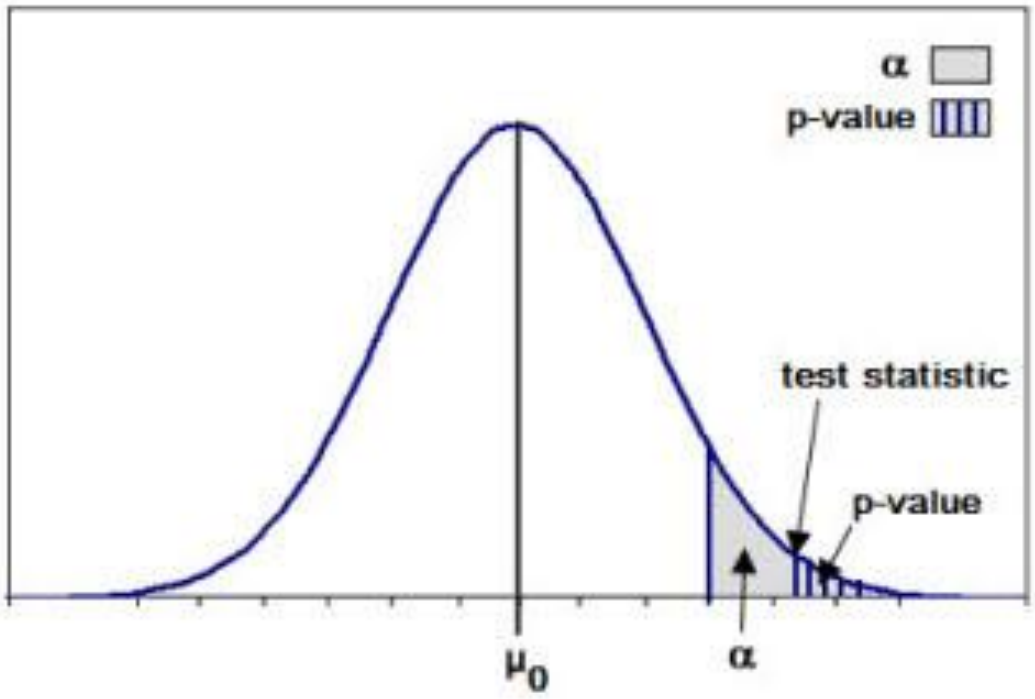
Ao mesmo tempo, o *p-valor* (soma das áreas em preto) é menor do que o nível do teste α (igual à soma das áreas hachuradas em vermelho)



Testes de Hipóteses e P-Valores

No caso de testes unicaudais, devemos considerar apenas a área de um dos lados da distribuição

Na figura ao lado, a estatística teste é maior do que o valor crítico, ao mesmo tempo em que o *p-valor* é menor do que o nível do teste α



Testes de Hipóteses e P-Valores

- Portanto, podemos considerar a seguinte regra de rejeição da hipótese nula, com base nos *p-valores*
    - Se p-valor < 0.05, então rejeitamos a hipótese nula com probabilidade de erro tipo I igual a 5%
    - Se p-valor < 0.10, então rejeitamos a hipótese nula com probabilidade de erro tipo I igual a 10%
    - Se p-valor < 0.01, então rejeitamos a hipótese nula com probabilidade de erro tipo I igual a 1%
- Diversos softwares estatísticos indicam o nível de significância da rejeição da hipótese nula, utilizando símbolos como, por exemplo, *, **, ***. O significado de cada um desses símbolos é indicado juntamente com a tabela de resultados

Significância dos Parâmetros em Modelos de Regressão

```
Call:
lm(formula = dados3$mort_infantil ~ dados3$renda_per_capita +
    dados3$indice_gini + dados3$salario_medio_mensal + dados3$perc_criancas_extrem_pobres +
    dados3$perc_criancas_pobres + dados3$perc_pessoas_dom_agua_estogo_inadequados +
    dados3$perc_pessoas_dom_paredes_inadequadas + dados3$perc_pop_dom_com_coleta_lixo)

Residuals:
     Min       1Q   Median       3Q      Max
-16.5530  -2.4952  -0.3666   1.9344  20.8067

Coefficients:
                                              Estimate Std. Error t value Pr(>|t|)
(Intercept)                                  1.936e+01  8.196e-01  23.627  < 2e-16 ***
dados3$renda_per_capita                     -1.278e-03  5.784e-04  -2.209  0.02721 *
dados3$indice_gini                          -1.430e+01  1.247e+00 -11.470  < 2e-16 ***
dados3$salario_medio_mensal                 -1.775e-01  9.515e-02  -1.866  0.06212 .
dados3$perc_criancas_extrem_pobres           3.854e-02  1.216e-02   3.169  0.00154 **
dados3$perc_criancas_pobres                  2.159e-01  1.148e-02  18.812  < 2e-16 ***
dados3$perc_pessoas_dom_agua_estogo_inadequados  5.055e-02  6.021e-03   8.397  < 2e-16 ***
dados3$perc_pessoas_dom_paredes_inadequadas  4.297e-02  7.924e-03   5.423 6.12e-08 ***
dados3$perc_pop_dom_com_coleta_lixo         -7.045e-03  6.520e-03  -1.080  0.27999
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.038 on 5555 degrees of freedom
Multiple R-squared:  0.6804,    Adjusted R-squared:  0.6799
F-statistic:  1478 on 8 and 5555 DF,  p-value: < 2.2e-16
```

Modelos de Regressão

- **Exercício 3 (continuação):**
  - Questão 3: para o modelo de regressão abaixo, quais os coeficientes estatisticamente significantes a 1%, 5%, e 10%? Quais coeficientes não são significantes nem mesmo a 10%?

```
mod3.ex <- lm(dados3$mort_infantil ~ dados3$renda_per_capita
        + dados3$indice_gini
        + dados3$salario_medio_mensal
        + dados3$perc_criancas_extrem_pobres
        + dados3$perc_criancas_pobres
        + dados3$perc_pessoas_dom_agua_estogo_inadequados
        + dados3$perc_pessoas_dom_paredes_inadequadas
        + dados3$perc_pop_dom_com_coleta_lixo
        + dados3$perc_pop_rural
        + as.factor(dados3$Regiao)
        + as.factor(dados3$Regiao)*dados3$renda_per_capita)
```

Modelos de Regressão

- **Exercício 3 (continuação):**
  - Questão 4: Com base na regressão da questão anterior, teste a hipótese nula de que o coeficiente da variável índice de Gini é menor ou igual a zero. Qual o p-valor para esse teste?

  - Questão 5: Com base na regressão da questão 3, teste a hipótese nula de que o coeficiente da variável percentual de crianças pobres é maior ou igual a 0.05. Qual o p-valor para esse teste?

Exemplos de Modelos de Regressão para Avaliação de Programas

Avaliação dos Fundos Constitucionais de Desenvolvimento Regional

- Discussão importante dado o montante de recursos destinados via fundos constitucionais de desenvolvimento regional

- Texto: “Avaliação dos Efeitos Econômicos dos Fundos Constitucionais de Financiamento do Nordeste, do Norte e do Centro-Oeste; uma Análise por Tipologia da PNDES entre 1999 e 2011”

- Regressão com dados de painel municipal, analisando o crescimento do PIB per capita municipal, em intervalos de tempo de 3 anos

    - Inclusão de variáveis *dummies* por município (efeitos fixos) e por ano

- Variáveis indicadoras da política são a proporção dos aportes dos fundos constitucionais sobre o PIB do município no início da janela de 3 anos

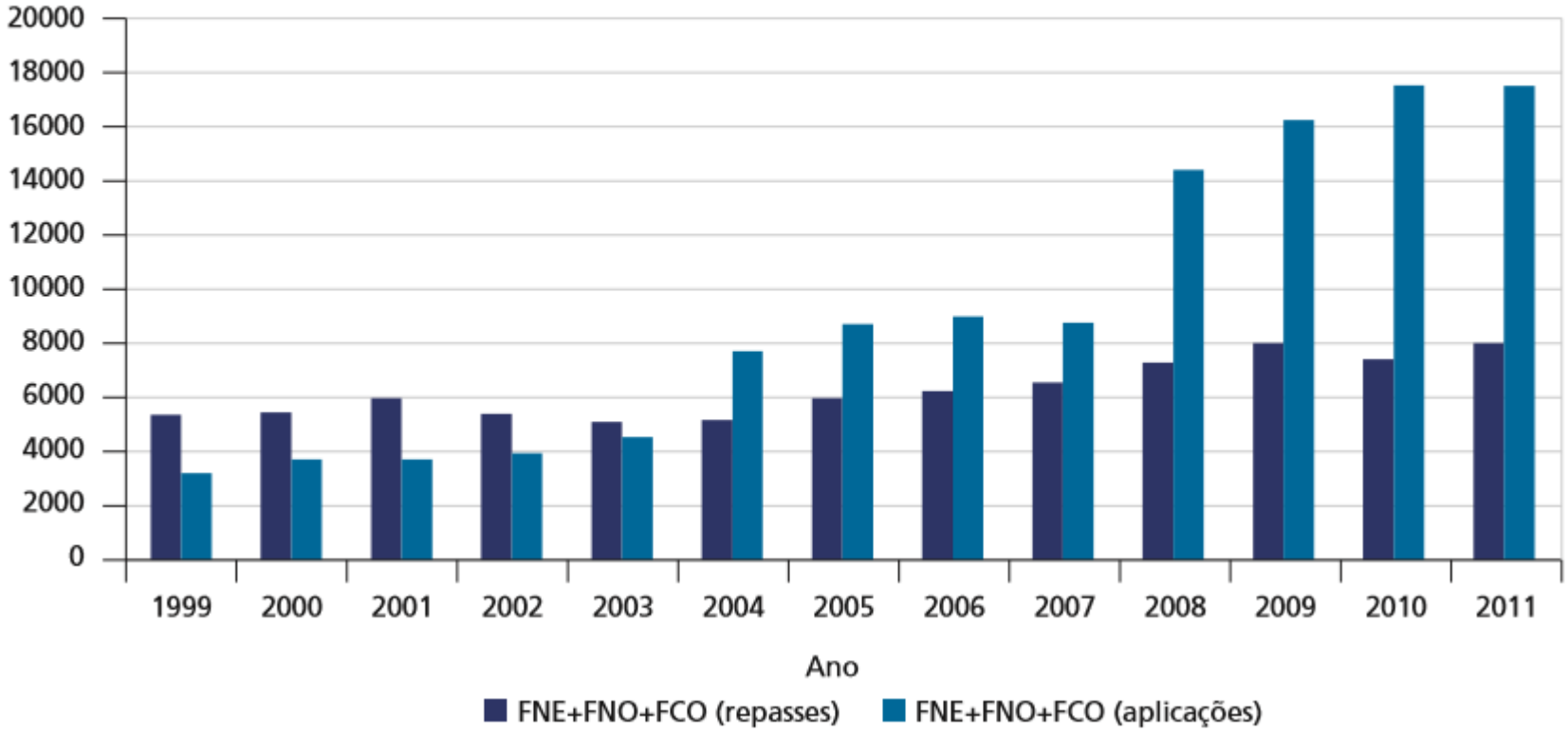
- Diversas variáveis explicativas foram incluídas no modelo para controlar para o efeito de outras variáveis

Avaliação dos Fundos Constitucionais de Desenvolvimento Regional

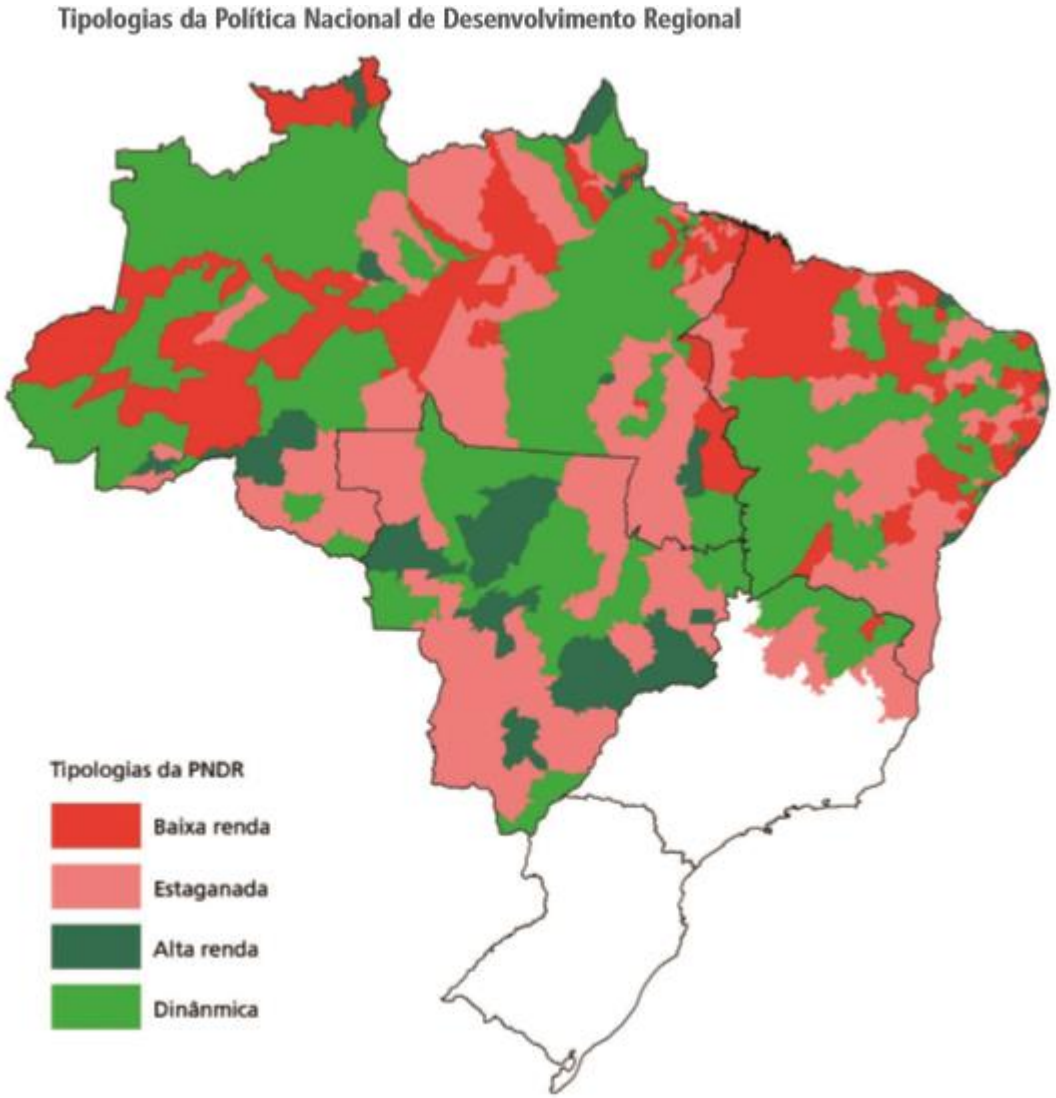
GRÁFICO 1

Repasses anuais do Tesouro Nacional e aplicações anuais dos recursos (1999-2011)

(R$ em milhões, a preços constantes de 2010)

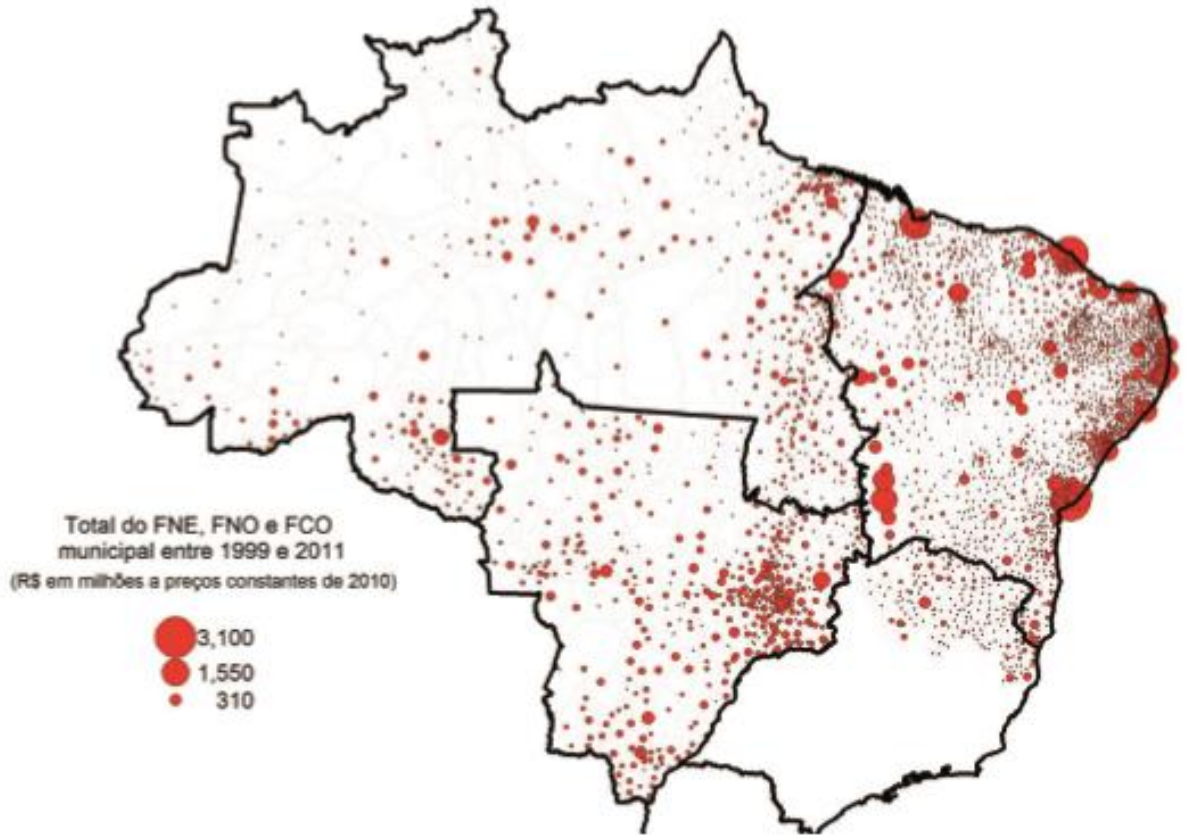


Avaliação dos Fundos Constitucionais de Desenvolvimento Regional



Avaliação dos Fundos Constitucionais de Desenvolvimento Regional

Distribuição espacial dos recursos do FNE, do FNO e do FCO no nível municipal (1999-2011)
(Em R$ milhões, a preços constantes de 2010)



Resultado dos efeitos do FNE sobre o crescimento médio anual do PIB *per capita* no nível municipal – método painel de efeitos fixos

	Variável dependente = taxa de crescimento anual média do PIB *per capita*				
Método de estimação	Painel efeitos fixos	Painel efeitos fixos		Painel efeitos fixos	Painel efeitos fixos
	(1)	(2)		(3)	(4)
Alta renda_Proporção do FNE início do período (1º ano) em relação ao PIB do início de cada período	0.9982**	0.8501**	Alta renda_Proporção do FNE início do período (1º + 2º ano) em relação ao PIB do início de cada período	-0.0122	-0.0380*
	(0.0157)	(0.0208)		(0.5977)	(0.0665)
Dinâmica_Proporção do FNE início do período (1º ano) em relação ao PIB do início de cada período	0.1407***	0.1225***	Dinâmica_Proporção do FNE início do período (1º + 2º ano) em relação ao PIB do início de cada período	0.1282***	0.1066***
	(0.0006)	(0.0010)		(0.0000)	(0.0000)
Baixa renda_Proporção do FNE início do período (1º ano) em relação ao PIB do início de cada período	0.4528***	0.2129***	Baixa renda_Proporção do FNE início do período (1º + 2º ano) em relação ao PIB do início de cada período	0.0934***	0.0273
	(0.0000)	(0.001)		(0.0002)	(0.2259)
Estagnada_Proporção do FNE início do período (1º ano) em relação ao PIB do início de cada período	0.1508**	-0.0191	Estagnada_Proporção do FNE início do período (1º + 2º ano) em relação ao PIB do início de cada período	0.1322***	0.0639***
	(0.0411)	0.7733		(0.0000)	(0.0099)
Ln (PIB per capita no início de cada período)	-0.1693***	-0.2944***	Ln (PIB *per capita* no início de cada período)	-0.1681***	-0.2936***
	(0.0000)	(0.0000)		(0.0000)	(0.0000)
Ln (anos médios de escolaridade no início de cada período, Rais)	0.0670***	-0.0103**	Ln (anos médios de escolaridade no início de cada período, Rais)	0.0653***	-0.01090***
	(0.0000)	(0.0138)		(0.0000)	(0.0091)
Ln (densidade populacional no início de cada período)	0.0926***	-0.1280***	Ln (densidade populacional no início de cada período)	0.0886***	-0.1280***
	(0.0000)	(0.0000)		(0.0000)	(0.0000)
Efeitos fixos	Sim	Sim	Efeitos fixos	Sim	Sim
Dummy de tempo	Não	Sim	*Dummy* de tempo	Não	Sim
Número de observações (municípios)	5.946	5.946		5.946	5.946
R2 ajustado	0.1739	0.3368		0.1779	0.3403

Resultado dos efeitos do FNO sobre o crescimento médio anual do PIB *per capita* no nível municipal – método painel de efeitos fixos

Método de estimação	Variável dependente = taxa de crescimento anual média do PIB *per capita*				
	Painel efeitos fixos	Painel efeitos fixos		Painel efeitos fixos	Painel efeitos fixos
	(1)	(2)		(3)	(4)
Alta renda_Proporção do FNO início do período (1º ano) em relação ao PIB do início de cada período	0.0866	-0.0115	Alta renda_Proporção do FNO início do período (1º + 2º ano) em relação ao PIB do início de cada período	0.2020*	0.1898*
	(0.5902)	(0.9406)		(0.0520)	(0.0566)
Dinâmica_Proporção do FNO início do período (1º ano) em relação ao PIB do início de cada período	0.1115**	0.0891**	Dinâmica_Proporção do FNO início do período (1º + 2º ano) em relação ao PIB do início de cada período	0.1167***	0.0981***
	(0.0147)	(0.0419)		(0.0017)	(0.0058)
Baixa renda_Proporção do FNO início do período (1º ano) em relação ao PIB do início de cada período	0.0721	0.0509	Baixa renda_Proporção do FNO início do período (1º + 2º ano) em relação ao PIB do início de cada período	0.0267	0.0194
	(0.1582)	(0.2987)		(0.1539)	(0.2795)
Estagnada_Proporção do FNO início do período (1º ano) em relação ao PIB do início de cada período	-0.0077	-0.0064	Estagnada_Proporção do FNO início do período (1º + 2º ano) em relação ao PIB do início de cada período	-0.0041	-0.0064
	(0.5416)	(0.5952)		(0.7329)	(0.5835)
Ln (PIB per capita no início de cada período)	-0.2100***	-0.2711***	Ln (PIB *per capita* no início de cada período)	-0.2087***	-0.2696***
	(0.0000)	(0.0000)		(0.0000)	(0.0000)
Ln (anos médios de escolaridade no início de cada período, Rais)	0.0418***	-0.0171	Ln (anos médios de escolaridade no início de cada período, Rais)	0.0439***	-0.0150
	(0.0015)	(0.2117)		(0.0008)	(0.2722)
Ln (densidade populacional no início de cada período)	0.0459*	-0.1205***	Ln (densidade populacional no início de cada período)	0.0412	-0.1234***
	(0.0690)	(0.0000)		(0.1019)	(0.0000)
Efeitos fixos	Sim	Sim	Efeitos fixos	Sim	Sim
Dummy de tempo	Não	Sim	*Dummy* de tempo	Não	Sim
Número de observações (municípios)	1.347	1.347		1.347	1.347
R2 ajustado	0.253	0.3037		0.257	0.3074

Resultado dos efeitos do FCO sobre o crescimento médio anual do PIB *per capita* no nível municipal – método painel de efeitos fixos

Método de estimação	Variável dependente = taxa de crescimento anual média do PIB *per capita*				
	Painel efeitos fixos	Painel efeitos fixos		Painel efeitos fixos	Painel efeitos fixos
	(1)	(2)		(3)	(4)
Alta renda_Proporção do FCO início do período (1º ano) em relação ao PIB do início de cada período	0.1068**	0.0457	Alta renda_Proporção do FCO início do período (1º + 2º ano) em relação ao PIB do início de cada período	0.1321***	0.0517*
	(0.0150)	(0.2716)		(0.0000)	(0.0980)
Dinâmica_Proporção do FCO início do período (1º ano) em relação ao PIB do início de cada período	0.2910***	0.0890	Dinâmica_Proporção do FCO início do período (1º + 2º ano) em relação ao PIB do início de cada período	0.0544	-0.0190
	(0.0011)	(0.2983)		(0.1316)	(0.5788)
Baixa renda_Proporção do FCO início do período (1º ano) em relação ao PIB do início de cada período	-	-	Baixa renda_Proporção do FCO início do período (1º + 2º ano) em relação ao PIB do início de cada período	-	-
	-	-		-	-
Estagnada_Proporção do FCO início do período (1º ano) em relação ao PIB do início de cada período	0.0579	-0.0326	Estagnada_Proporção do FCO início do período (1º + 2º ano) em relação ao PIB do início de cada período	0.0187	-0.0451**
	(0.1257)	(0.3709)		(0.3783)	(0.0309)
Ln (PIB *per capita* no início de cada período)	-0.2090***	-0.2188***	Ln (PIB *per capita* no início de cada período)	-0.2054***	-0.2204***
	(0.0000)	(0.0000)		(0.0000)	(0.0000)
Ln (anos médios de escolaridade no início de cada período, Rais)	0.1318***	0.0103	Ln (anos médios de escolaridade no início de cada período, Rais)	0.1247***	0.0076
	(0.0015)	(0.6049)		(0.0000)	(0.7008)
Ln (densidade populacional no início de cada período)	0.0430	-0.0632**	Ln (densidade populacional no início de cada período)	0.0413	-0.0717**
	(0.1175)	(0.0231)		(0.1328)	(0.0105)
Efeitos fixos	Sim	Sim	Efeitos fixos	Sim	Sim
Dummy de tempo	Não	Sim	*Dummy* de tempo	Não	Sim
Número de observações (municípios)	1.338	1.338		1.338	1.338
R2 ajustado	0.2574	0.2257		0.2578	0.2287

Regressão versus Identificação de Causalidade

- Uma das maiores críticas aos modelos de regressão é que estamos estudando apenas "correlação" entre variáveis, e não necessariamente causalidade

- Uma vasta literatura tem sido desenvolvida nas últimas décadas para tentar "identificar" causalidade entre variáveis de política sobre indicadores de performance

- O mantra da identificação de causalidade é a comparação da performance no grupo de tratamento (que foi submetido à política pública, por exemplo) e o grupo de controle

- Para que essa comparação seja válida, algumas das principais técnicas buscam comparar grupos de tratamento e controle homogêneos, a menos da exposição ou não à política pública

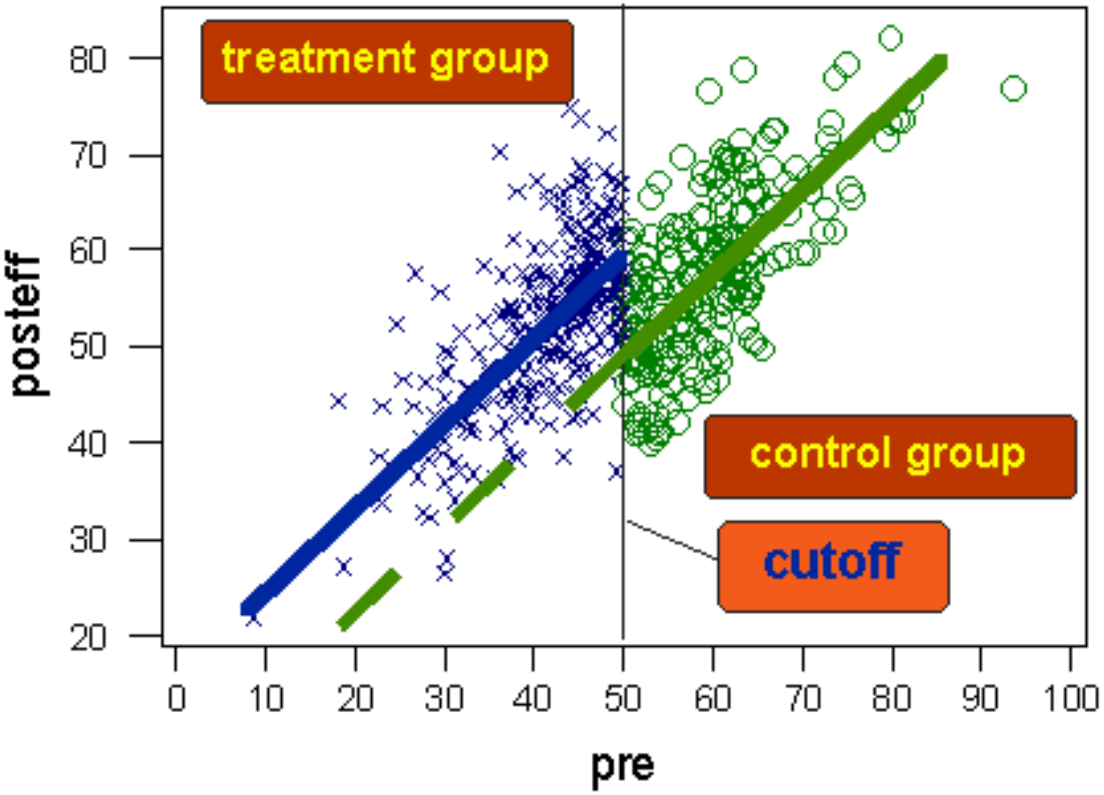
    - Mesmo que os grupos sejam diferentes, é importante que essas diferenças sejam capturadas totalmente pelo que chamamos de variáveis observáveis. Essas variáveis observáveis entram na regressão como variáveis de controle adicionais

    - O maior problema para identificação de causalidade acontece justamente quando há diferenças entre os grupos de controle e tratamento, que não podem ser capturadas pelas variáveis observáveis

Regressão de Descontinuidade

- A regressão de descontinuidade é um método bastante utilizando para identificação de causalidade em avaliação de políticas públicas

- Os grupos de controle e tratamento são montados a redor do corte de elegibilidade dos indivíduos à política pública



Efeito do SIMPLES Federal sobre a Geração de Empregos

- Texto: “O SIMPLES Federal e a Geração de Empregos na Indústria”, Carlos Henrique Courseuil e Rodrigo de Moura
- Lei número 9.317, de dezembro de 1996 – simplificação tributária, aplicação de alíquotas reduzidas, visando potencializar o desempenho dos estabelecimentos alvos
- Empregam regressão de descontinuidade para identificar o efeito do SIMPLES sobre o número de empregos das firmas no Brasil
    - Empresas com faturamento anual abaixo de um valor de corte são elegíveis ao imposto
- Dados da Pesquisa Industrial Anual (PIA) do IBGE
- O estudo olha os impactos do SIMPLES em dois instantes do tempo:
    - 1997, quando o SIMPLES foi implementado (valor de corte R$ 720.000 anual)
    - 1999, quando houve um aumento da receita máxima que torna a firma elegível (valor de corte R$ 1.200.000 anual)

Tabela 3. Estimativas do modelo FD – PIA 1996/97

Variável Dependente

Variação do Número de Empregados entre 1996 e 1997

	Janela (em milhares)					
Regressores	300	250	200	150	100	50
Simples (T_{i97})	1,69	1,96	1,64	1,29	1,86	1,80
	0,80	*0,85*	*0,79*	*0,76*	*0,87*	*1,20*
ΔF_{i96}	1E-08	2E-08	9E-09	8E-09	7E-10	-7E-07
	5E-09	*5E-09*	*5E-09*	*4E-09*	*3E-09*	*2E-06*
D_{i97} (F_{i96} –c)	-1E-06	-1E-06	9E-06	-1E-06	-1E-05	-2E-06
	4E-06	*7E-06*	*7E-06*	*7E-06*	*2E-05*	*5E-05*
Constante	-0,88	-1,18	-1,80	-0,88	-2,17	-2,59
	1,01	*1,12*	*1,14*	*1,29*	*1,77*	*2,22*
Nº de obs.	3499	2880	2312	1728	1165	603

Nota: Foram incluídos também como regressores dummies para setores de atividade e para as Unidades Federativas.
Simples (T i97) = 1 se a firma optou pelo SIMPLES em 1997 e 0 se não optou;
ΔF_{i96} é a varição do faturamento bruto de 1995 para 1996
(F i96 –c) = (Receita Bruta em 1996)-720000;
Erro Padrão em *itálico* .

Tabela 4. Estimativas do modelo FD – PIA 1998/99

Variável Dependente						
Variação do Número de Empregados entre 1998 e 1999	Janela (em milhares)					
Regressores	300	250	200	150	100	50
Simples (T_{i99})	2,77	2,24	2,82	0,98	0,69	2,52
	1,34	*1,14*	*1,15*	*1,07*	*1,25*	*2,11*
ΔF_{i98}	3E-09	6E-08	1E-06	6E-08	5E-07	2E-07
	1E-06	*1E-06*	*8E-07*	*8E-07*	*1E-06*	*1E-06*
D_{i99} (F_{i98} –c)	-6E-06	-6E-06	-2E-05	4E-06	7E-06	3E-05
	3E-06	*5E-06*	*7E-06*	*1E-05*	*2E-05*	*5E-05*
Constante	0,64	0,66	-0,83	0,00	-5,54	-6,31
	2,72	*2,65*	*2,07*	*4,03*	*2,56*	*3,30*
Nº de obs.	2111	1734	1394	1054	740	393

Nota: Foram incluídos também como regressores dummies para setores de atividade e para as Unidades Federativas.
Simples (T i99) = 1 se a firma optou pelo SIMPLES em 1999 e 0 se não optou;
ΔF_{i98} é a varição do faturamento bruto de 1997 para 1998
(F i98 –c) = (Receita Bruta em 1998)-1200000;
Erro Padrão em *itálico* .

Impacto do PBF sobre Indicadores Educacionais

- Texto: “Avaliação do Impacto do Programa Bolsa Família sobre Indicadores Educacionais”, Julio Alfredo Romero e Ana Maria Hermeto

- Estuda o efeito do PBF sobre indicadores educacionais das crianças de 7 a 14 anos, utilizando regressão de descontinuidade

    - O benefício básico (de R$ 50,00) era pago a famílias consideradas extremamente pobres (aquelas com renda mensal de até R$ 50,00 por pessoal);
    - O benefício variável era pago às famílias pobres, com renda mensal de até R$ 100,00 por pessoa
    - As regressões de descontinuidade consideraram dois cortes: R$ 50,00 e R$ 100,00

- Dados da pesquisa de Avaliação de Impacto do Bolsa Família (AIBF), de 2005, junto com o Cadastro Único

- Impactos satisfatórios sobre indicadores de curto prazo após a implantação do PBF: redução de evasão escolar de mulheres e aumento da aprovação para homens no Nordeste

Impacto do PBF sobre Indicadores Educacionais

TABELA 1 – Variáveis dependentes: Indicadores para avaliar os diferenciais do PBF na educação para crianças entre 7 e 14 anos de idade

Variáveis	Descrição
Não deixaram de ir à escola no último mês (ou o complemento deste)	Proporção de meninas e meninos no domicílio que não deixaram de ir à escola no último mês.
Evasão ou abandono	Proporção de meninas e meninos no domicílio que evadiram do sistema de ensino entre 2004 e 2005.
Progressão	Proporção de meninas e meninos no domicílio que foram aprovados entre 2004 e 2005.
Alocação entre trabalho e estudo	Proporção de meninas e meninos no domicílio que declararam só estudar atualmente, vis-à-vis aqueles que declararam só trabalhar, trabalhar e estudar e não trabalhar nem estudar.
Retenção	Proporção de meninas e meninos que foram reprovados entre 2004 e 2005.

Impacto do PBF sobre Indicadores Educacionais

TABELA 2 – Variáveis independentes: variáveis utilizadas na especificação dos modelos de regressão descontínua para avaliar os diferenciais do PBF na educação

Atributos do chefe de família:	
Raça do chefe de família	Branca Não Branca
Sexo do chefe de família	Masculino Feminino
Escolaridade do chefe de família	Até 3 anos de estudos* Até 4 anos de estudos* Até 7 anos de estudos*
Idade do chefe de família	Menor e igual há 50 anos Mais que 50 anos
Altura em metros do chefe de família	Medida em metros (mts)
Escolaridade da mãe do chefe de família	Mãe alfabetizada Mãe não alfabetizada
Tempo de permanência do chefe de família no município	Menos de 10 anos* Menos de 5 anos*
Tempo de permanência do chefe de família na área rural.	Viveu até os 14 anos Não viveu até os 14 anos

Impacto do PBF sobre Indicadores Educacionais

Características da família:	
Número de membros da família	Número de membros no domicilio
Crianças entre 0 a 3 anos de idade	Proporção de crianças de 0 a 3 anos
Crianças entre 0 a 6 anos de idade	Proporção de crianças de 0 a 6 anos
Crianças mulheres 7a14/ criança 0 a 14 anos	Proporção crianças mulheres 7 a 14/ crianças 0 a 14
Casal com filhos até 14 anos	O Casal tem filhos até 14 anos O Casal não tem filhos até 14 anos
Presença de pessoas de 60 anos ou mais	Há pessoa de 60 anos e mais no domicílio Há pessoa menor de 60 anos no domicílio.
Características do domicilio:	
Qualidade de domicilio[1]	Qualidade inferior* Qualidade media*
Área de residência do domicilio	Urbana Rural
Região de residência do domicílio	Nordeste* Norte – Centro Oeste*

*Para cada uma destas categorias foi construída uma variável *dummy*.

[1] Esta variável foi gerada através do método Grade of Membership (GOM), com três categorias para a qualidade das condições dos domicílios, classificadas em: muito boa, regular e ruim.

Impacto do PBF sobre Indicadores Educacionais

TABELA 5: Estimação da regressão descontínua dos indicadores para avaliar os diferenciais do PBF na educação de crianças de 7 a 14 anos. Brasil e Regiões, 2005

Variáveis/Regiões	Ponto de corte					
	Até R$100.00			Até R$50.00		
	Total	Homem	Mulheres	Total	Homem	Mulheres
a) Crianças que evadiram a escola em 2004 (evasão)						
Brasil			-0,015**			-0,017*
Nordeste	-0,026*					
Norte/centro-oeste				-0,023*		
b) Crianças que foram aprovados a escola entre 2004 e 2005						
Brasil						
Nordeste					0,283*	
Norte/centro-oeste						
c) Crianças que repetiram a escola entre 2004 e 2005 (repetência)						
Brasil						
Nordeste	-0,097*				-0,290*	
Norte/centro-oeste						
d) Crianças que só estudavam em 2005						
Brasil				-0,218***		
Nordeste			-0,134***			
Norte/centro-oeste						

Fonte: AIBF e CadÚnico, 2005.
* valor significativo a 10%; ** valor significativo a 5%; *** valor significativo a 1%.

Testes de Hipóteses para Vários Parâmetros ao Mesmo Tempo

Testes de Hipóteses para Vários Parâmetros

- Considere o teste de hipótese, com hipóteses nulas e alternativas conforme abaixo:

$$H_0: \gamma_1 = a$$
$$H_A: \gamma_1 \neq a$$

- Nesse caso, estamos testando o valor para apenas um parâmetro do modelo de regressão

- Em geral, é possível testar hipóteses correspondentes a vários parâmetros simultaneamente

- Por exemplo, considere novamente o nosso modelo de regressão com variáveis *dummies* para os efeitos da macrorregiões:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ \ldots + \beta_k x_{ki}$$
$$+\delta_1 D_{SU} + \delta_2 D_{NO} + \delta_3 D_{SE} + \delta_4 D_{NE} + \epsilon_i$$

- Gostaríamos agora de testar se todas as *dummies* para regiões são nulas. Isso equivale a dizermos que as diferenças para y_i entre as regiões são explicadas totalmente pelas demais variáveis explicativas na regressão

Testes de Hipóteses para Vários Parâmetros

- Nesse caso, as hipóteses nulas e alternativas podem ser escritas conforme abaixo:

$$H_0: \delta_1 = 0, \delta_2 = 0, \delta_3 = 0, \delta_4 = 0$$
$$H_A: \text{pelos menos um dos coefientes testados é diferente de zero}$$

- A estatística teste é dada por:

$$F = \frac{\left(R^2_{\text{irrestrito}} - R^2_{\text{restrito}}\right) \times (n - k - 1)}{\left(1 - R^2_{\text{irrestrito}}\right) \times m}$$

- $R^2_{\text{irrestrito}}$ é o coeficiente de determinação da regressão irrestrita (incluindo as *dummies*)
- R^2_{restrito} é o coeficiente de determinação da regressão restrita (excluindo as *dummies*)
- *n* é o número de observações na amostra
- *k* é o número de variáveis explicativas da regressão irrestrita (incluindo as dummies)
- *m* é o número de restrições testadas; no exemplo, é o número de coeficientes das dummies

Testes de Hipóteses para Vários Parâmetros

- Implementação no R:

- Equação do modelo “irrestrito” (com as *dummies* de regiões)

```
mod2.ex <- lm(dados3$mort_infantil ~ dados3$renda_per_capita
              + dados3$indice_gini
              + dados3$salario_medio_mensal
              + dados3$perc_criancas_extrem_pobres
              + dados3$perc_criancas_pobres
              + dados3$perc_pessoas_dom_agua_estogo_inadequados
              + dados3$perc_pessoas_dom_paredes_inadequadas
              + dados3$perc_pop_dom_com_coleta_lixo
              + dados3$perc_pop_rural
              + as.factor(dados3$Regiao))
summary(mod2.ex)
```

Testes de Hipóteses para Vários Parâmetros

- Equação do modelo “restrito” (sem as *dummies* de regiões)

```
mod2.ex.rest <- lm(dados3$mort_infantil ~ dados3$renda_per_capita
                   + dados3$indice_gini
                   + dados3$salario_medio_mensal
                   + dados3$perc_criancas_extrem_pobres
                   + dados3$perc_criancas_pobres
                   + dados3$perc_pessoas_dom_agua_estogo_inadequados
                   + dados3$perc_pessoas_dom_paredes_inadequadas
                   + dados3$perc_pop_dom_com_coleta_lixo
                   + dados3$perc_pop_rural)
summary(mod2.ex.rest)
```

- Testando a exclusão das variáveis *dummy*:

```
anova(mod2.ex.rest, mod2.ex, test='LRT')
```

Testes de Hipóteses para Vários Parâmetros

- Resultados:

```
Analysis of Variance Table

Model 1: dados3$mort_infantil ~ dados3$renda_per_capita + dados3$indice_gini +
    dados3$salario_medio_mensal + dados3$perc_criancas_extrem_pobres +
    dados3$perc_criancas_pobres + dados3$perc_pessoas_dom_agua_estogo_inadequados +
    dados3$perc_pessoas_dom_paredes_inadequadas + dados3$perc_pop_dom_com_coleta_lixo +
    dados3$perc_pop_rural
Model 2: dados3$mort_infantil ~ dados3$renda_per_capita + dados3$indice_gini +
    dados3$salario_medio_mensal + dados3$perc_criancas_extrem_pobres +
    dados3$perc_criancas_pobres + dados3$perc_pessoas_dom_agua_estogo_inadequados +
    dados3$perc_pessoas_dom_paredes_inadequadas + dados3$perc_pop_dom_com_coleta_lixo +
    dados3$perc_pop_rural + as.factor(dados3$Regiao)
  Res.Df   RSS Df Sum of Sq  Pr(>Chi)
1   5554 86683
2   5550 67741  4     18942 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

- Qual a conclusão a partir dos resultados do teste de hipótese? Os coeficientes das dummies de regiões são significativos conjuntamente ou não?

Testes de Hipóteses para Vários Parâmetros

- Implementação no R (alternativamente):
- Equação: (com função “linearHypothesis” do pacote “car”)

```
mod2.ex <- lm(dados3$mort_infantil ~ dados3$renda_per_capita
              + dados3$indice_gini
              + dados3$salario_medio_mensal
              + dados3$perc_criancas_extrem_pobres
              + dados3$perc_criancas_pobres
              + dados3$perc_pessoas_dom_agua_estogo_inadequados
              + dados3$perc_pessoas_dom_paredes_inadequadas
              + dados3$perc_pop_dom_com_coleta_lixo
              + dados3$perc_pop_rural
              + as.factor(dados3$Regiao))
summary(mod2.ex)

linearHypothesis(mod2.ex, c("dados3$indice_gini = 0",
                            "dados3$salario_medio_mensal = 0",
                            "dados3$perc_pop_rural"))
```

Testes de Hipóteses para Vários Parâmetros

- Implementação no R (alternativamente):

  Resultados:

  Linear hypothesis test

```
Hypothesis:
dados3$indice_gini = 0
dados3$salario_medio_mensal = 0
dados3$perc_pop_rural = 0

Model 1: restricted model
Model 2: dados3$mort_infantil ~ dados3$renda_per_capita + dados3$indice_gini +
   dados3$salario_medio_mensal + dados3$perc_criancas_extrem_pobres +
   dados3$perc_criancas_pobres + dados3$perc_pessoas_dom_agua_estogo_inadequados +
   dados3$perc_pessoas_dom_paredes_inadequadas + dados3$perc_pop_dom_com_coleta_lixo +
   dados3$perc_pop_rural + as.factor(dados3$Regiao)

  Res.Df   RSS Df Sum of Sq      F    Pr(>F)
1   5553 68663
2   5550 67741  3    922.43 25.191 3.498e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Testes de Hipóteses para Vários Parâmetros

- **Exercício prático para fixação**. Considere o modelo de regressão abaixo. Teste as hipóteses conjuntamente:

    - Teste a hipótese conjunta: (coeficiente do índice de gini) = 0, (coeficiente do salario médio mensal) = 1, (coeficiente prec crianças pobres) = 0. Qual o p-valor do teste? Você rejeita a hipótese nula? Com que nível de significância?

```
mod2.ex <- lm(dados3$mort_infantil ~ dados3$renda_per_capita
        + dados3$indice_gini
        + dados3$salario_medio_mensal
        + dados3$perc_criancas_extrem_pobres
        + dados3$perc_criancas_pobres
        + dados3$perc_pessoas_dom_agua_estogo_inadequados
        + dados3$perc_pessoas_dom_paredes_inadequadas
        + dados3$perc_pop_dom_com_coleta_lixo
        + dados3$perc_pop_rural
        + as.factor(dados3$Regiao))
```

Testes de Hipóteses para Vários Parâmetros

- Sob a hipótese nula, a estatística teste possui distribuição aproximadamente qui-quadrada, com número de graus de liberdade igual ao número de restrições no modelo

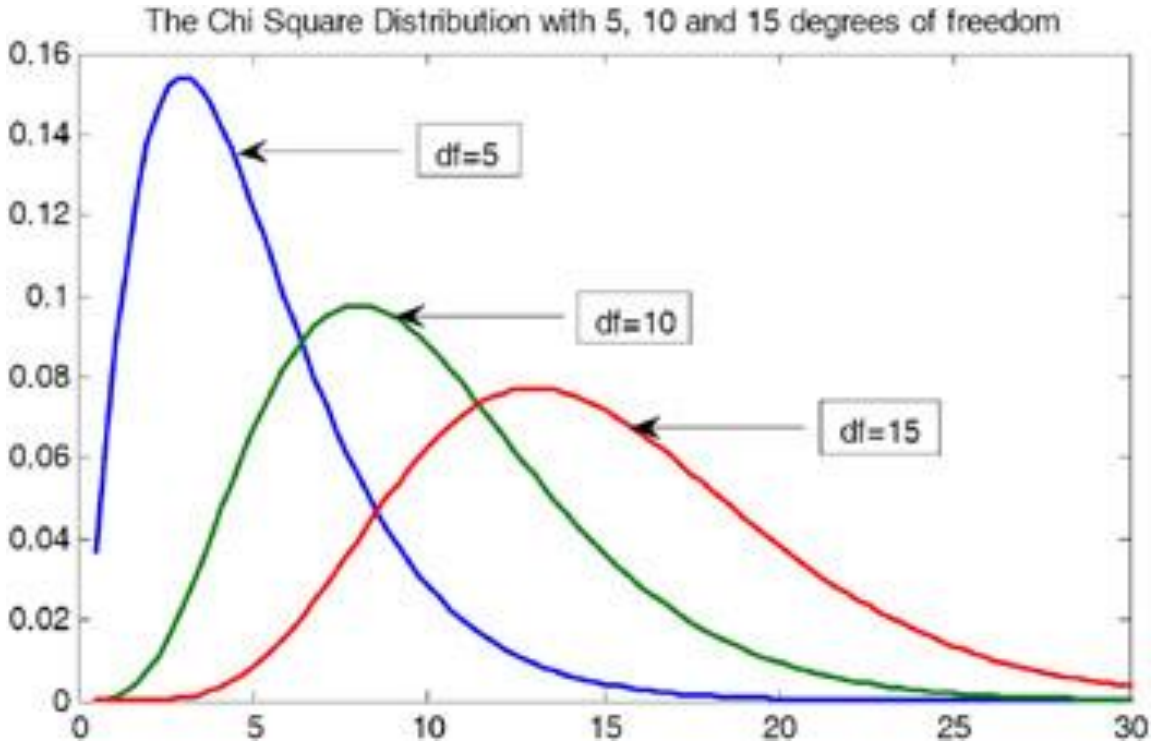
- Por exemplo, se estivermos testando a significância de quatro parâmetros conjuntamente, a estatística teste tem distribuição qui-quadrada com quatro graus de liberdade

- Para pequenas amostras, da mesma forma que utilizamos a distribuição *t*-Student, ao invés da distribuição normal padronizada, para testar múltiplos parâmetros, nós utilizamos a distribuição *F* ao invés da distribuição qui-quadrada

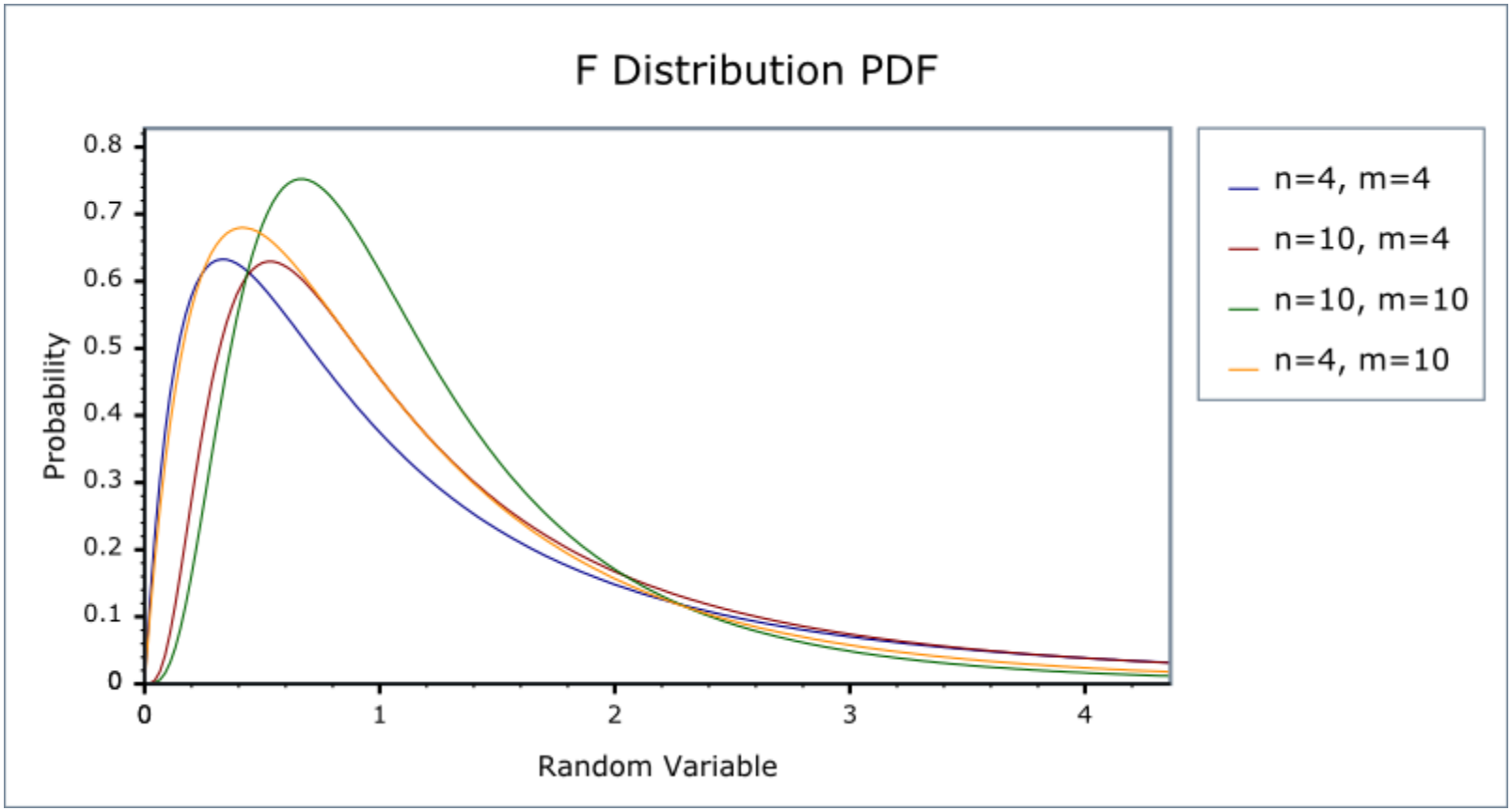
- Nesse caso, assumimos que, sob a hipótese nula, a distribuição da estatística teste é aproximadamente uma distribuição *F*, com número de graus de liberdade no numerador igual ao número de restrições. No denominado, o número de graus de liberdade é igual a $n-k-1$ (k é o número de parâmetros do modelo irrestrito)

- Pode-se mostrar que, quando n vai para o infinito, a distribuição *F* converge para uma qui-quadrada dividido pelo seu número de graus de liberdade

Testes de Hipóteses para Vários Parâmetros



Testes de Hipóteses para Vários Parâmetros



Testes de Hipóteses para Vários Parâmetros

- Considere agora o exemplo com termos quadrático e cúbico para uma das variáveis – vimos que dessa forma podemos capturar não-linearidades na relação entre a variável preditora e a variável resposta:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ \ldots + \beta_k x_{ki} + \boldsymbol{\beta_{(k+1)} x_{1i}^2} + \boldsymbol{\beta_{(k+2)} x_{1i}^3} + \epsilon_i$$

- Podemos testar se os termos não-lineares são necessários

$$H_0\colon \boldsymbol{\beta_{(k+1)} = \beta_{(k+2)} = 0}$$

$$H_A\colon \text{pelos menos um dos parâmetros é diferente de zero}$$

- Para isso, podemos proceder da mesma forma que o exemplo anterior: (i) rodamos o modelo irrestrito; (ii) rodamos o modelo restrito; (iii) fazemos a comparação entre os dois modelos

Testes de Hipóteses para Vários Parâmetros

- Implementação no R:

- Equação do modelo “irrestrito” (com os termos quadrático e cúbico)

```
mod1b.ex <- lm(dados3$mort_infantil ~ dados3$renda_per_capita
               + I(renda_per_capita^2)
               + I(renda_per_capita^3)
               + dados3$indice_gini
               + dados3$salario_medio_mensal
               + dados3$perc_criancas_extrem_pobres
               + dados3$perc_criancas_pobres
               + dados3$perc_pessoas_dom_agua_estogo_inadequados
               + dados3$perc_pessoas_dom_paredes_inadequadas
               + dados3$perc_pop_dom_com_coleta_lixo, data = dados)
summary(mod1b.ex)
```

Testes de Hipóteses para Vários Parâmetros

- Resultados do modelo irrestrito:

```
Coefficients:
                                              Estimate Std. Error t value Pr(>|t|)
(Intercept)                                  3.519e+01  1.202e+00  29.278  < 2e-16 ***
dados3$renda_per_capita                     -6.809e-02  4.012e-03 -16.975  < 2e-16 ***
I(renda_per_capita^2)                        6.170e-05  4.105e-06  15.030  < 2e-16 ***
I(renda_per_capita^3)                       -1.753e-08  1.400e-09 -12.523  < 2e-16 ***
dados3$indice_gini                           1.178e+00  1.492e+00   0.789 0.429906
dados3$salario_medio_mensal                 -9.563e-02  9.277e-02  -1.031 0.302675
dados3$perc_criancas_extrem_pobres          -2.928e-02  1.280e-02  -2.287 0.022230 *
dados3$perc_criancas_pobres                  6.767e-02  1.418e-02   4.772 1.87e-06 ***
dados3$perc_pessoas_dom_agua_estogo_inadequados 2.714e-02  6.010e-03   4.517 6.40e-06 ***
dados3$perc_pessoas_dom_paredes_inadequadas  2.632e-02  7.768e-03   3.389 0.000708 ***
dados3$perc_pop_dom_com_coleta_lixo          2.948e-03  6.368e-03   0.463 0.643460
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.928 on 5553 degrees of freedom
Multiple R-squared:  0.6977,  Adjusted R-squared:  0.6971
F-statistic:  1281 on 10 and 5553 DF,  p-value: < 2.2e-16
```

- Os termos quadrático e cúbico são estatisticamente significantes individualmente?
- De acordo com a significância dos termos quadrático e cúbico acima, você acha que a hipótese nula H_0: $\beta_{(k+1)} = \beta_{(k+2)} = 0$ vai ser rejeitada ou aceita?

Testes de Hipóteses para Vários Parâmetros

- Implementação no R:

- Equação do modelo “restrito” (sem os termos quadrático e cúbico)

```
mod1b.ex.rest <- lm(dados3$mort_infantil ~ dados3$renda_per_capita
                    + dados3$indice_gini
                    + dados3$salario_medio_mensal
                    + dados3$perc_criancas_extrem_pobres
                    + dados3$perc_criancas_pobres
                    + dados3$perc_pessoas_dom_agua_estogo_inadequados
                    + dados3$perc_pessoas_dom_paredes_inadequadas
                    + dados3$perc_pop_dom_com_coleta_lixo, data = dados)
summary(mod1b.ex.rest)
```

- Comparação entre os modelos:

```
anova(mod1b.ex.rest, mod1b.ex, test='LRT')
```

Testes de Hipóteses para Vários Parâmetros

- Resultados do teste de exclusão de variáveis:

```
Analysis of Variance Table

Model 1: dados3$mort_infantil ~ dados3$renda_per_capita + dados3$indice_gini +
    dados3$salario_medio_mensal + dados3$perc_criancas_extrem_pobres +
    dados3$perc_criancas_pobres + dados3$perc_pessoas_dom_agua_estogo_inadequados +
    dados3$perc_pessoas_dom_paredes_inadequadas + dados3$perc_pop_dom_com_coleta_lixo
Model 2: dados3$mort_infantil ~ dados3$renda_per_capita + I(renda_per_capita^2) +
    I(renda_per_capita^3) + dados3$indice_gini + dados3$salario_medio_mensal +
    dados3$perc_criancas_extrem_pobres + dados3$perc_criancas_pobres +
    dados3$perc_pessoas_dom_agua_estogo_inadequados +
dados3$perc_pessoas_dom_paredes_inadequadas +
    dados3$perc_pop_dom_com_coleta_lixo
  Res.Df   RSS Df Sum of Sq  Pr(>Chi)
1   5555 90564
2   5553 85670  2    4893.6 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

- De acordo com os resultados do teste acima, você rejeita a hipótese nula com nível de significância de 1%? E de 5%? E de 10%

- Qual a nossa conclusão sobre a necessidade de inclusão de termos quadrático e cúbico na equação?

Testes de Hipóteses para Vários Parâmetros

- Voltando aos resultados do modelo irrestrito:

```
Coefficients:
                                              Estimate Std. Error t value Pr(>|t|)
(Intercept)                                  3.519e+01  1.202e+00  29.278  < 2e-16 ***
dados3$renda_per_capita                     -6.809e-02  4.012e-03 -16.975  < 2e-16 ***
I(renda_per_capita^2)                        6.170e-05  4.105e-06  15.030  < 2e-16 ***
I(renda_per_capita^3)                       -1.753e-08  1.400e-09 -12.523  < 2e-16 ***
dados3$indice_gini                           1.178e+00  1.492e+00   0.789 0.429906
dados3$salario_medio_mensal                 -9.563e-02  9.277e-02  -1.031 0.302675
dados3$perc_criancas_extrem_pobres          -2.928e-02  1.280e-02  -2.287 0.022230 *
dados3$perc_criancas_pobres                  6.767e-02  1.418e-02   4.772 1.87e-06 ***
dados3$perc_pessoas_dom_agua_estogo_inadequados 2.714e-02  6.010e-03   4.517 6.40e-06 ***
dados3$perc_pessoas_dom_paredes_inadequadas  2.632e-02  7.768e-03   3.389 0.000708 ***
dados3$perc_pop_dom_com_coleta_lixo          2.948e-03  6.368e-03   0.463 0.643460
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.928 on 5553 degrees of freedom
Multiple R-squared:  0.6977,  Adjusted R-squared:  0.6971
F-statistic:  1281 on 10 and 5553 DF,  p-value: < 2.2e-16
```

- Qual o significado do termo *F-statistic* e do respectivo *p-value* na última linha do output da regressão?

- Na verdade, essa linha corresponde um teste de hipótese conjunto de vários parâmetros

Testes de Hipóteses para Vários Parâmetros

- **Exercício prático para fixação**. Considere o modelo de regressão abaixo. Teste as hipóteses conjuntamente:

    - Teste a hipótese: (coeficiente do índice de gini) + 2 * (coeficiente do salário médio) = 0. Qual o p-valor do teste? Você rejeita a hipótese nula? Com que nível de significância?

      (dica: use o comando ?linearHypothesis e veja os exemplos no help dessa função)

```
mod2.ex <- lm(dados3$mort_infantil ~ dados3$renda_per_capita
        + dados3$indice_gini
        + dados3$salario_medio_mensal
        + dados3$perc_criancas_extrem_pobres
        + dados3$perc_criancas_pobres
        + dados3$perc_pessoas_dom_agua_estogo_inadequados
        + dados3$perc_pessoas_dom_paredes_inadequadas
        + dados3$perc_pop_dom_com_coleta_lixo
        + dados3$perc_pop_rural
        + as.factor(dados3$Regiao))
```

Testes de Hipóteses para Vários Parâmetros

- Considere agora o modelo geral de regressão:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ \ldots + \beta_k x_{ki} + \epsilon_i$$

- Nesse modelo geral, as variáveis preditoras podem incluir termos quadrático, cúbico, etc., podem incluir variáveis *dummy*, e podem incluir interações entre variáveis
- O termo *F-statistic* e o respectivo *p-value* correspondem justamente à hipótese nula conjunta

$$H_0: \beta_1 = \beta_2 = \ \ldots = \beta_k = 0$$

$$H_A: \text{pelo menos um dos coeficientes é diferente de zero}$$

- Note que o intercepto β_0 não está sendo testado. Portanto, a estatística *F* nesse caso está testando um modelo com apenas o intercepto versus um modelo com o intercepto mais as variáveis preditoras

Modelos de Regressão

- **Exercício 4 - para entregar em 2 semanas:**
  - Como de costume, os exercícios podem ser entregues em grupos de 2 ou três alunos, e o grupo deve submeter o código em R utilizado para responder ao exercício, juntamente com a discussão dos resultados
  - Utilize como base o código em R 'Analise_de_Regressao_Linear_Exercicios_Praticos_2'
  - Rode a regressão de acordo com o modelo abaixo:

```
mod1.ex <- lm(dados3$mort_infantil ~ dados3$renda_per_capita
                                   + dados3$indice_gini
                                   + dados3$salario_medio_mensal
                                   + dados3$perc_criancas_extrem_pobres
                                   + dados3$perc_criancas_pobres
                                   + dados3$perc_pessoas_dom_agua_estogo_inadequados
                                   + dados3$perc_pessoas_dom_paredes_inadequadas
                                   + dados3$perc_pop_dom_com_coleta_lixo)
```

Modelos de Regressão

- **Exercício 4 (continuação):**
  - Questão 1: Teste a hipótese nula conjunta de que todos os coeficientes da regressão são nulos, exceto o intercepto. Qual o p-valor para esse teste? Você rejeita a hipótese nula com nível de significância de 1%? Você rejeita com nível de significância de 5%? Escreva as hipótese nula e alternativa.
  - Questão 2: Com base na regressão da questão anterior, aumente o modelo de regressão incluindo um termo quadrático para a renda per capita e outro para o índice Gini. Teste a significância conjunta desses dois termos quadráticos. Qual o p-valor para esse teste? Você rejeita a hipótese nula com nível de significância de 1%? Você rejeita com nível de significância de 5%? Escreva as hipótese nula e alternativa.
  - Questão 3: Com base na regressão básica do slide anterior, acrescente ao modelo uma interação entre a variável perc_criancas_extrem_pobres e a macrorregião na qual o município se encontra. Teste a hipótese nula de que essa interação é nula. Qual o p-valor para esse teste? Você rejeita a hipótese nula com nível de significância de 1%?
  - Questão 4: Teste a hipótese de que o efeito da variável salario_medio_mensal sobre a mortalidade infantil nos municípios é o mesmo para todas as macrorregiões no Brasil.

Modelos de Regressão na Forma Matricial

Modelos de Regressão em Notação Matricial

- Considere agora o modelo geral de regressão:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ \ldots + \beta_k x_{ki} + \epsilon_i$$

- Na equação acima, o termo ϵ_i corresponde ao erro da regressão, e incorpora todos os demais fatores, não presentes na equação linear, que explicam a variável y_i
- Até agora, utilizamos notações mais simplificadas, para fins de apresentação dos conceitos e utilização dos modelos de regressão
- Na maioria dos livros de regressão, e em vários trabalhos publicados, utiliza-se a notação matricial
- A notação matricial, além de simplificar a apresentação dos resultados, também indica expressões que podem ser utilizadas no software R
- O R tem todo um arcabouço para somas, multiplicação, inversão, subtração, transposição etc. de matrizes
- Conforme veremos mais adiante, fórmulas para o estimador de mínimos quadrados ordinário são bem simples, quando utilizamos a notação matricial
- Nesta seção, apresentaremos alguns conceitos básicos da notação matricial, que são muito utilizados na literatura

Modelos de Regressão – Forma Matricial

Modelo de regressão linear simples (apenas um preditor)

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

$$\epsilon_i \sim^{iid} N(0, \sigma^2)$$

Amostra com *n* observações

$$\begin{aligned} Y_1 &= \beta_0 + \beta_1 X_1 + \epsilon_1 \\ Y_2 &= \beta_0 + \beta_1 X_2 + \epsilon_2 \\ \vdots & \vdots \quad \vdots \\ Y_n &= \beta_0 + \beta_1 X_n + \epsilon_n \end{aligned}$$

Modelos de Regressão – Forma Matricial

Escrevendo na forma de matrizes:

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} \beta_0 + \beta_1 X_1 \\ \beta_0 + \beta_1 X_2 \\ \vdots \\ \beta_0 + \beta_1 X_n \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 & X_1 \\ 1 & X_2 \\ \vdots & \vdots \\ 1 & X_n \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

Design matriz = $\mathbf{X}_{n\times 2} = \begin{bmatrix} 1 & X_1 \\ 1 & X_2 \\ \vdots & \vdots \\ 1 & X_n \end{bmatrix}$

No caso de regressão linear múltipla, com várias variáveis preditoras, as fórmulas são totalmente similares, aumentando-se apenas o número de colunas da matriz de desenho

Modelos de Regressão – Forma Matricial

Vetor de parâmetros = $$\beta_{2\times 1} = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}$$

Vetor de resíduos = $$\epsilon_{n\times 1} = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

Vetor de respostas = $$\mathbf{Y}_{n\times 1} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}$$

Modelos de Regressão – Forma Matricial

Equação matricial:

$$\begin{aligned} \mathbf{Y} &= \mathbf{X}\beta + \epsilon \\ \mathbf{Y}_{n\times 1} &= \mathbf{X}_{n\times 2}\beta_{2\times 1} + \epsilon_{n\times 1} \end{aligned}$$

Os resíduos são normais, independentes e identicamente distribuídos (possuem correlação igual a zero). A matriz de covariância do vetor de resíduos é dada por:

$$\sigma^{\mathbf{2}}\{\epsilon\}_{n\times n} = Cov \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix} = \sigma^2 \mathbf{I}_{n\times n} = \begin{bmatrix} \sigma^2 & 0 & \cdots & 0 \\ 0 & \sigma^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma^2 \end{bmatrix}$$

Portanto, os resíduos possuem distribuição normal multivariada, com médias zero e matriz de covariância dada pela matriz anterior

Modelos de Regressão – Forma Matricial

Equação matricial:

$$\begin{aligned} \mathbf{Y} &= \mathbf{X}\beta + \epsilon \\ \mathbf{Y}_{n\times 1} &= \mathbf{X}_{n\times 2}\beta_{2\times 1} + \epsilon_{n\times 1} \end{aligned}$$

A matriz de covariância do vetor de respostas resíduos é dada por:

$$\sigma^{\mathbf{2}}\{\mathbf{Y}\}_{n\times n} = Cov \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = \sigma^2 \mathbf{I}_{n\times n}$$

Portanto, a matriz de respostas possui distribuição normal multivariada, com médias dadas pelo vetor $\mathbf{X} \times \beta$, e matriz de covariância dada pela matriz anterior

Modelos de Regressão – Forma Matricial

Para um vetor de parâmetros β, os resíduos podem ser escritos como:

$$\epsilon = \mathbf{Y} - \mathbf{X}\beta$$

A soma dos quadrados dos resíduos é dada pelo produto:

$$\sum \epsilon_i^2 = [\epsilon_1 \, \epsilon_2 \, \cdots \epsilon_n] \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix} = \epsilon' \epsilon$$

Pode se mostrar que o estimador de mínimos quadrados ordinários para o vetor de parâmetros β é dado por:

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$$

Modelos de Regressão – Forma Matricial

Para uma regressão linear simples, pode-se mostrar que a expressão

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$$

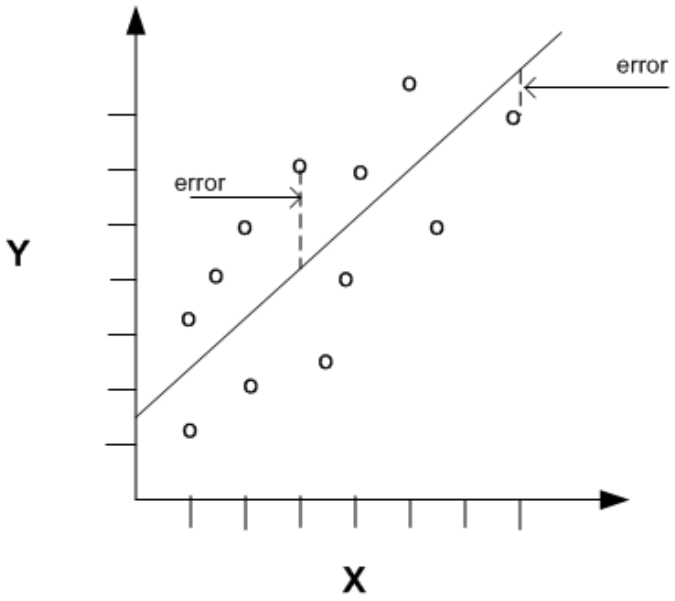
se reduz a:

$$\begin{aligned} b_1 &= \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sum(X_i - \bar{X})^2} \equiv \frac{SS_{XY}}{SS_X} \\ b_0 &= \bar{Y} - b_1\bar{X} \end{aligned}$$

Os valores preditos com a regressão são escritos como:

$$\hat{\mathbf{Y}} = \begin{bmatrix} \hat{Y}_1 \\ \hat{Y}_2 \\ \vdots \\ \hat{Y}_n \end{bmatrix} = \begin{bmatrix} b_0 + b_1X_1 \\ b_0 + b_1X_2 \\ \vdots \\ b_0 + b_1X_n \end{bmatrix} = \begin{bmatrix} 1 & X_1 \\ 1 & X_2 \\ \vdots & \vdots \\ 1 & X_n \end{bmatrix} \begin{bmatrix} b_0 \\ b_1 \end{bmatrix} = \mathbf{Xb}$$

Modelos de Regressão – Forma Matricial



Erro padrão dos resíduos da regressão:

$$S_{YX} = \sqrt{\frac{SSE}{n-2}} = \sqrt{\frac{\sum_{i=1}^{n}\left(Y - \hat{Y}_i\right)^2}{n-2}}$$

```
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.928 on 5553 degrees of freedom
Multiple R-squared:  0.6977,      Adjusted R-squared:  0.6971
F-statistic:  1281 on 10 and 5553 DF,  p-value: < 2.2e-16
```

Modelos de Regressão – Forma Matricial

- De onde vem a coluna de erros padrões dos parâmetros da regressão?

```
Coefficients:
                                             Estimate Std. Error t value Pr(>|t|)
(Intercept)                                 3.519e+01  1.202e+00  29.278  < 2e-16 ***
dados3$renda_per_capita                    -6.809e-02  4.012e-03 -16.975  < 2e-16 ***
I(renda_per_capita^2)                       6.170e-05  4.105e-06  15.030  < 2e-16 ***
I(renda_per_capita^3)                      -1.753e-08  1.400e-09 -12.523  < 2e-16 ***
dados3$indice_gini                          1.178e+00  1.492e+00   0.789 0.429906
dados3$salario_medio_mensal                -9.563e-02  9.277e-02  -1.031 0.302675
dados3$perc_criancas_extrem_pobres         -2.928e-02  1.280e-02  -2.287 0.022230 *
dados3$perc_criancas_pobres                 6.767e-02  1.418e-02   4.772 1.87e-06 ***
dados3$perc_pessoas_dom_agua_estogo_inadequados  2.714e-02  6.010e-03   4.517 6.40e-06 ***
dados3$perc_pessoas_dom_paredes_inadequadas  2.632e-02  7.768e-03   3.389 0.000708 ***
dados3$perc_pop_dom_com_coleta_lixo         2.948e-03  6.368e-03   0.463 0.643460
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.928 on 5553 degrees of freedom
Multiple R-squared:  0.6977,  Adjusted R-squared:  0.6971
F-statistic:  1281 on 10 and 5553 DF,  p-value: < 2.2e-16
```

- A partir dessa coluna de erros padrões, encontram-se:
    - intervalos de confiança
    - estatísticas testes
    - p-valores

Modelos de Regressão – Forma Matricial

Matriz de variância-covariância para os estimadores dos coeficientes da regressão:

$$\hat{\Sigma} = \begin{bmatrix} \hat{\mathrm{Var}}(\hat{\beta}_0) & \hat{\mathrm{Cov}}(\hat{\beta}_0, \hat{\beta}_1) & \hat{\mathrm{Cov}}(\hat{\beta}_0, \hat{\beta}_2) & \hat{\mathrm{Cov}}(\hat{\beta}_0, \hat{\beta}_3) \\ \hat{\mathrm{Cov}}(\hat{\beta}_1, \hat{\beta}_0) & \hat{\mathrm{Var}}(\hat{\beta}_1) & \hat{\mathrm{Cov}}(\hat{\beta}_1, \hat{\beta}_2) & \hat{\mathrm{Cov}}(\hat{\beta}_1, \hat{\beta}_3) \\ \hat{\mathrm{Cov}}(\hat{\beta}_2, \hat{\beta}_0) & \hat{\mathrm{Cov}}(\hat{\beta}_2, \hat{\beta}_1) & \hat{\mathrm{Var}}(\hat{\beta}_2) & \hat{\mathrm{Cov}}(\hat{\beta}_2, \hat{\beta}_3) \\ \hat{\mathrm{Cov}}(\hat{\beta}_3, \hat{\beta}_0) & \hat{\mathrm{Cov}}(\hat{\beta}_3, \hat{\beta}_1) & \hat{\mathrm{Cov}}(\hat{\beta}_3, \hat{\beta}_2) & \hat{\mathrm{Var}}(\hat{\beta}_3) \end{bmatrix}$$

Na diagonal principal, temos as variâncias das estimativas para cada coeficiente.
Fora da diagonal principal, temos as covariâncias entre as estimativas.

Fórmula matricial para a matriz de variância-covariância:

$$\hat{\Sigma} = \hat{\sigma}^2(\mathrm{X}'\mathrm{X})^{-1}$$

Na qual $\hat{\sigma}^2$ é a variância estimada para os erros da regressão (com *k* variáveis explicativas):

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n}[y_i - \hat{y}_i]^2}{n-k-1}$$

Modelos de Regressão – Forma Matricial

- Exemplo no programa “Analise_de_Regressao_Linear_Exercicios_Praticos_2.R”

```
#-------------------------------------------------------------------------
#---- exemplos de expressões matriciais em R para modelos de regressão
#-------------------------------------------------------------------------

mod1.X <- lm(mort_infantil ~ renda_per_capita
             + salario_medio_mensal
             + perc_criancas_extrem_pobres
             + perc_pessoas_dom_agua_estogo_inadequados
             + perc_pop_dom_com_coleta_lixo, data = dados3)
summary(mod1.X)

X1 <- model.matrix(mod1.X) #---- design matrix para o modelo de regressão
head(X1)
tail(X1)
df.X1 <- as.data.frame(X1)  #---- transformando em data.frame para visualização mais fácil
View(df.X1)

mod2.X <- lm(mort_infantil ~ renda_per_capita + as.factor(Regiao), data = dados3)
summary(mod2.X)

X2 <- model.matrix(mod2.X)
tail(X2)
head(X2)
df.X2 <- as.data.frame(X2)  #---- transformando em data.frame para visualização mais fácil
View(df.X2)
```

Modelos de Regressão – Forma Matricial

- Exemplo no programa “Analise_de_Regressao_Linear_Exercicios_Praticos_2.R”

```
#--- desvio padrão e variância dos resíduos da regressão - cálculo manual

n <- nrow(X1)      #--- número de observações
k <- ncol(X1) - 1  #--- número de var explicativas
n;k

mod1.residuos <- mod1.X$residuals
head(mod1.residuos)
tail(mod1.residuos)
hist(mod1.residuos, col = 'red', breaks = 20)

mod1.residuos.var <- (t(mod1.residuos) %*% mod1.residuos) / (n-k-1)
mod1.residuos.var
mod1.residuos.desvpad <- sqrt(mod1.residuos.var)
mod1.residuos.desvpad
```

- Resultado no R:

```
> mod1.residuos.desvpad
         [,1]
[1,] 4.172513
```

Modelos de Regressão – Forma Matricial

- Comparação com o resultado direto no sumário da regressão:

```
> summary(mod1.X)

Call:
lm(formula = mort_infantil ~ renda_per_capita + salario_medio_mensal +
    perc_criancas_extrem_pobres + perc_pessoas_dom_agua_estogo_inadequados +
    perc_pop_dom_com_coleta_lixo, data = dados3)

Residuals:
     Min       1Q   Median       3Q      Max
-19.1132  -2.4589  -0.4745   1.8734  21.8341

Coefficients:
                                          Estimate Std. Error t value Pr(>|t|)
(Intercept)                              21.838201   0.677444  32.236  < 2e-16 ***
renda_per_capita                         -0.009314   0.000402 -23.166  < 2e-16 ***
salario_medio_mensal                     -0.329859   0.097846  -3.371 0.000754 ***
perc_criancas_extrem_pobres               0.198273   0.006851  28.942  < 2e-16 ***
perc_pessoas_dom_agua_estogo_inadequados  0.062409   0.006126  10.187  < 2e-16 ***
perc_pop_dom_com_coleta_lixo             -0.011618   0.006085  -1.909 0.056268 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.173 on 5558 degrees of freedom
Multiple R-squared:  0.6585,    Adjusted R-squared:  0.6582
F-statistic:  2144 on 5 and 5558 DF,  p-value: < 2.2e-16
```

Modelos de Regressão – Forma Matricial

- Exemplo no programa “Analise_de_Regressao_Linear_Exercicios_Praticos_2.R”
- Expressão para a matriz de variância-covariância: $\hat{\Sigma} = \hat{\sigma}^2(\mathrm{X}'\mathrm{X})^{-1}$

```
#--- matriz de variância-covariância e erros padrões dos coeficientes

mod1.residuos.var <- as.numeric(mod1.residuos.var)
mod1.residuos.var
sqrt(mod1.residuos.var)

cov1 <- mod1.residuos.var * (solve(t(X1) %*% X1))
cov1
diag(cov1)
erropadrao1 <- sqrt(diag(cov1))
erropadrao1
```

- Resultado no R:

```
> erropadrao1
                                (Intercept)                                renda_per_capita
salario_medio_mensal                    perc_criancas_extrem_pobres
                               0.6774444491                                0.0004020388
0.0978460997                                0.0068507226
perc_pessoas_dom_agua_estogo_inadequados                    perc_pop_dom_com_coleta_lixo
                               0.0061265144                                0.0060846035
```

Modelos de Regressão – Forma Matricial

- Comparação com o resultado direto no sumário da regressão:

```
> summary(mod1.X)

Call:
lm(formula = mort_infantil ~ renda_per_capita + salario_medio_mensal +
    perc_criancas_extrem_pobres + perc_pessoas_dom_agua_estogo_inadequados +
    perc_pop_dom_com_coleta_lixo, data = dados3)

Residuals:
     Min       1Q   Median       3Q      Max
-19.1132  -2.4589  -0.4745   1.8734  21.8341

Coefficients:
                                          Estimate Std. Error t value Pr(>|t|)
(Intercept)                              21.838201   0.677444  32.236  < 2e-16 ***
renda_per_capita                         -0.009314   0.000402 -23.166  < 2e-16 ***
salario_medio_mensal                     -0.329859   0.097846  -3.371 0.000754 ***
perc_criancas_extrem_pobres               0.198273   0.006851  28.942  < 2e-16 ***
perc_pessoas_dom_agua_estogo_inadequados  0.062409   0.006126  10.187  < 2e-16 ***
perc_pop_dom_com_coleta_lixo             -0.011618   0.006085  -1.909 0.056268 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.173 on 5558 degrees of freedom
Multiple R-squared:  0.6585,  Adjusted R-squared:  0.6582
F-statistic:  2144 on 5 and 5558 DF,  p-value: < 2.2e-16
```

Modelos de Regressão – Forma Matricial

- Exemplo no programa “Analise_de_Regressao_Linear_Exercicios_Praticos_2.R”
- Expressão para os coeficientes estimados: $\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$

```
#--- coeficientes estimados, estatística teste e pvalores

Y1 <- dados3$mort_infantil
beta1 <- (solve(t(X1) %*% X1)) %*% (t(X1) %*% Y1) #--- coeficientes
beta1

estatistica_t1 <- beta1 / erropadrao1 #--- estatística teste t
estatistica_t1

pvalor1 <- 2*(1 - pt(abs(estatistica_t1), n-k-1)) #--- p-valores (com t-Student)
pvalor1

resultados1 <- cbind(beta1, erropadrao1, estatistica_t1, pvalor1) #--- juntando tudo
resultados1
```

Modelos de Regressão – Forma Matricial

- Resultado no R:

```
> resultados1
                                                    erropadrao1
(Intercept)                                  21.838200883 0.6774444491  32.236150 0.0000000000
renda_per_capita                             -0.009313502 0.0004020388 -23.165679 0.0000000000
salario_medio_mensal                         -0.329858957 0.0978460997  -3.371202 0.0007535145
perc_criancas_extrem_pobres                   0.198273447 0.0068507226  28.941976 0.0000000000
perc_pessoas_dom_agua_estogo_inadequados      0.062409485 0.0061265144  10.186785 0.0000000000
perc_pop_dom_com_coleta_lixo                 -0.011617670 0.0060846035  -1.909355 0.0562677132
```

- Output tradicional usando *summary* da regressão (para comparação):

```
Coefficients:
                                              Estimate Std. Error t value Pr(>|t|)
(Intercept)                                  21.838201   0.677444  32.236  < 2e-16 ***
renda_per_capita                             -0.009314   0.000402 -23.166  < 2e-16 ***
salario_medio_mensal                         -0.329859   0.097846  -3.371 0.000754 ***
perc_criancas_extrem_pobres                   0.198273   0.006851  28.942  < 2e-16 ***
perc_pessoas_dom_agua_estogo_inadequados      0.062409   0.006126  10.187  < 2e-16 ***
perc_pop_dom_com_coleta_lixo                 -0.011618   0.006085  -1.909 0.056268 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Modelos de Regressão – Forma Matricial

- **Exercício 5 - para entregar em 2 semanas:**
  - Como de costume, os exercícios podem ser entregues em grupos de 2 ou três alunos, e o grupo deve submeter o código em R utilizado para responder ao exercício, juntamente com a discussão dos resultados
  - Utilize como base o código em R 'Analise_de_Regressao_Linear_Exercicios_Praticos_2'
  - Rode a regressão de acordo com o modelo abaixo:

```
mod2.X <- lm(mort_infantil ~ renda_per_capita + as.factor(Regiao), data = dados3)
summary(mod2.X)

X2 <- model.matrix(mod2.X)
tail(X2)
head(X2)
df.X2 <- as.data.frame(X2)  #---- transformando em data.frame para visualização mais fácil
View(df.X2)
```

Modelos de Regressão – Forma Matricial

- **Exercício 5 (continuação):**
  - Questão 1: Encontre os coeficientes estimados da regressão, utilizando a fórmula matricial, conforme exemplo anterior.
  - Questão 2: Encontre os erros padrões para os coeficientes estimados da regressão, utilizando a fórmula matricial, conforme exemplo anterior.
  - Questão 3: Encontre as estatísticas teste para os coeficientes estimados da regressão, utilizando a fórmula matricial, conforme exemplo anterior.
  - Questão 4: Encontre os p-valores, utilizando a distribuição t-Student, para os coeficientes estimados da regressão, utilizando a fórmula matricial, conforme exemplo anterior.
  - Questão 5: Compare os resultados obtidos com as fórmulas matriciais, com os resultados obtidos via *summary* na regressão utilizada.

Modelos de Regressão – Forma Matricial

- **Exercício 5 (continuação):**
  - Questão 6: Conforme você observará neste exercício, a variável dummy para o Sudeste é não significativa. Vamos então excluir somente esta dummy da regressão. Para isso, você irá seguir os seguintes passos:

1. Exclua a coluna para a dummy do Sudeste da matriz de desenho deste exercício. Para isso, utilize os comandos:

```
#--- excluindo uma coluna da matriz de desenho X2

head(X2) #-- antes da exclusão
X2 <- X2[,!(colnames(X2) %in% c("as.factor(Regiao)Sudeste"))]
head(X2) #-- depois da exclusão
```

2. Refaça os passos de 1 a 4 anteriores, com a nova matriz de desenho X2, depois de excluir a coluna para a *dummy* do Sudeste. Os coeficientes remanescentes são todos significativos a 1%?

Seleção de Variáveis em Modelos de Regressão

Seleção de Variáveis

- Considere agora o modelo geral de regressão:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ \ldots + \beta_k x_{ki} + \epsilon_i$$

- Imagine que o nosso objetivo é de fazer previsões sobre a variável y_i com base em conjunto possível de variáveis preditoras
- Em geral, a inclusão adicional de variáveis preditoras na regressão, como já mencionamos aumenta o coeficiente de determinação (R^2)
- A inclusão adicional de variáveis preditoras também reduz (mesmo que marginalmente) o erro de previsão (*dentro da amostra*) mesmo que as variáveis preditoras não façam sentido
- Portanto, quanto mais incluímos variáveis explicativas na regressão, o R^2 aumenta e a soma do quadrado dos erros diminui

$$SQE = \sum_{i=1}^{n} [y_i - \hat{y}_i]^2$$

Seleção de Variáveis

- O problema é que a soma do quadrados do erros SQE corresponde aos erros da regressão dentro da amostra (*in-sample error*)
- Nós gostaríamos de ter um modelo de regressão que possa ter boas previsões para dados fora da amostra
    - Exemplo: queremos um modelo para avaliar a probabilidade de sucesso de novos cursos, com base em uma base de dados histórica de cursos anteriores, que fracassaram ou foram bem sucedidos
- Portanto, nós gostaríamos de ter um modelo que apresentasse baixo erro de previsão fora da amostra (*out-of-sample error*)
- Essa ideia de termos um bom modelo para previsão fora da amostra está intrinsicamente ligada aos procedimentos de validação cruzada (*cross-validation*) de um determinado modelo de regressão
    - A ideia da validação cruzada é dividir a amostra disponível em duas subamostras; por exemplo, uma delas com 80% das observações, e a outra com 20%.
    - Essa divisão tem que ser cuidadosa, para manter um certo balanço das informações em cada uma delas.

Seleção de Variáveis

- Validação cruzada:
  - Dividimos a amostra em duas partes – podemos fazer uma divisão aleatória entre as observações que vão entrar em cada subamostra; a primeira amostra com n_1 observações e a segunda com n_2 observações
  - A primeira subamostra é usada para estimar os coeficientes da regressão $(\beta_0, \beta_1, \beta_2, \dots, \beta_k)$
  - Usamos os coeficientes estimados na primeira amostra para prever a variável reposta na segunda amostra
  - Calculamos agora o erro médio quadrático de previsão (*mean square prediction error*) com base apenas na segunda amostra

$$MSPE = \frac{1}{n_2}\sum_{i=1}^{n_2}[y_i - \hat{y}_i]^2$$

  - Podemos então procurar o modelo de regressão, com as variáveis preditoras, que nos forneça o menor MSPE
  - O MSPE nos dá uma ideia do erro fora da amostra
- Um dos pacotes em R para previsão: "caret" (http://topepo.github.io/caret/index.html)

Façam suas Apostas!

- Vamos agora testar o erro de previsão via *cross-validation* para três modelos
- Vamos dividir a amostra em 20% e 80% aleatoriamente, baleando-se por macrorregião; evitamos assim que uma região fique subrepresentada em uma das amostras
    - Amostra de treinamento: “dadosTrain”
    - Amostra de test: “dadosTest”
- Usaremos o pacote “caret” em R
- Usaremos 80% da amostra para estimação e 20% para testar os erros de previsão de cada modelo
- Três modelos serão avaliados – qual a sua aposta?
    - Valendo camarote VIP do Weslley Safadão

Façam suas Apostas!

- Vamos agora testar o erro de previsão via *cross-validation* para três modelos

- Vamos dividir a amostra em 20% e 80% aleatoriamente, baleando-se por macrorregião; evitamos assim que uma região fique subrepresentada em uma das amostras

    - Amostra de treinamento: “dadosTrain”
    - Amostra de test: “dadosTest”

```
set.seed(2104)
trainIndex <- createDataPartition(dados3$Regiao,
               p = .8, list = FALSE, times = 1) #-- balanceando entre regiões

head(trainIndex)

dadosTrain <- dados3[ trainIndex,] #--- amostra de treinamento
dadosTest  <- dados3[-trainIndex,] #--- amostra usada para testar a previsão

table(dadosTrain$Regiao)
table(dadosTest$Regiao)
```

Façam suas Apostas!

- Modelo 1:

```
mod1 <- lm(mort_infantil ~ renda_per_capita
        + I(renda_per_capita^2)
        + I(renda_per_capita^3)
        + indice_gini
        + salario_medio_mensal
        + perc_criancas_extrem_pobres
        + perc_criancas_pobres
        + perc_pessoas_dom_agua_estogo_inadequados
        + perc_pessoas_dom_paredes_inadequadas
        + perc_pop_dom_com_coleta_lixo, data = dadosTrain)
summary(mod1)
```

Façam suas Apostas!

- Modelo 2:

```
mod2 <- lm(mort_infantil ~ renda_per_capita
           + indice_gini
           + salario_medio_mensal
           + perc_criancas_extrem_pobres
           + perc_criancas_pobres
           + perc_pessoas_dom_agua_estogo_inadequados
           + perc_pessoas_dom_paredes_inadequadas
           + perc_pop_dom_com_coleta_lixo
           + perc_pop_rural
           + as.factor(Regiao)
           + as.factor(Regiao)*renda_per_capita, data = dadosTrain)
summary(mod2)
```

Façam suas Apostas!

- Modelo 3:

```
mod3 <- lm(mort_infantil ~ renda_per_capita
        + indice_gini
        + salario_medio_mensal
        + perc_criancas_extrem_pobres
        + perc_criancas_pobres
        + perc_pessoas_dom_agua_estogo_inadequados
        + perc_pessoas_dom_paredes_inadequadas
        + perc_pop_dom_com_coleta_lixo
        + perc_pop_rural, data = dadosTrain)
summary(mod3)
```

Façam suas Apostas!

- Comparando os três modelos:

```
mod1.pred <- predict(mod1, newdata = dadosTest, se.fit = T)
mod2.pred <- predict(mod2, newdata = dadosTest, se.fit = T)
mod3.pred <- predict(mod3, newdata = dadosTest, se.fit = T)

mod1.pred.error <- mod1.pred$fit - dadosTest$mort_infantil
mod2.pred.error <- mod2.pred$fit - dadosTest$mort_infantil
mod3.pred.error <- mod3.pred$fit - dadosTest$mort_infantil

mod1.mspe <- mean(mod1.pred.error^2)
mod2.mspe <- mean(mod2.pred.error^2)
mod3.mspe <- mean(mod3.pred.error^2)

mod1.mspe
mod2.mspe
mod3.mspe
```

Seleção de Variáveis

- *K-fold cross-validation*:

  Atualmente, uma regra de ouro para a seleção de modelos de previsão baseia-se na metodologia chamada *K-fold cross-validation*

  1. Nesse caso, dividimos (em geral, aleatoriamente) a amostra total de dados em K subamostras; Em geral, usa-se $K = 10$; podemos usar também $K = 5$ ou $K = 20$

  2. Depois de dividir a amostra em $K = 10$ partes, separamos a primeira dessas partes, e estimamos os coeficientes da regressão com base nas outras nove partes conjuntamente

  3. Calculamos agora o erro médio quadrático de previsão (*mean square prediction error*) com base apenas na primeira das 10 partes

$$MSPE_1 = \frac{1}{n_1}\sum_{i=1}^{n_1}[y_i - \hat{y}_i]^2$$

  4. Repetimos os passos 2 e 3 mais nove vezes, cada uma das vezes deixando um 1/10 da amostra de fora estimações, e depois calculando o erro médio de previsão justamente na subamostra deixada de fora

  5. Combinamos os erros médios quadráticos das 10 partes para chegarmos a uma medida agregada do erro de previsão fora da amostra

Seleção de Variáveis

- Em geral, a inclusão indiscriminada de novas variáveis preditoras, apesar de reduzir o erro dentro da amostra, acaba aumentando o erro fora da amostra
- Por outro lado, a não inclusão de variáveis preditoras importantes pode também causar também uma perda de poder de previsão fora da amostra
- Portanto, todos os métodos de seleção automática de variáveis na literatura consideram essa relação de compromisso entre o aumento do poder explicatório da regressão versus a parcimônia na especificação
- Chamamos de ***trade-off viés-variância***
    - Quando incluímos variáveis, ***reduzimos o viés*** do modelo (é possível capturar mais especificidades da relação entre preditores e variável resposta)
    - No entanto, quando incluímos variáveis, temos que estimar mais parâmetros e a imprecisão (***variância***) de cada um deles aumenta
- Uma série de técnicas e indicadores foram criados para encontrarmos modelos para atender a essa relação de compromisso, sem necessariamente termos que recorrer à validação cruzada
    - Vamos estudar algumas dessas técnicas nos próximos slides

Método de Máxima Verossimilhança

- Para um conjunto de parâmetros $\beta_0, \beta_1, \beta_2, \ldots, \beta_k$ para um modelo de regressão
- A função de verossimilhança, assumindo que os resíduos da regressão são normais, independentes e identicamente distribuídos, é escrita como

$$L(\beta_0, \beta_1, \ldots, \beta_k, \sigma^2) = \prod_{i=1}^{n} \frac{1}{(2\pi\sigma^2)^{\frac{1}{2}}} \mathrm{e}^{\left[-\frac{1}{2\sigma^2}(y_i - \beta_0 - \beta_1 x_{1i} - \ldots - \beta_k x_{ki})^2\right]}$$

- O termo σ^2 corresponde à variância dos resíduos da regressão, que também precisa ser estimada
- O método de máxima verossilhança é comumente empregado para estimar os parâmetros do modelo de regressão linear
- Pode-se mostrar que a fórmula para a estimativa dos coeficientes via máxima verossilhança é idêntica à fórmula para estimativa via método de mínimos quadrados ordinários
- A diferença entre os dois métodos está na estimativa da variância σ^2

Método de Máxima Verossimilhança

- Pelo método de mínimos quadrados ordinário (MQO), a estimativa de σ^2 é dada por

$$s^2 = \frac{SQE}{n-k-1} = \frac{\sum_{i=1}^{n}[y_i - \hat{y}_i]^2}{n-k-1}$$

- Pelo método de máxima verossimilhança, a estimativa de σ^2 tem expressão

$$\hat{\sigma}^2 = \frac{SQE}{n} = \frac{\sum_{i=1}^{n}[y_i - \hat{y}_i]^2}{n}$$

- A estimativa via MQO para σ^2 é não viesada; porém, o viés na estimativa via máxima verossimilhança desaparece quando o número de observações na amostra aumenta

- Por motivos numéricos e analíticos, trabalhamos com o log da função de verossimilhança, ao invés da função original

$$\log L(\beta_0, \beta_1, \ldots, \beta_k, \sigma^2) = -\frac{n}{2}\log 2\pi\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_{1i} - \ \ldots - \beta_k x_{ki})^2$$

Método de Máxima Verossimilhança

- Por motivos numéricos e analíticos, trabalhamos com o log da função de verossimilhança, ao invés da função original

$$\log L(\beta_0, \beta_1, \dots, \beta_k, \sigma^2) = -\frac{n}{2}\log 2\pi\hat{\sigma}^2 - \frac{n}{2} = -\frac{\mathrm{n}}{2}(1 + \log 2\pi\hat{\sigma}^2)$$

onde $\hat{\sigma}^2 = \frac{SQE}{n} = \frac{\sum_{i=1}^{n}[y_i - \hat{y}_i]^2}{n}$

- Portanto, quanto menor a SQE, maior a função de log verossimilhança
- O parâmetro $\hat{\sigma}^2$ é calculado usando-se os erros dentro da amostra; portanto, quando incluímos mais variáveis (mesmo desnecessárias) no modelo, o erro diminui e a função de log verossimilhança aumenta
- Diversos indicadores surgiram com base em penalizações da função de log verossimilhança para a inclusão de mais variáveis na regressão
- Critérios comuns: AIC, BIC

Seleção de Variáveis

- Critério de Informação de Akaike - AIC

$$AIC = -2 \log L(\beta_0, \beta_1, \dots, \beta_k, \sigma^2) + 2 \times p$$

O número *p* corresponde ao número de parâmetros livres na regressão. No caso da regressão linear, temos: um intercepto, k variáveis preditoras, a variância dos resíduos

$$p = 1 + k + 1 = 2 + k$$

No caso de regressão linear, o AIC é equivalente o critério *Cp* de Mallow

- Critério de Informação Bayesiano - BIC

$$BIC = -2 \log L(\beta_0, \beta_1, \dots, \beta_k, \sigma^2) + \log n \times p$$

- Os termos $[2 \times p]$ e $[\log n \times p]$, no AIC e BIC, correspondem a pênaltis para a inclusão adicional de variáveis
- Portanto, a inclusão de variáveis vai aumentar $\log L(\beta_0, \beta_1, \dots, \beta_k, \sigma^2)$, mas aumenta também os pênaltis $[2 \times p]$ e $[\log n \times p]$

Seleção de Variáveis

- Portanto, quando da escolha de um melhor modelo, selecionar aquele que resulte em menor BIC ou menor AIC
- O pênalti no BIC é mais pesado do que no AIC; consequência: o BIC em geral indica a escolha de modelos mais parcimoniosos
- Outros critérios existem, considerando-se outros pênaltis para o número de parâmetros livres *p*, entre eles:
    - Critério de informação de Hannan-Quinn
    - AIC corrigido
- No R:

```
AIC(mod1)
AIC(mod2)
AIC(mod3)

BIC(mod1)
BIC(mod2)
BIC(mod3)
```

Seleção Automática de Variáveis

- Imagine agora queremos encontrar automaticamente um conjunto de variáveis que resulte em um melhor modelo para fins de previsão
- Diversas possibilidades existem na literatura, entre elas:
  - Seleção *best subset* *
  - Seleção *stepwise*
  - Seleção *backwards*
  - Seleção *foward*
  - Regressão *ridge*
  - Lasso
- Todas elas buscam satisfazer a relação de compromisso entre erro dentro da amostra e parcimônia do modelo
- O R possui ferramentas para utilização dos métodos acima

Seleção Automática de Variáveis

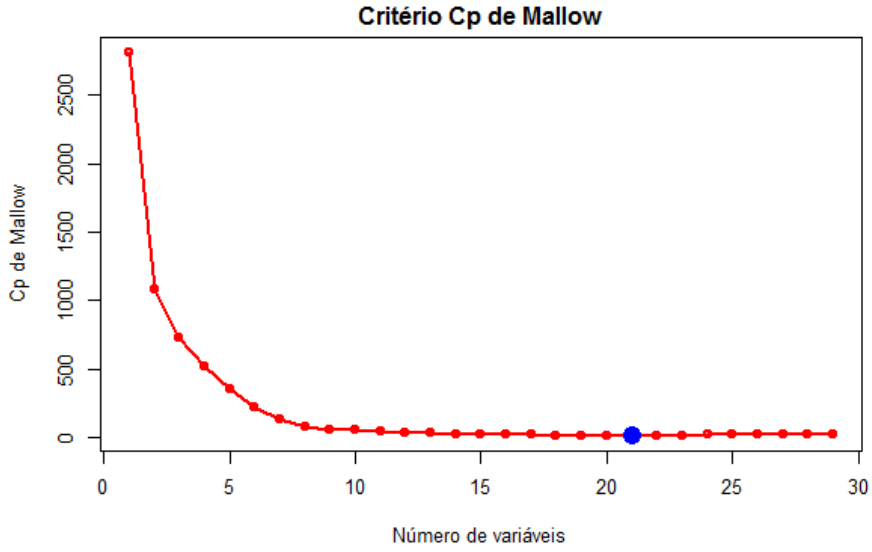
- Seleção *best subset*
  - Varre todas as combinações possíveis de variáveis preditoras para encontrar o conjunto com melhor R^2 ajustado ou melhor critério Cp de Mallow, por exemplo
  - Computacionalmente, pode ser bastante demandante, e pode se tornar inviável quando temos muitas variáveis candidatas
  - Se tivermos M variáveis candidatas, há 2^M conjuntos possíveis; por exemplo, M = 100, há 1,268 x 10^{30} regressões possíveis
- No R:
  - Pacote “leaps”
  - Para cada número de variáveis, encontra o modelo com menos SQE
  - Podemos analisar o valor do Cp (AIC), do BIC, do R^2 ajustado e do R^2 para cada número de variáveis
  - Para cada número de variáveis na regressão, o pacote “leaps” encontra o melhor conjunto de variáveis

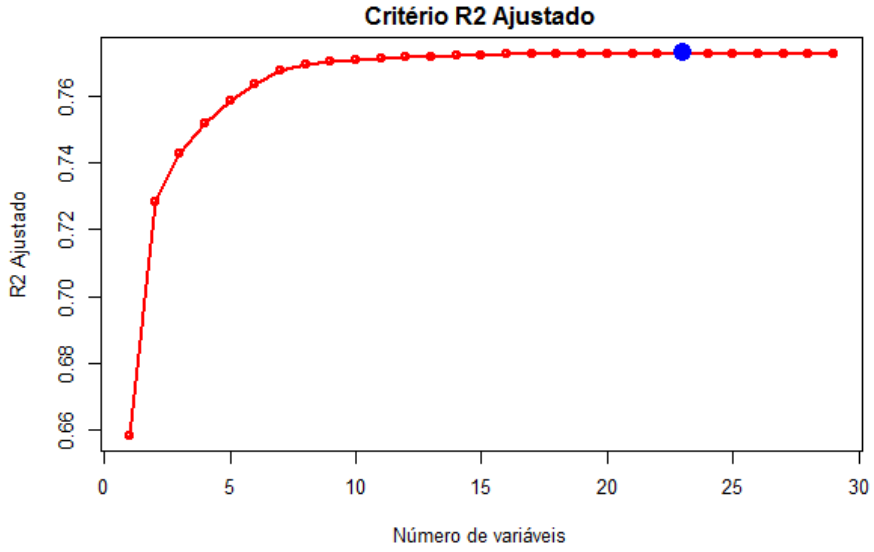
Seleção Automática de Variáveis

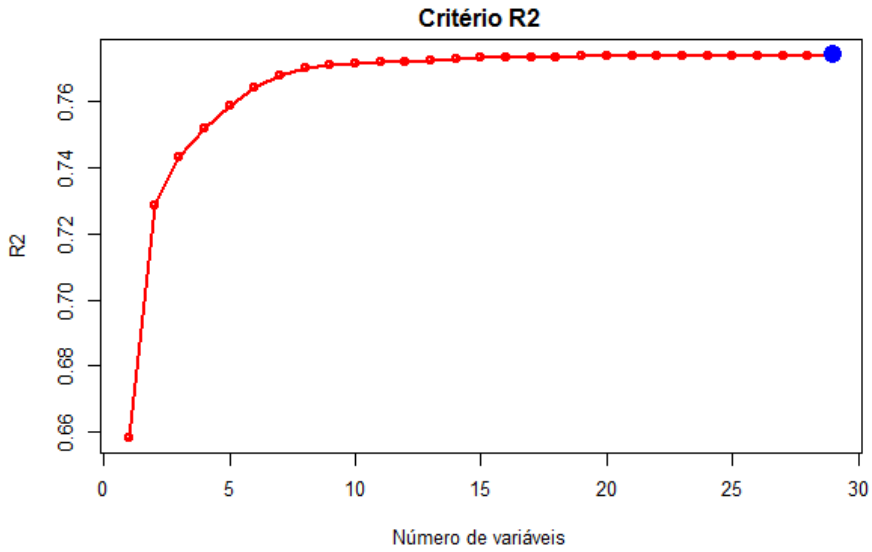
- Modelo completo:

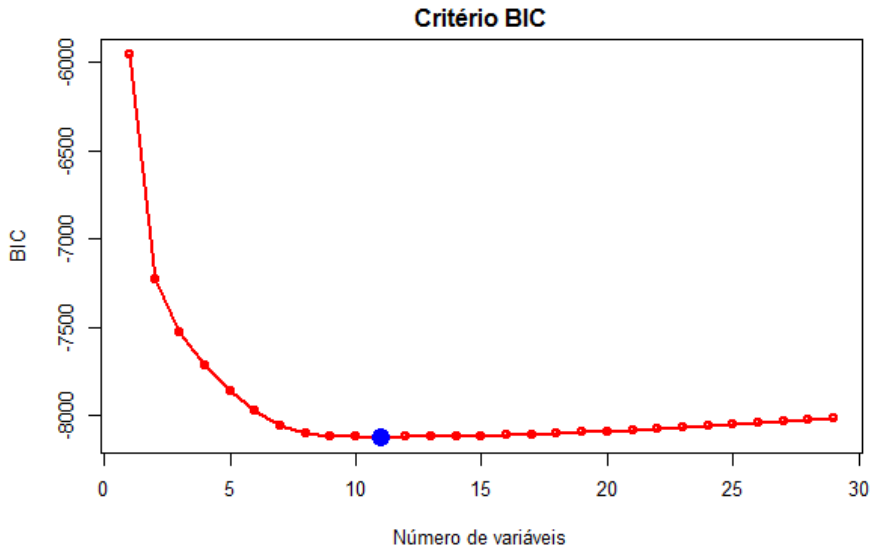
```
bestsub <- regsubsets(mort_infantil ~ renda_per_capita
          + I(renda_per_capita^2)
          + I(renda_per_capita^3)
          + I(renda_per_capita^4)
          + I(renda_per_capita^5)
          + indice_gini
          + I(indice_gini^2)
          + I(indice_gini^3)
          + I(indice_gini^4)
          + I(indice_gini^5)
          + salario_medio_mensal
          + I(salario_medio_mensal^2)
          + I(salario_medio_mensal^3)
          + I(salario_medio_mensal^4)
          + I(salario_medio_mensal^5)
          + perc_criancas_extrem_pobres
          + perc_criancas_pobres
          + perc_pessoas_dom_agua_estogo_inadequados
          + perc_pessoas_dom_paredes_inadequadas
          + perc_pop_dom_com_coleta_lixo
          + perc_pop_rural
          + as.factor(Regiao)
          + as.factor(Regiao)*renda_per_capita, data = dados3,
          nvmax = 50)
```

Seleção Automática de Variáveis









Seleção Automática de Variáveis

- Para um número grande (maior do que 50 ou 100) de potenciais preditores, a opção de *best subset* pode ser inviável computacionalmente
- Alternativas computacionalmente viáveis incluem:
  - Seleção *stepwise*
  - Seleção *backwards*
  - Seleção *foward*
- Seleção *forward*:
  1. Comece com uma regressão com apenas o intercepto
  2. Para as demais variáveis candidatas, escolha aquela cuja inclusão implica em maior aumento de R^2
  3. Se essa nova adição foi estatisticamente significante, mantenha a variável; caso contrário, retire a variável, volte ao modelo anterior, e pare o algoritmo
  4. Repita os passos 2 e 3 até que a adição de qualquer nova variável não seja estatisticamente significante (a um nível de significância pré-especificado)

Seleção Automática de Variáveis

- Seleção *backwards*:

    1. Comece com uma regressão com todas as variáveis candidatas
    2. Se houver alguma variável cujo coeficiente é estatisticamente não significativo, elimine a variável que tenha menor nível de significância no modelo (maior p-valor); caso contrário, esse é o modelo final
    3. Repita o passo 2 até atingir um modelo no qual todas as variáveis são estatisticamente significantes (a um nível de significância pré-definido)

- Seleção *Stepwise*:

    1. Trata-se de uma combinação das seleções do tipo *forward* e *backwards*
    2. Os passos *forward* e *backwards* são intercalados, de forma a adicionarmos variáveis que sejam significativas e retirarmos variáveis que não sejam estatisticamente significativas
    3. O algoritmo para quando não for mais possível adicionar variáveis novas que sejam estatisticamente significantes, ou retirar variáveis incluídas que forem estatisticamente não significantes

- Os passos acima dão uma ideia geral dos algoritmos; diferentes softwares possuem versões que são variações ao redor dessa ideia geral

```r
#-----------------------------------------------------------------------
#---- Backwards, forward e stepwise selection
#-----------------------------------------------------------------------

mod.full <- lm(mort_infantil ~ renda_per_capita
               + I(renda_per_capita^2)
               + I(renda_per_capita^3)
               + indice_gini
               ......
               + as.factor(Regiao)
               + as.factor(Regiao)*renda_per_capita, data = dados3)
summary(mod.full)

step1 <- step(mod.full, direction = "backward")
summary(step1)

step2 <- step(mod.full, direction = "forward")
summary(step2)

step3 <- step(mod.full, direction = "both")
summary(step3)
formula(step3)

mod.step3 <- lm(formula = formula(step3), data = dados3)
summary(mod.step3)
```

- **Exercício 6 - para entregar em 2 semanas:**
  - Utilize como base o código em R 'Analise_de_Regressao_Linear_Exercicios_Praticos_2'. Considere o modelo completo abaixo. Usando os diversos métodos aprendidos em sala de aula, encontre um modelo, subconjunto do modelo abaixo, que apresente o menor AIC. O grupo de alunos que obtiver o modelo com AIC de menor valor terá a nota deste exercício multiplicada por dois. No resultado entregue, você deverá incluir o código em R para obter o melhor modelo, e deverá incluir também a fórmula em R para essa "melhor" regressão

```
mod.full <- lm(mort_infantil ~ renda_per_capita
               + I(renda_per_capita^2)
               + I(renda_per_capita^3)
               + I(renda_per_capita^4)
               + I(renda_per_capita^5)
               + indice_gini
               + I(indice_gini^2)
               + I(indice_gini^3)
               + I(indice_gini^4)
               + I(indice_gini^5)
               + salario_medio_mensal
               + I(salario_medio_mensal^2)
               + I(salario_medio_mensal^3)
               + I(salario_medio_mensal^4)
               + I(salario_medio_mensal^5)
               + perc_criancas_extrem_pobres
               + perc_criancas_pobres
               + perc_pessoas_dom_agua_estogo_inadequados
               + perc_pessoas_dom_paredes_inadequadas
               + perc_pop_dom_com_coleta_lixo
               + perc_pop_rural
               + as.factor(Regiao)
               + as.factor(Regiao)*renda_per_capita, data = dados3)
```

Apêndice 1

Testes de Hipóteses
para Médias Amostrais

Testes de Hipóteses

- Imagine agora o seguinte problema: um acordo internacional entre países desenvolvidos e em desenvolvimento estabeleceu um sistema de transferências de recursos para unidades da federação de países em desenvolvimento. No entanto, entre os condicionantes das transferências está a exigência de que a média populacional de anos de estudo seja maior do que 6 anos, para indivíduos maiores do que 15 anos

- Para uma determinada UF, a média populacional estimada anos antes era de 4,2 anos de estudo. Foi então coletada uma nova amostra de indivíduos maiores de 15 anos, com *n* = 82. A média dessa amostra de 82 observações foi igual a 6.12 anos. Portanto, a estimativa de anos de estudo nessa nova amostra é $\bar{x} = 6.12$.

    - Pergunta: é justo ou não efetuar a transferência, de acordo com os resultados dessa amostra?

    - É possível que a UF de fato tenha média de escolaridade até menor do que 6 anos de estudo, mas por sorte a amostra coletada resultou em uma média de 6.12 anos?

Testes de Hipóteses

- As instituições internacionais podem argumentar que amostras com média amostral de 6.12 ou maiores até são comuns, e não necessariamente a UF cumpriu o requerimento para as transferências

- E agora? Como resolver esse impasse? Digamos que de fato a instituição internacional estivesse correta e que 6.12 até seja um valor "comum" para amostras com 82 observações, mesmo com média populacional (desconhecida) menor do que 6

    - Pergunta: qual seria então um valor de corte razoável, a partir do qual poderíamos argumentar com certo grau de "certeza" que a UF de fato tem média populacional de escolaridade maior do que 6 anos de estudo?

    - Será que para esse valor de corte razoável, há também a probabilidade de ainda assim a UF estar recebendo injustamente as transferências de recursos?

    - **Nos deparamos então com dois tipos de erros possíveis:**

        - **Não fazemos a transferência, quando de fato a UF é merecedora da mesma**

        - **Fazemos a transferência, quando a UF não atingiu de fato a meta de anos de escolaridade**

Testes de Hipóteses

- Como proceder então?

- Precisamos construir uma regra de procedimentos para “julgar” em que situações seria justo ou não fazer a transferência de recursos, a depender do resultado obtido na amostra - esse tipo de procedimento chamamos de **testes de hipóteses**

Testes de Hipóteses

- Hipótese nula versus hipótese alternativa:
  - Hipótese nula – denominada H_0, corresponde à situação "status quo", ou a situação contra a qual se necessita de evidências suficientes. Em geral, a hipótese nula seria o equivalente à assertiva "todo cidadão é inocente até que se prove o contrário"
  - Hipótese alternativa – denominada H_a, corresponde à situação em geral contrária à hipótese nula. Para aceitar a hipótese alternativa, precisamos encontrar evidências suficientes na amostra
- No exemplo das transferências de recursos para as UF's, a hipótese nula seria de que a média populacional de anos de estudo na UF é menor ou igual a 6 anos
- A hipótese alternativa é de que a média populacional de anos de estudo é maior do que 6 anos
- Portanto, denominando por μ a média populacional, podemos expressar as hipóteses nula e alternativa na forma:

$$H_0: \mu \leq \mu_0 = 6$$
$$H_a: \mu > \mu_0 = 6$$

- **As hipóteses nula e alternativa** são o primeiro elemento para a construção de um teste de hipóteses

Testes de Hipóteses

- O segundo elemento para a construção de um teste de hipóteses é o que chamamos de **estatística teste**. Para diferentes testes de hipóteses, os estatísticos criaram estatísticas testes das mais variadas.

- No caso de testes da média, a estatística teste mais utilizada é a estatística teste *t*

$$t_{stat} = \frac{\bar{X} - \mu_0}{\hat{\sigma}/\sqrt{n}} = \frac{\bar{X} - 6}{\hat{\sigma}/\sqrt{n}}$$

- O terceiro elemento necessário para a construção do teste de hipóteses é a **distribuição da estatística teste sob a hipótese nula** (assumindo que a hipótese nula é verdadeira)

- Antes de continuar com a distribuição da estatística teste t_{stat}, há um fato importante sobre a distribuição normal:

  Seja ***U*** uma variável aleatória com distribuição normal com média ***m*** e desvio padrão ***s***. Então, a variável ***Z***, dada por

$$\boldsymbol{Z = \frac{U - m}{s}}$$

  tem distribuição normal padronizada (média 0 e desvio padrão 1)

Testes de Hipóteses

- Lembrando que, sob a hipótese nula, a média amostral $\bar{X}$ tem distribuição aproximadamente normal com média μ_0 (média populacional assumindo a hipótese nula) e desvio padrão $\hat{\sigma}/\sqrt{n}$.

- Portanto, utilizando o fato anterior sobre a distribuição normal, concluímos que a variável

$$t_{stat} = \frac{\bar{X} - \mu_0}{\hat{\sigma}/\sqrt{n}}$$

tem distribuição normal padronizada (com média 0 e desvio padrão igual a 1) aproximadamente. Portanto, assumindo que a hipótese nula é verdadeira, temos

$$t_{stat} \approx N(0,1)$$

- Finalmente, o quarto elemento para a construção de um teste de hipóteses é a **regra de rejeição da hipótese nula**

    - Em geral a regra de rejeição tem a forma $t_{stat} > \mathrm{c}$, onde c é um valor crítico de corte

Testes de Hipóteses

- Como escolher esse valor crítico de corte?
    - Se escolhermos um valor de corte muito alto, muito raramente iremos considerar a UF cumpridora da regra de anos de estudo > 6 anos, e portanto elegível para recebimento de transferências
    - Por outro lado, se considerarmos um valor de corte muito baixo, poderemos estar transferindo recursos para UF's não merecedoras de fato
- Vamos então pensar nos dois tipos de erros que podemos estar incorrendo quando utilizamos uma regra de rejeição da hipótese nula
    - Erro do tipo I – rejeitamos a hipótese nula quando na verdade a hipótese nula é verdadeira. No nosso exemplo, o erro do tipo I seria aceitar que a UF tem média de anos de estudo maior do que 6, quando na verdade a média da UF é menor ou igual 6

      (Em resumo: "prende o inocente")
    - Erro do tipo II – aceitamos a hipótese nula (ou seja, não transferimos recursos), quando na verdade a média populacional de anos de estudos é maior do que 6 de fato

      (Em resumo: "deixa o culpado solto")

Testes de Hipóteses

- De maneira geral, a determinação do valor de corte ***c*** considera a probabilidade de cometermos o erro tipo I.
- Tenta-se controlar a probabilidade de estarmos "condenando um inocente".
- Por um costume tribal, em geral escolhem-se probabilidades de erro tipo I (de condenar o inocente) com valores iguais a 10%, 5% ou 1%
- A probabilidade de se cometer um erro do tipo I é conhecida como **nível do teste de hipóteses**
- Vamos então assumir que a regra a ser adotada assume uma probabilidade de erro do tipo I igual a 1%. A partir daí, vamos determinar o valor de corte ***c*** na regra, de acordo com a qual rejeitamos a hipótese nula (e transferimos recursos para a UF) quando:

$$t_{stat} = \frac{\bar{X} - \mu_0}{\hat{\sigma}/\sqrt{n}} > c$$

- Para um determinado valor de corte ***c***, qual a probabilidade de rejeitar a hipótese nula, quando na verdade ela é verdadeira?

Testes de Hipóteses

- Lembrando que rejeitamos a hipótese nula quando coletamos uma amostra para a qual calculamos $\bar{X}$ e $\hat{\sigma}$ e obtemos um valor t_{stat}, com

$$t_{stat} = \frac{\bar{X} - \mu_0}{\hat{\sigma}/\sqrt{n}} > c$$

- Porém, assumindo que a hipótese nula é verdadeira, ou seja, a média populacional é igual a 6, vimos que a variável t_{stat} tem distribuição aproximadamente normal padronizada (média 0 e desvio padrão 1)

- Portanto, a probabilidade de rejeitarmos a hipótese nula, quando ela é verdadeira, corresponde à probabilidade

$$\text{Prob}\left[\frac{\bar{X} - \mu_0}{\hat{\sigma}/\sqrt{n}} > c\right] = \text{Prob}[t_{stat} > c]$$

- Essa probabilidade corresponde à cauda superior de uma distribuição normal padronizada – podemos então usar as funções no R para a distribuição normal

Testes de Hipóteses

- Podemos então calcular qual o valor de corte ***c***, para a nossa regra de rejeição da hipótese nula
- Estabelecemos que a probabilidade de erro tipo I que estamos dispostos a aceitar é igual a 1%
- Portanto, podemos escolher o valor de corte ***c*** tal que

$$\text{Prob}\left[\frac{\bar{X}-\mu_0}{\hat{\sigma}/\sqrt{n}} > c\right] = \text{Prob}[t_{stat} > c] = 1\% = 0.01$$

$$\Rightarrow \text{Prob}[\text{t}_{\text{stat}} \leq \text{c}] = 1 - 0.01 = 0.99$$

- Usando o R, queremos achar o valor c para o qual a *F(c)* = 0.99. O comando é simplesmente:

```
c <- qnorm(0.99);
```

- O obtemos um valor *c* = 2.326348
- Portanto, rejeitamos a hipótese nula (e damos recursos para a UF) quando

$$\frac{\bar{X}-\mu_0}{\hat{\sigma}/\sqrt{n}} > 2.326348$$

Testes de Hipóteses

- **Exemplo 1 – teste de hipóteses:** considere uma amostra de n = 15 observações abaixo:

  X = 5.3, 6.1, 7.2, 3.3, 5.2, 8.1, 5.3, 8.2, 5.1, 9.4, 8.3, 4.1, 5.3, 10.2, 4.5

- A média amostral é dada por $\bar{X} = 6.373$ e $\hat{\sigma} = 2.056$, e temos que o valor da estatística teste é dado por

$$t_{stat} = \frac{\bar{X} - 6}{\hat{\sigma}/\sqrt{15}} = \frac{0.373}{2.056/\sqrt{15}} = 0.7033$$

- Comparando com o valor de corte *c* = 2.326348, temos $t_{stat} < c$ e não rejeitamos a hipótese nula. Portanto, nesse exemplo, a UF não receberá os recursos (mesmo com $\bar{X} = 6.373 > 3$)

- Considere agora a amostra: X = 7.3, 6.1, 7.2, 6.3, 5.2, 8.1, 11.3, 8.2, 15.1, 9.4, 8.3, 4.1, 5.3, 10.6, 6.5

- Nesse caso, $\bar{X} = 7.933$, $\hat{\sigma} = 2.807$ e $t_{stat} = 2.667 > c = 2.326348$

- Portanto, para essa segunda amostra, rejeitamos a hipótese nula, com um teste de nível igual a 1% (probabilidade de erro do tipo I)

- Esse tipo de teste é conhecido como **teste uni-caudal** – estamos olhando a probabilidade em apenas uma das caudas

Testes de Hipóteses

- Os valores de corte para os testes de hipótese podem ser obtidos a partir dos quantis da distribuição normal, assumindo amostras com muitas observações
    - Vimos a aproximação da distribuição normal para a média amostral (média das amostras)
- Da mesma forma que no caso dos intervalos de confiança, quando o número de observações na amostra não é muito elevado (por exemplo, n < 30 ou 40), ao invés de usar os quantis da distribuição normal, podemos usar os quantis da distribuição *t-Student*
- Por exemplo, para uma amostra de n = 15 observações, podemos usar uma distribuição *t-Student* com $\nu = n - 1 = 14$ graus de liberdade
- Para um teste com probabilidade de erro tipo I igual a 1%, o valor crítico pode ser obtido com R pela função

```
c1t <- qt(0.99, df=14);
```

- Para testes com níveis de 5% ou 10% (outras probabilidades de erro tipo I), usamos comandos:

```
c1t <- qt(0.95, df=14);      ou      c1n <- qnorm(0.95);
c2t <- qt(0.90, df=14);      ou      c2n <- qnorm(0.90);
```

Testes de Hipóteses e P-Valores

- No procedimento acima para teste de hipóteses, calculamos o valor da estatística teste $t_{stat} = \frac{\bar{X}-\mu_0}{\hat{\sigma}/\sqrt{n}}$ (com base na amostra disponível) e checamos se esse valor excede um determinado valor de corte, ou **valor crítico do teste**, para um determinado nível (probabilidade de erro tipo I) de 10%, 5% ou 1%

- Alternativamente, podemos simplesmente encontrar a probabilidade da cauda à direita do valor da estatística teste t_{stat}, usando uma distribuição normal padronizada ou usando uma distribuição *t-Student* com (n-1) graus de liberdade

- Para entender a lógica desse procedimento, considere um nível do teste igual a 5%, com valor crítico igual a c = 1.645 (usando distribuição normal padronizada). Temos então duas situações:

    - Ou t_{stat} > 1.645, e nesse caso a área à direita de t_{stat} será menor do que 5% => rejeitamos a hipótese nula

    - Ou $t_{stat} \leq 1.645$, e nesse caso a área à direita de t_{stat} será maior ou igual a 5% => aceitamos a hipótese nula

- Portanto, para verificar se rejeitamos ou não a hipótese nula, em um teste de nível 5%, basta checar se a área à direita do valor t_{stat} é maior ou menor que 5%. Essa área a direita é conhecida como **p-valor** do teste

Testes de Hipóteses e P-Valores

- Usando um comando no R (primeira amostra):

```
amostra1 <- c(5.3, 6.1, 7.2, 3.3, 5.2, 8.1, 5.3, 8.2, 5.1, 9.4, 8.3, 4.1, 5.3, 10.2, 4.5)
t.test(amostra1, mu = 6, alternative = "greater")
```

- Resultado:

```
> t.test(amostra1, mu = 6, alternative = "greater")

        One Sample t-test

data:  amostra1
t = 0.70333, df = 14, p-value = 0.2467
alternative hypothesis: true mean is greater than 6
95 percent confidence interval:
 5.438413      Inf
sample estimates:
mean of x
 6.373333
```

- p-value = 0.2467 > 10% e portanto aceitamos a hipótese nula (não transferimos os recursos), mesmo com nível do teste de 10%

Testes de Hipóteses e P-Valores

- Usando um comando no R (segunda amostra):

```
amostra2 <- c(7.3, 6.1, 7.2, 6.3, 5.2, 8.1, 11.3, 8.2, 15.1, 9.4, 8.3, 4.1, 5.3, 10.6, 6.5)
t.test(amostra2, mu = 6, alternative = "greater")
```

- Resultado:

```
> t.test(amostra2, mu = 6, alternative = "greater")

        One Sample t-test

data:  amostra2
t = 2.6675, df = 14, p-value = 0.009195
alternative hypothesis: true mean is greater than 6
95 percent confidence interval:
 6.656776      Inf
sample estimates:
mean of x
 7.933333
```

- p-value = 0.009195 < 1% e portanto rejeitamos a hipótese nula (multamos a empresa), com nível do teste de 1%

Testes de Hipóteses e P-Valores

- **Exemplo 2 – testes de hipóteses:** considere uma empresa que produz e vende cerveja de trigo. Cada garrafa deveria conter 500 ml de cerveja. No entanto, uma denúncia levou a uma investigação de se a quantidade de cerveja na garrafa é de fato 500 ml, e a unidade de investigação coletou uma amostra de n = 22 garrafas. Para cada garrafa, anotou-se o total de conteúdo ml. Os números coletados, em ml, estão abaixo:

  X = 508.8, 500.0, 491.2, 509.5, 521.2, 489.3, 489.0, 515.5, 486.1, 493.5, 493.6,
  497.3, 500.0, 496.6, 499.3, 499.9, 496.4, 482.4, 498.0, 496.3, 501.8, 498.6

- A empresa será multada se ficar constatado que a média de conteúdo nas garrafas é menor do que 500 ml

- As hipóteses nula e alternativa nesse caso são (teste também **uni-caudal**)

$$H_0: \mu \geq \mu_0 = 500$$
$$H_a: \mu < \mu_0 = 500$$

- Para rejeitar a hipótese nula nesse caso, utilizamos a seguinte regra de rejeição: $t_{stat} < c$

- Note que essa regra de rejeição é o contrário da regra de rejeição do exemplo anterior

Testes de Hipóteses e P-Valores

- Temos que encontrar o valor crítico ***c***, para essa regra de rejeição. Para isso, da mesma forma que no exemplo anterior, fixamos uma probabilidade de erro tipo I, que assumiremos igual a 1% nesse caso

- Portanto, a probabilidade de condenarmos a empresa de cerveja, ela sendo inocente, é igual a 1%

- Temos então que encontrar o valor *c* tal que:

$$\text{Prob}\left[\frac{\bar{X} - \mu_0}{\hat{\sigma}/\sqrt{n}} < c\right] = \text{Prob}[t_{stat} < c] = 1\% = 0.01$$

- Se usarmos uma distribuição normal, podemos usar o R, com a função “c <- qnorm(0.01);”, obtendo c = -2.326348.

- Se usarmos uma distribuição *t-Student*, com n-1 = 21 graus de liberdade, podemos usar o R, com o comando “c <- qt(0.01, df = 21);”, obtendo c = -2.517648

- Note que, no caso de usarmos a distribuição *t-Student*, necessita-se de valores ainda mais longe de 500 ml para termos evidências para rejeitar a hipótese nula

Testes de Hipóteses e P-Valores

- Para a amostra considerada, a média amostral é igual a 498.38, e o desvio padrão amostral é igual a 9.14. O valor da estatística teste é dado por:

$$t_{stat} = \frac{\bar{X} - 500}{\hat{\sigma}/\sqrt{22}} = \frac{-1.622727}{9.138507/\sqrt{22}} = -0.8329$$

- Usando a função t.test no R, temos o comando:

```
amostra3 = c(508.8, 500.0, 491.2, 509.5, 521.2, 489.3, 489.0, 515.5, 486.1, 493.5, 493.6,
             497.3, 500.0, 496.6, 499.3, 499.9, 496.4, 482.4, 498.0, 496.3, 501.8, 498.6);
t.test(amostra3, mu = 500, alternative = "less")
```

- O resultado (não rejeitamos a hipótese nula nesse caso):

```
        One Sample t-test

data:  amostra3
t = -0.83288, df = 21, p-value = 0.2071
alternative hypothesis: true mean is less than 500
95 percent confidence interval:
     -Inf 501.7299
sample estimates:
mean of x
 498.3773
```

Testes de Hipóteses e P-Valores

- **Exemplo 3 - testes de hipótese:** considere a mesma empresa que produz e vende cerveja de trigo. Cada garrafa deveria conter 500 ml de cerveja. Do ponto de vista da empresa, para ela não é bom nem que a quantidade média seja menor do que 500 ml (poderia ser multada), nem que a quantidade média seja maior do que 500 ml. Se for maior que 500 ml, ela está perdendo dinheiro.

- O controle de qualidade da empresa coleta uma amostra de 25 observações:

X = 506.7, 505.5, 504.2, 505.9, 495.8, 508.4, 502.6, 510.1, 491.5, 498.0, 507.0, 515.3, 502.4, 509.2, 519.2, 508.4, 496.7, 503.3, 502.4, 499.8, 503.8, 503.9, 501.5, 497.3, 501.9

- Nesse caso, importa se o valor da média populacional é diferente de 500 ml (erros dos dois lados é ruim para empresa). As hipóteses nula e alternativa nesse caso são (teste também **bi-caudal**)

$$H_0: \mu = \mu_0 = 500$$
$$H_a: \mu \neq \mu_0 = 500$$

- Para rejeitar a hipótese nula nesse caso, utilizamos a seguinte regra de rejeição, com base no valor absoluto da estatística teste:

$$|t_{stat}| = \left| \frac{\bar{X} - \mu_0}{\hat{\sigma}/\sqrt{n}} \right| > \mathrm{c}$$

Testes de Hipóteses e P-Valores

- Para encontrar o valor crítico *c*, vamos escolher uma probabilidade de erro do tipo I de 5%
- Nesse caso, queremos achar o valor *c* tal que

$$\text{Prob}\left[\left|\frac{\bar{X}-\mu_0}{\hat{\sigma}/\sqrt{n}}\right| > \text{c}\right] = \text{Prob}\left[\frac{\bar{X}-\mu_0}{\hat{\sigma}/\sqrt{n}} > \text{c}\right] + \text{Prob}\left[\frac{\bar{X}-\mu_0}{\hat{\sigma}/\sqrt{n}} < -\text{c}\right] = 5\%$$

- Como estamos trabalhando com distribuições simétrica em torno do valor zero (normal padronizada ou *t-Student*), temos que:

$$\text{Prob}\left[\frac{\bar{X}-\mu_0}{\hat{\sigma}/\sqrt{n}} > \text{c}\right] = \text{Prob}\left[\frac{\bar{X}-\mu_0}{\hat{\sigma}/\sqrt{n}} < -\text{c}\right] = \frac{0.05}{2} = 0.025$$

- Podemos então achar o valor de *c* usando os comandos:

```
c <- qnorm(0.0975);  #---- no caso da normal padronizada
c <- qt(0.0975, df=24); #---- no caso da t-Student
```

- Os resultados são: 1.96 e 2.063 respectivamente

Testes de Hipóteses e P-Valores

- Para a amostra considerada no terceiro exemplo, a média amostral é igual a 504.032, e o desvio padrão amostral é igual a 6.047. O valor da estatística teste é dado por:

$$t_{stat} = \frac{\bar{X} - 500}{\hat{\sigma}/\sqrt{25}} = \frac{4.032}{6.047/\sqrt{25}} = 3.33414$$

- Usando a função t.test no R, temos o comando:

```
amostra4 <- c(506.7, 505.5, 504.2, 505.9, 495.8, 508.4, 502.6, 510.1, 491.5, 498.0, 507.0, 515.3,
502.4, 509.2, 519.2, 508.4, 496.7, 503.3, 502.4, 499.8, 503.8, 503.9, 501.5, 497.3, 501.9)
t.test(amostra4, mu = 500, alternative = "two.sided")
```

- O resultado (rejeitamos a hipótese nula nesse caso):

```
        One Sample t-test

data:  amostra4
t = 3.3341, df = 24, p-value = 0.002771
alternative hypothesis: true mean is not equal to 500
95 percent confidence interval:
 501.5361 506.5279
sample estimates:
mean of x
  504.032
```