ANÁLISE DE DADOS MULTIVARIADOS I - REGRESSÃO

(AULA 02)

Novembro e dezembro de 2018

Reinaldo Soares de Camargo

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, independente, 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

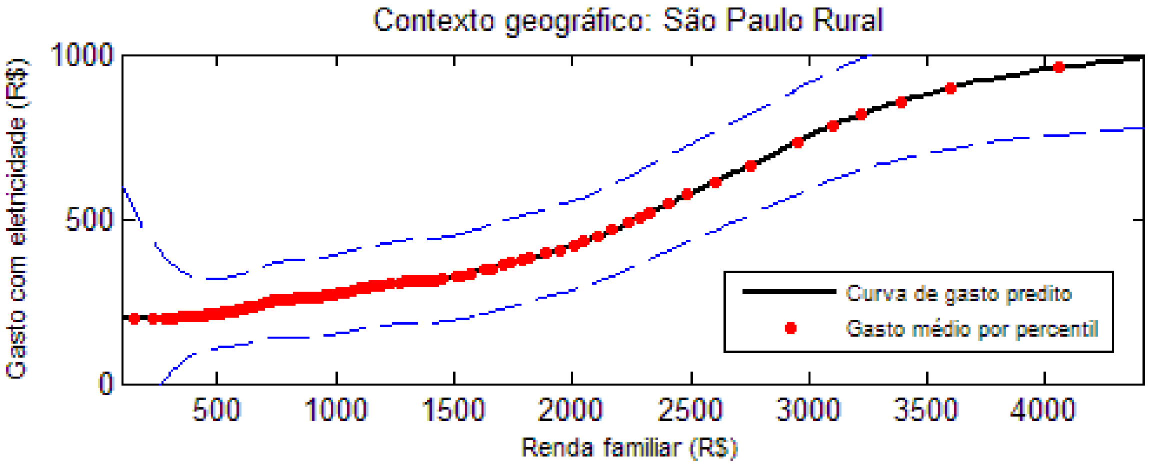
- Quanto à forma funcional para g(.),

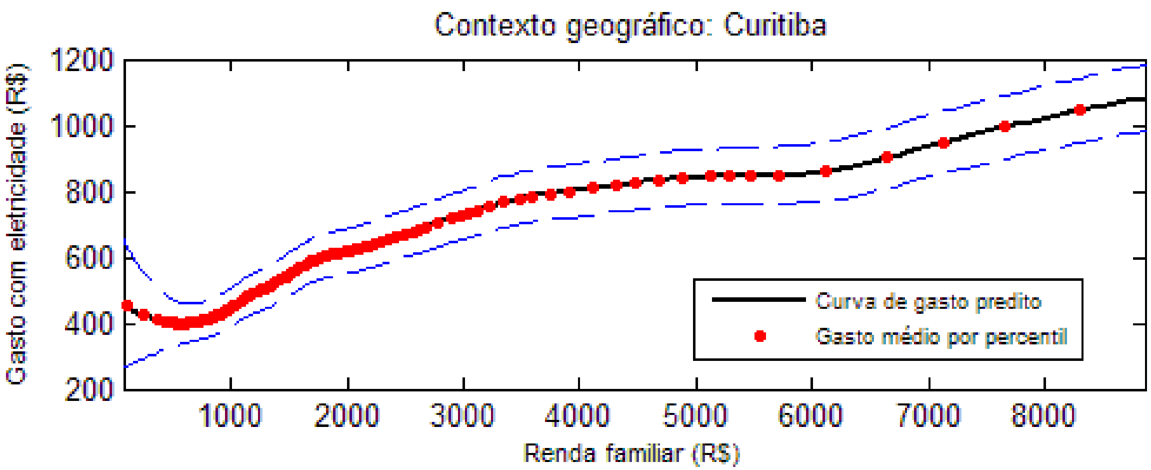
  - Quando a função g(.) é pré-especificada, chamamos de **regressão paramétrica**
  - Quando a função g(.) é desconhecida e é estimada pelos dados, chamamos de regressão **não-paramétrica** ou **semi-paramétrica**

- 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

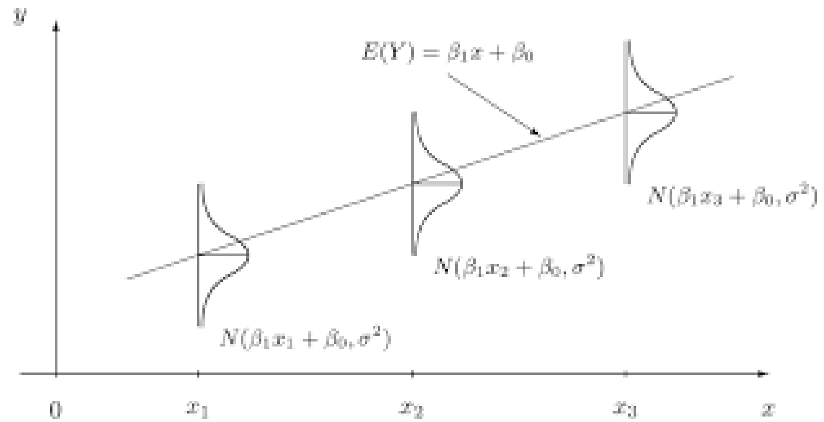
- Modelos de regressão linear simples
  - O tipo mais comum de modelo de regressão é modelo de regressão linear
  - Forma funcional é simplesmente uma expressão linear das variáveis explicativas
  - O termo ϵ_i corresponde à parte do que observamos para a variável resposta, que não é explicada pelas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

$$y_i = \beta_0 + \beta_1 x_{1i} + \epsilon_i$$

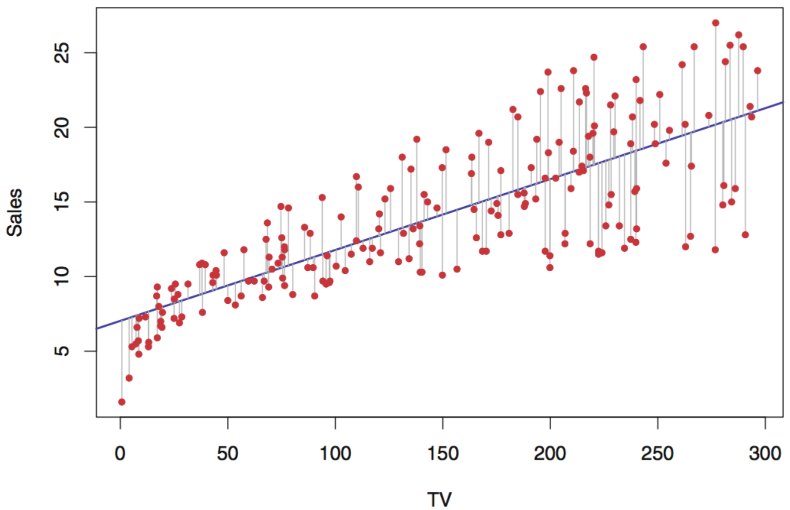
- Diversas hipóteses são descritas na literatura para o modelo de regressão linear na sua forma mais básica (pressupostos do modelo)
  - **Os termos ϵ_i têm distribuição normal**
  - Os termos ϵ_i são não correlacionados entre eles
  - Os termos ϵ_i possuem variância constante (erros homoscedásticos)
  - Os termos ϵ_i são não correlacionados com a variável explicativa x_{1i}
- Na prática, essas hipóteses básicas não se confirmam, 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

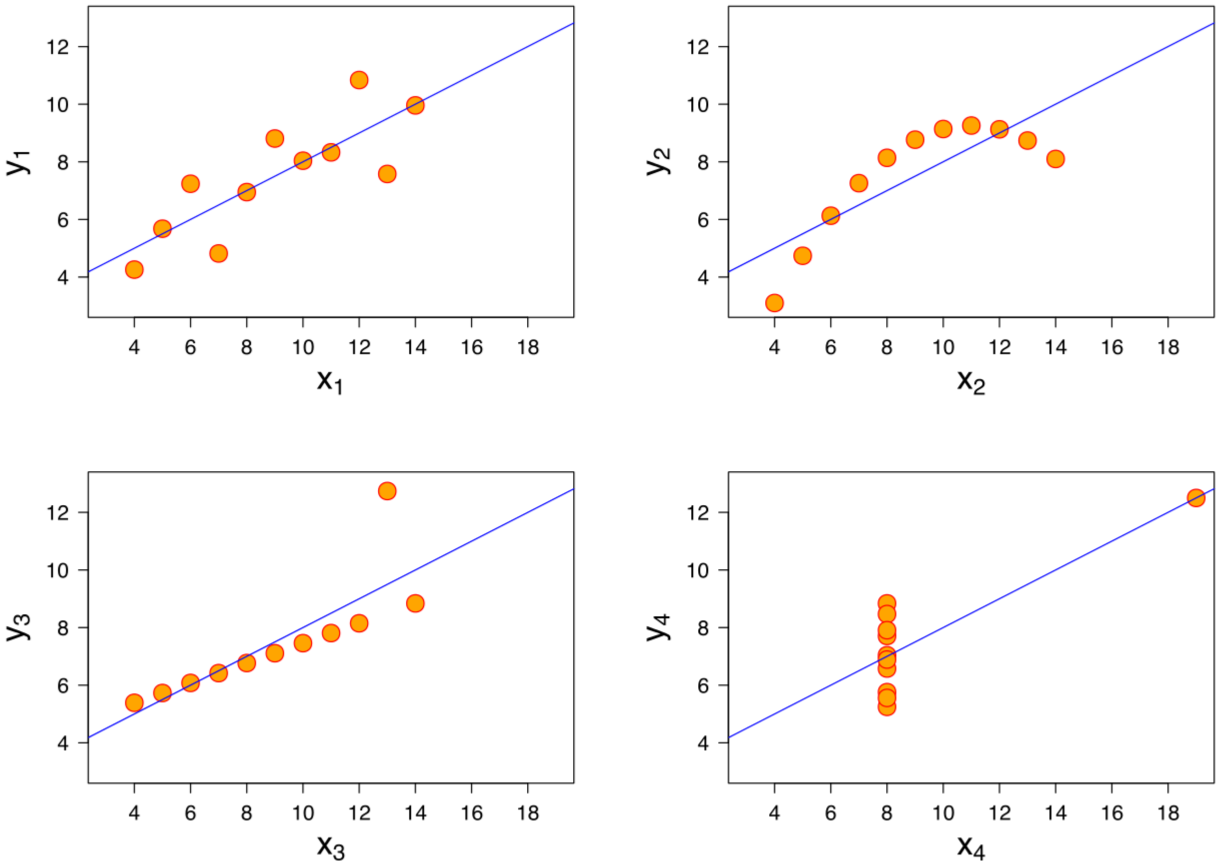


- Hipótese de homoscedasticidade é violada



Modelos de Regressão

- Regressão linear (simples) com uma variável explicativa



Modelos de Regressão

- **Estimação dos coeficientes desconhecidos β_0 , β_1 para cada variável no modelo de regressão linear simples**

  - Considere um conjunto especifico de valores para β_0 , β_1
  - Com base nesses valores, podemos calcular o valor previsto para a variável resposta.

$$\hat{y}_i = \beta_0 + \beta_1 x_{1i}$$

  - O erro de previsão é dado por

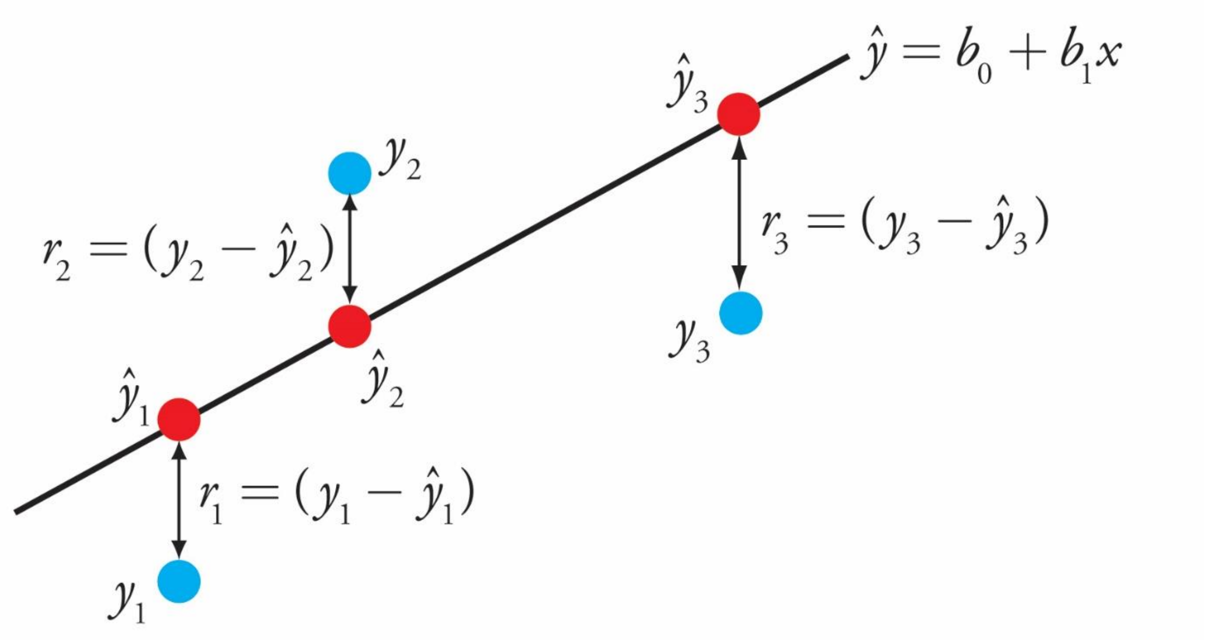
$$r_i = y_i - \hat{y}_i$$

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

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

  - Podemos então escolher um conjunto de valores dos coeficientes β_0 , β_1 de forma a minimizar a soma dos erros quadráticos SQE.

Modelos de Regressão



- O método de estimação dos coeficientes β_0 , β_1 pela minimização da soma dos erros ao quadrado é conhecido como **método de mínimos quadrados ordinários (MQO) ou OLS** do inglês **(Ordinary Least Squares).**

Modelos de Regressão

- **Estimação dos coeficientes desconhecidos β_0 , β_1 via método de mínimos quadrados ordinário possui forma fechada a partir da amostra observada.**

  - No caso de uma variável explicativa (regressão linear simples),

$$\hat{y}_i = \beta_0 + \beta_1 x_1 + \varepsilon_i$$

  - podemos estimar β1 a expressão

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

  - O coeficientes **β_0 é estimado por**

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

  - O nome “mínimos quadrados ordinários” é utilizado porque as estimativas do intercepto ($\hat{\beta}_0$) e da inclinação ($\hat{\beta}_1$) minimizam a soma dos resíduos quadrados..

Modelos de Regressão

- **Aplicação em R**

- Obtenha os dados de peso, altura, gênero da turma
- Esboçar um gráfico das variáveis peso e altura
- Obtenha as estatísticas descritivas das variáveis peso e altura.
- Estime um modelo de regressão linear simples para explicar as variações de peso (y) em função da altura (x) dos participantes do curso.
- Esboce a reta de regressão com os dados estimados
- Interprete os resultados.

Modelos de Regressão

- **Interpretação dos coeficientes (β_1)**

- **Quando as variáveis y e x estão em níveis**

$$y = \beta_0 + \beta_1 x + u$$

- **O aumento de uma unidade em x aumenta y em β_1 Vezes.**

Modelos de Regressão

- **Exemplo de interpretação dos coeficientes (β_1)**

- Modelo que explique os gastos do consumidor (y), medido em bilhões de R$, em função da renda disponível (x), também em bilhões de R$, obteve as seguintes estimativas.

$$\hat{y} = -27{,}53 + 0{,}93x$$

- O aumento de um R$ 1 bilhão de dólares (uma unidade) na renda disponível eleva os gastos dos consumidores em R$ R$ 930 milhões.

Modelos de Regressão

- **Interpretação dos coeficientes (β_1)**

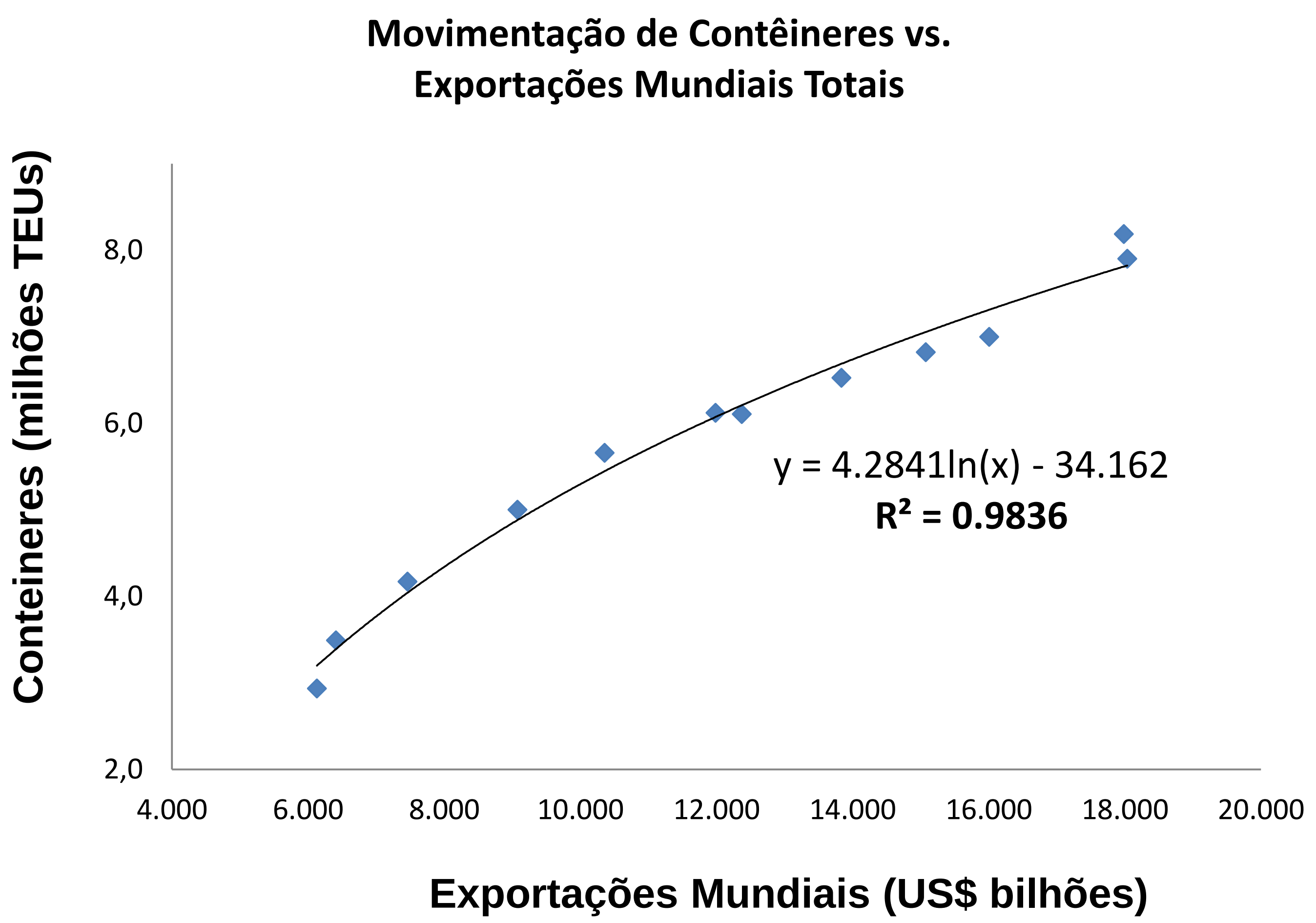
- **Quando a variável y estiver em nível e a x estiver em log**

$$y = \beta_0 + \beta_1 \ln(x) + u$$

- **Ln(x) = logaritmo natural: $\log_e$ x , em que *e* ~= 2,71.**

- **O aumento de uma unidade em x aumenta y em ($\beta_1/100$)%.**

Modelos de Regressão



Modelos de Regressão

- **Interpretação dos coeficientes (β_1)**

- **Quando as variáveis y estiver em log e a variável x em nível**

$$\log(y) = \beta_0 + \beta_1 x + u$$

- **O aumento de uma unidade em x aumenta y em** $100x\beta_1\%$.

Modelos de Regressão

- **Exemplo de interpretação dos coeficientes (β_1)**

- Para estudar a demanda de habitação numa dada localidade, foi especificado o seguinte modelo de regressão.

$$\log(y) = \beta_0 + \beta_1 x + u$$

- Em que y = quantidade de habitações demandadas ao ano; x = Preço da unidade de habitação na localidade.

- Os resultados no processo de estimação foram:

$$\widehat{logY} = 4{,}17 - 0{,}247x$$

- O aumento de uma unidade monetária no preço da unidade, reduz a demanda por habitações em 24,7%.

Modelos de Regressão

- **Interpretação dos coeficientes (β_1)**

- **Quando as variáveis y e x estiverem em log**

$$\log(y) = \beta_0 + \beta_1 \log(x) + u$$

- **O aumento de 1% em x aumenta y em β_1%.**

Este é o modelo de elasticidade constante. Elasticidade é a razão entre o percentual de mudança em uma variável e o percentual de mudança em outra variável.

Modelos de Regressão

- **Exemplo de interpretação dos coeficientes** (β_1)

- Para estudar a demanda por modelo de longo e de curto prazo, foi especificado seguinte modelo.

$$\ln(M) = \beta_0 + \beta_1 \ln(R) + u$$

- Em que M = estoque total de moeda, R = taxa de juros

- Os resultados no processo de estimação foram:

$$\widehat{lnM} = 0{,}1365 - 0{,}7476\ln(R)$$

- O aumento de 1% na taxa de juros, reduz a demanda por moeda em 0,7476%.

Modelos de Regressão

- **Uso de log em econometria.**

- O uso de logaritmos de variáveis dependentes ou independentes pode permitir relações não-lineares entre a variável explicada e as variáveis explicativas (Wooldrigde, 2006, p. 179).

- O uso de logs pode aliviar ou até eliminar problemas de heterocedasticidade (quando a variância dos erros não é constante, ou seja, não há homoscedasticidade) ou concentração em distribuições condicionais advindas de variáveis estritamente positivas. As estimativas com o uso de logs são menos sensíveis a observações desiguais (ou extremas) devido ao estreitamento considerável que pode ocorrer na amplitude dos valores das variáveis (Wooldrigde, 2006, p. 181).

Modelos de Regressão

- Algumas regras práticas para o uso de logs, conforme Wooldridge (2006, p. 181):

- **Geralmente usam log:**
  - valores monetários positivos frequentemente são transformados em log (salários, vendas de empresas, valor de mercado de empresas)
  - grandes valores inteiros também costumam ser usados em forma logarítmica: população, número total de funcionários, matrículas escolares.

- **Geralmente não usam log:**
  - variáveis medidas em anos geralmente não levam a forma logarítmica: educação, experiência, tempo de permanência, idade.

- **Podem usar ou não o log:**
  - variáveis que são proporções ou percentagens podem usar ou não o log, mas a tendência é utilizá-las em sua forma original para possibilitar uma interpretação em termos de pontos percentuais: taxa de desemprego, taxa de participação em planos de aposentadoria, taxa de aprovação em exames de escolaridade padronizados, taxa de detenção por crimes registrados.

Modelos de Regressão

- **1ª Lista de exercícios para entregar em 12/11/2018.**
  - Os exercícios podem ser entregues em grupos de 2 alunos, e o grupo deve submeter o código em R utilizado para responder ao exercício, juntamente com a discussão dos resultados.
  - Utilize a base de dados do IDH brasil 2010 (IDH_Brasil_2010.csv)
  - 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)

summary(mod1.ex)
```

Modelos de Regressão

- **1ª Lista de exercícios para entregar em 12/11/2018.**
  - 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

- Coeficiente de determinação (R^2) da regressão

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

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

  - 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

- 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

Obrigado.