Ciência dos Dados

Aula 29 – Projeto 3

Modelo de regressão linear

Projeto 3

O Projeto 3 é composto por três etapas:

1ª. Etapa: Escolha das variáveis

2ª. Etapa: Desenvolvimento teórico dos coeficientes linear e angular de um modelo de regressão simples e generalização para um modelo de regressão múltipla.

3ª. Etapa: Analise descritiva e análise de regressão nos dados definidos na Etapa 1 e sob o modelo teórico estudado na Etapa 2. E ainda avaliação se o modelo de regressão obtido é igualmente bom quando os países são separados em subgrupos (com critérios consistentes a definir).

Insper

Projeto 3

Cada grupo deverá ter uma das variáveis resposta a seguir:

- Fertilidade (Children per women)
- Expectativa de Vida (Life expectancy)
- Mortalidade infantil (Child mortality)
- Índice de percepção de corrupção (Corruption Perception Index - CPI)
- Taxa de emprego (Employment rate)
- Taxa de desemprego (Unemployment rate)
- Score de democracia (Democracy score)

Os slides a seguir descrevem as características e cuidados com uma Análise de Regressão

Pesquise alguma referência bibliográfica para mais detalhes!!

Objetivo de uma Análise de Regressão

Estudar relação entre variáveis quantitativas.

Para o Projeto 3, essas devem ser extraídas do GapMinder.

Exemplos:

Expectativa de vida e Gasto com Saúde

Expectativa de vida e % da população com acesso ao saneamento

Taxa de criminalidade e Taxa de desemprego

Índice de percepção de corrupção e IDH

CO_2 e PIB

Insper

Objetivo de uma Análise de Regressão

A presença ou ausência de **relação linear** pode ser investigada sob dois pontos de vista:

a) Quantificando a força dessa relação: correlação.

b) Explicitando a forma dessa relação: regressão.

Graficamente, a relação entre duas variáveis quantitativas pode ser feita via **Gráfico de Dispersão.**

Insper

Objetivo – Um particular problema

Para o Projeto 3, é necessário que o grupo trace um problema/pergunta que deseja avaliar!!

Exemplo:

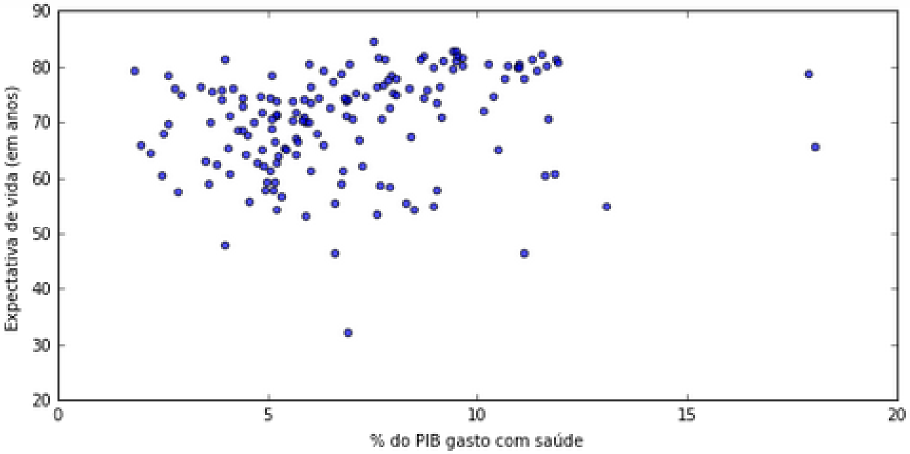
Investimentos na saúde e saneamento básico podem aumentar sobrevida de uma população de um país?

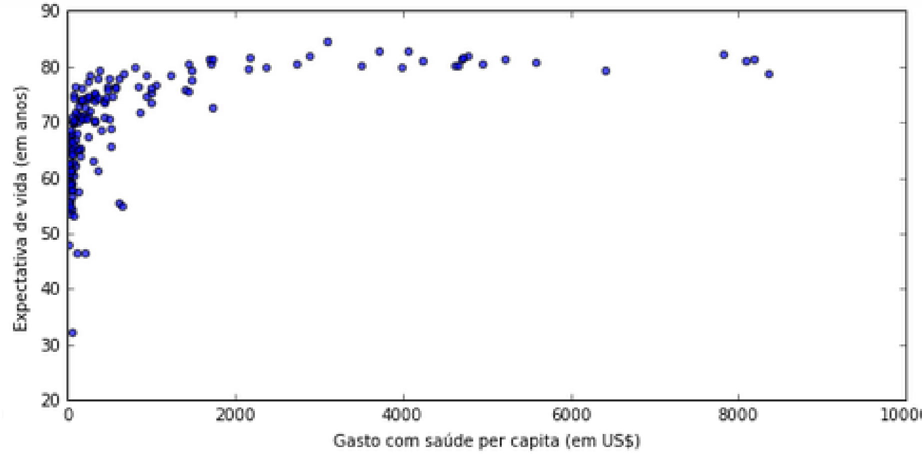
Variáveis selecionadas que podem auxiliar na análise:

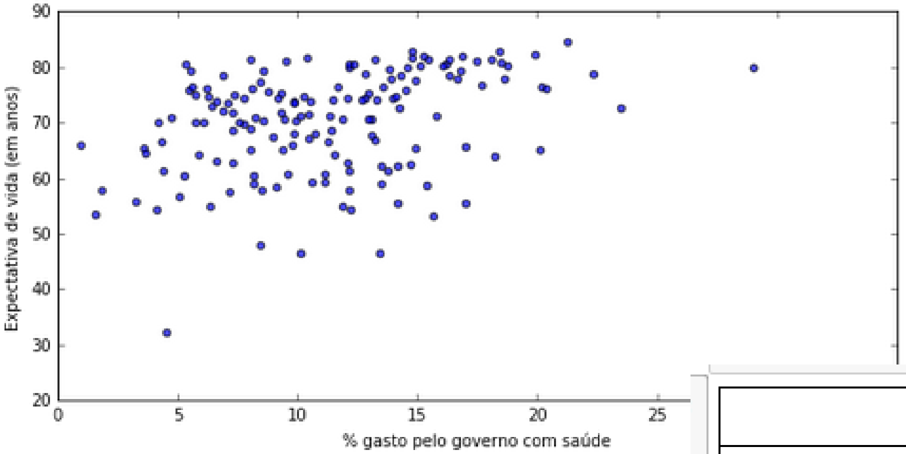
Expectativa de vida
Gasto com Saúde per capita (em US$)
% do PIB investido na saúde
% gasto pelo governo com a saúde
% da população com acesso ao saneamento

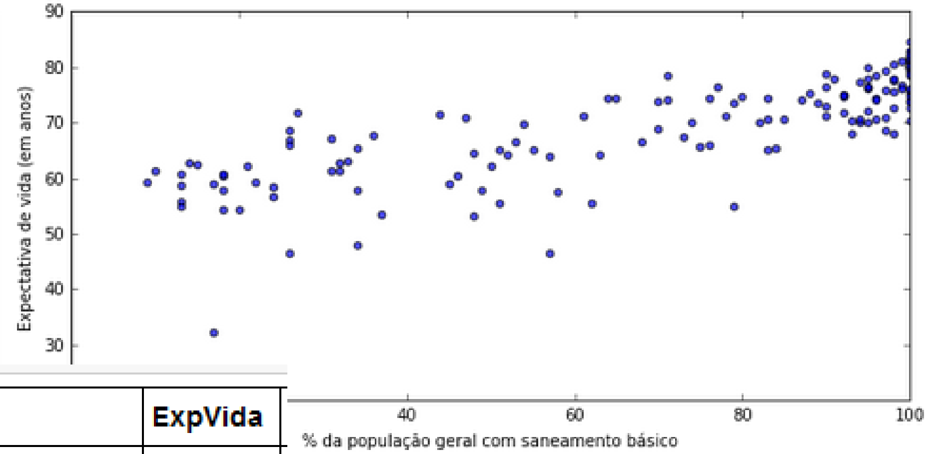
Insper

Análise Descritiva





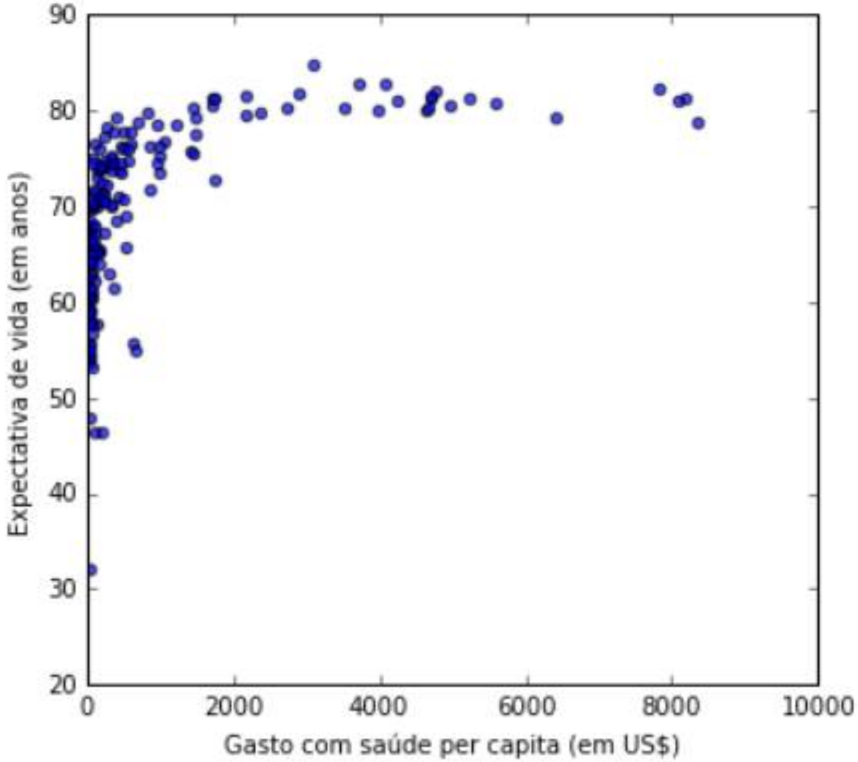


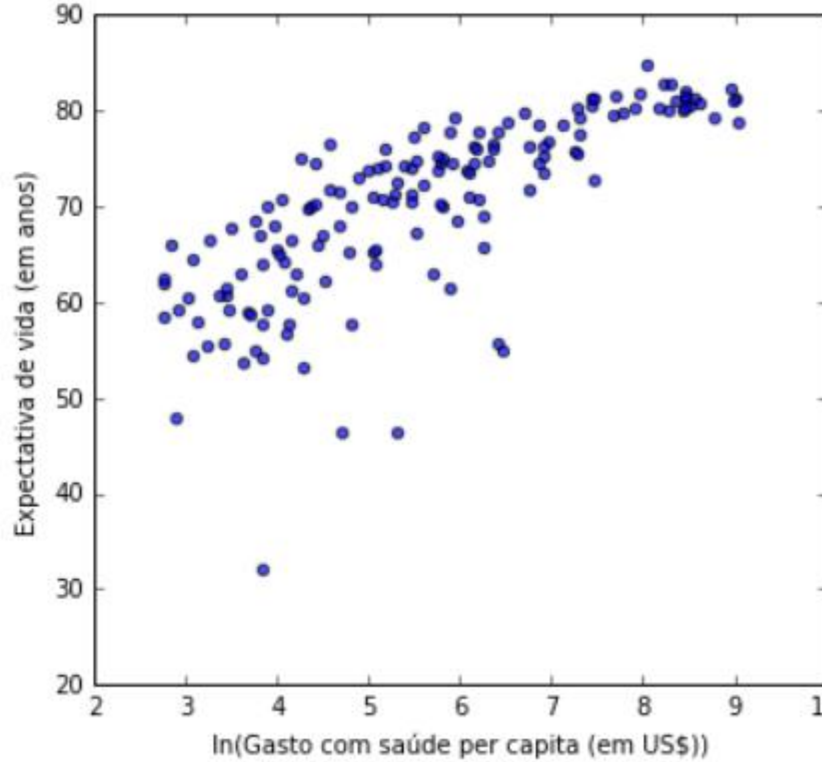


	ExpVida
ExpVida	1.000000
PercSaudePIB	0.236452
GastoSaudePerCap	0.553312
PercSaudeGov	0.361530
PropPopSanea	0.802367

Insper

Transformação na variável





	ExpVida
ExpVida	1.000000
PercSaudePIB	0.236452
GastoSaudePerCap	0.553312
PercSaudeGov	0.361530
PropPopSanea	0.802367
LNGastoSaudePerCap	0.763843

Análise de regressão

"A coleção de ferramentas estatísticas que são usadas para modelar e explorar relações entre variáveis que estão relacionadas de maneira não determinística é chamada de análise de regressão."

Montgomery, D.C. e Runger, G.C. **Estatística aplicada e probabilidade para engenheiros.** 6ª. Edição. Rio de Janeiro: LTC, 2016.

Análise de regressão

Objetivo: Explicar como uma ou mais variáveis se comportam em função de outra.

Variável dependente (resposta) - y: variável de interesse, cujo comportamento se deseja explicar.

Variável independente (explicativa) - x: variável ou variáveis que são utilizadas para explicar a variável dependente.

Modelo de regressão: equação (reta) que associa **y** e um ou vários **x**.

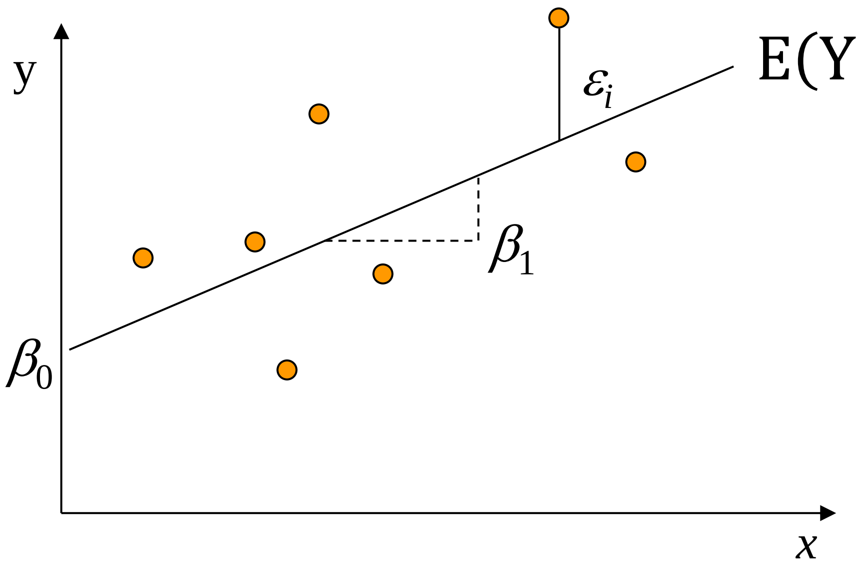
Análise de regressão

Metodologia estatística que estuda (modela) a relação entre duas ou mais variáveis

1. Expectativa de vida $\Rightarrow$ variável resposta
   Gasto com saúde (per capita) $\Rightarrow$ variável explicativa
   $\Downarrow$
   modelo de regressão linear simples

2. Expectativa de vida $\Rightarrow$ variável resposta
   Gasto com saúde (per capita) $\Rightarrow$ variável explicativa
   % população com saneamento $\Rightarrow$ variável explicativa
   $\Downarrow$
   modelo de regressão linear múltipla

Modelo de Regressão Linear Simples



$$E(Y|x) = \hat{y} = \beta_0 + \beta_1 x$$

Inclinação populacional

Intercepto populacional

Erro Aleatório

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

Variável Dependente

Variável Independente

Método dos Mínimos Quadrados

Os valores populacionais de β_0 e β_1 são desconhecidos.

O método utilizado na estimação desses parâmetros é o método dos mínimos quadrados, o qual considera os erros dos Y_i de seu valor esperado:

$$\varepsilon_i = Y_i - (\beta_0 + \beta_1 x_i)$$

Em particular, o método dos mínimos quadrados requer que consideremos a soma dos n erros quadrados, denotado por SQ:

$$SQ = \sum_{i=1}^{n} \left(Y_i - \hat{Y}_i\right)^2 = \sum_{i=1}^{n} \left(Y_i - \beta_0 - \beta_1 x_i\right)^2$$

Insper

Inferência em Análise de Regressão

Usualmente, uma das hipóteses em análise de regressão é avaliar a significância da regressão.

Ou seja,

H_0: $\beta_1 = 0$ ➔ não há relação entre x e Y

H_1: $\beta_1 \neq 0$ ➔ há relação entre x e Y

Para realizar esse teste de hipóteses, será necessário atribuir distribuição aos erros ε_i, além de outras suposições ao modelo.

Suposições do modelo linear simples

- Os erros têm distribuição normal com média e variância constante, ou seja,

$$\varepsilon_i \sim N(0,\sigma^2).$$

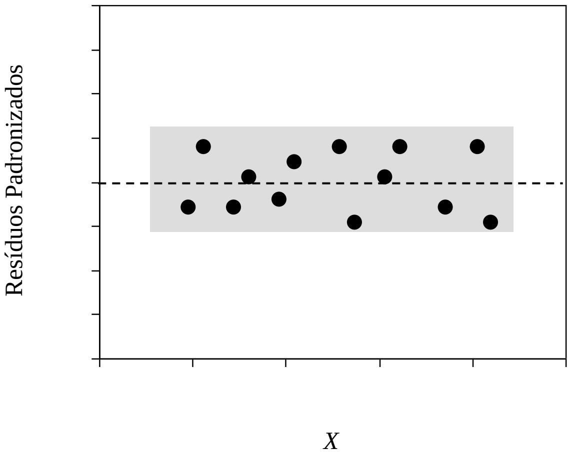
- Os erros são independentes entre si, ou seja,

$$Corr(\varepsilon_i , \varepsilon_j)=0$$

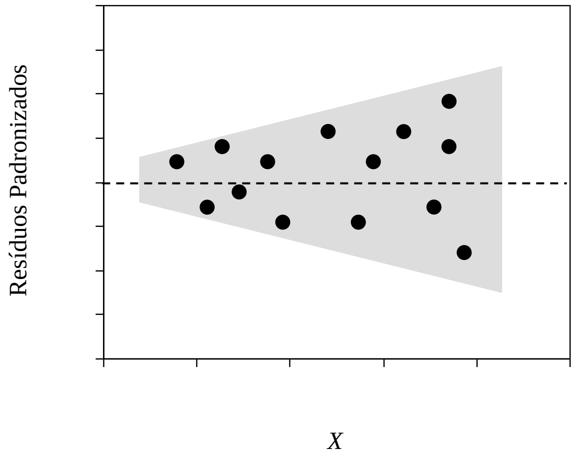
- Modelo é linear nos parâmetros.
- Homocedasticidade: $Var(\varepsilon_i) = \sigma^2$

Análise de Resíduos

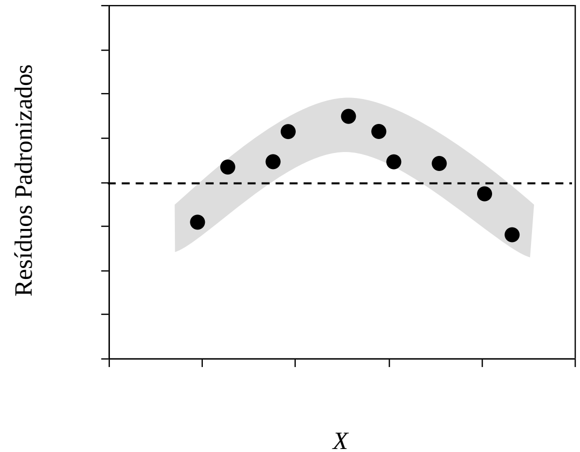
“ideal”



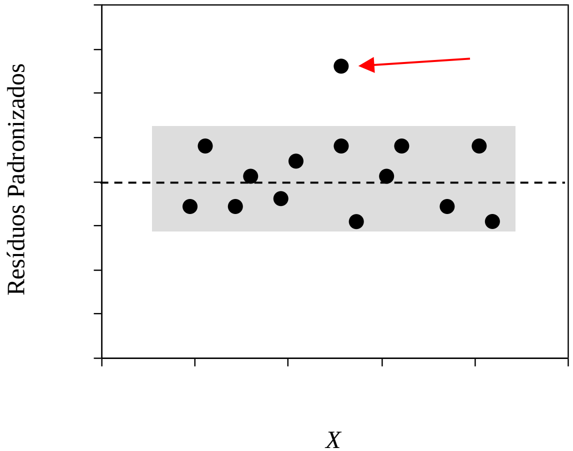
σ^2 não constante



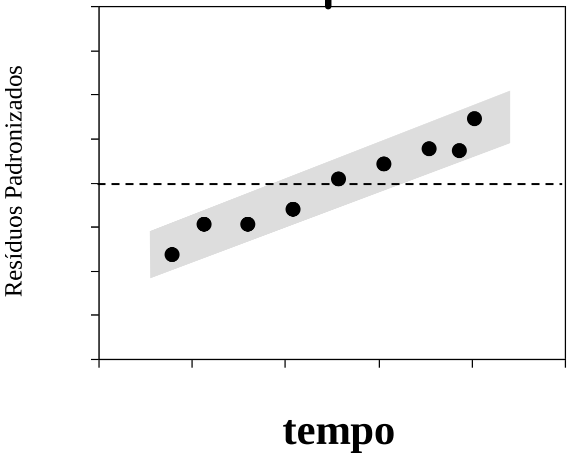
não linearidade



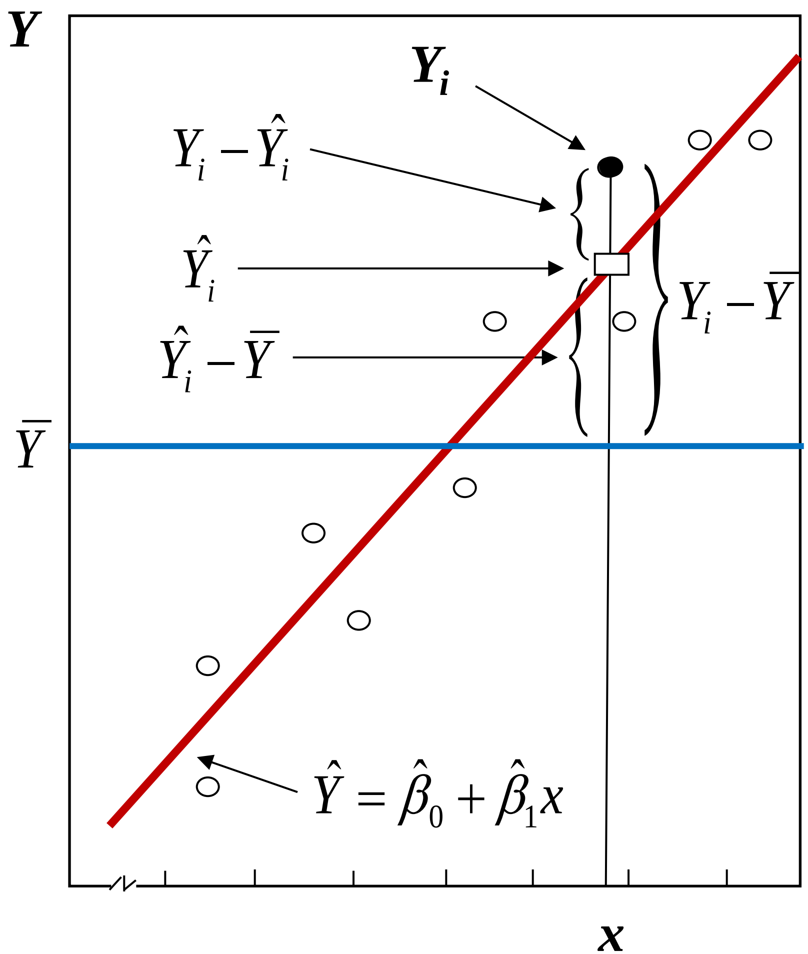
“outlier”



não independência



Qualidade do ajuste



$$\sum_{i=1}^{n}(Y_i - \bar{Y})^2 = \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2 + \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$$

$$\text{SQT} = \text{SQReg} + \text{SQRes}$$

$$R^2 = \frac{\text{SQReg}}{\text{SQT}} = \frac{\text{SQT - SQRes}}{\text{SQT}} = 1 - \frac{\text{SQRes}}{\text{SQT}}$$

Coeficiente de determinação

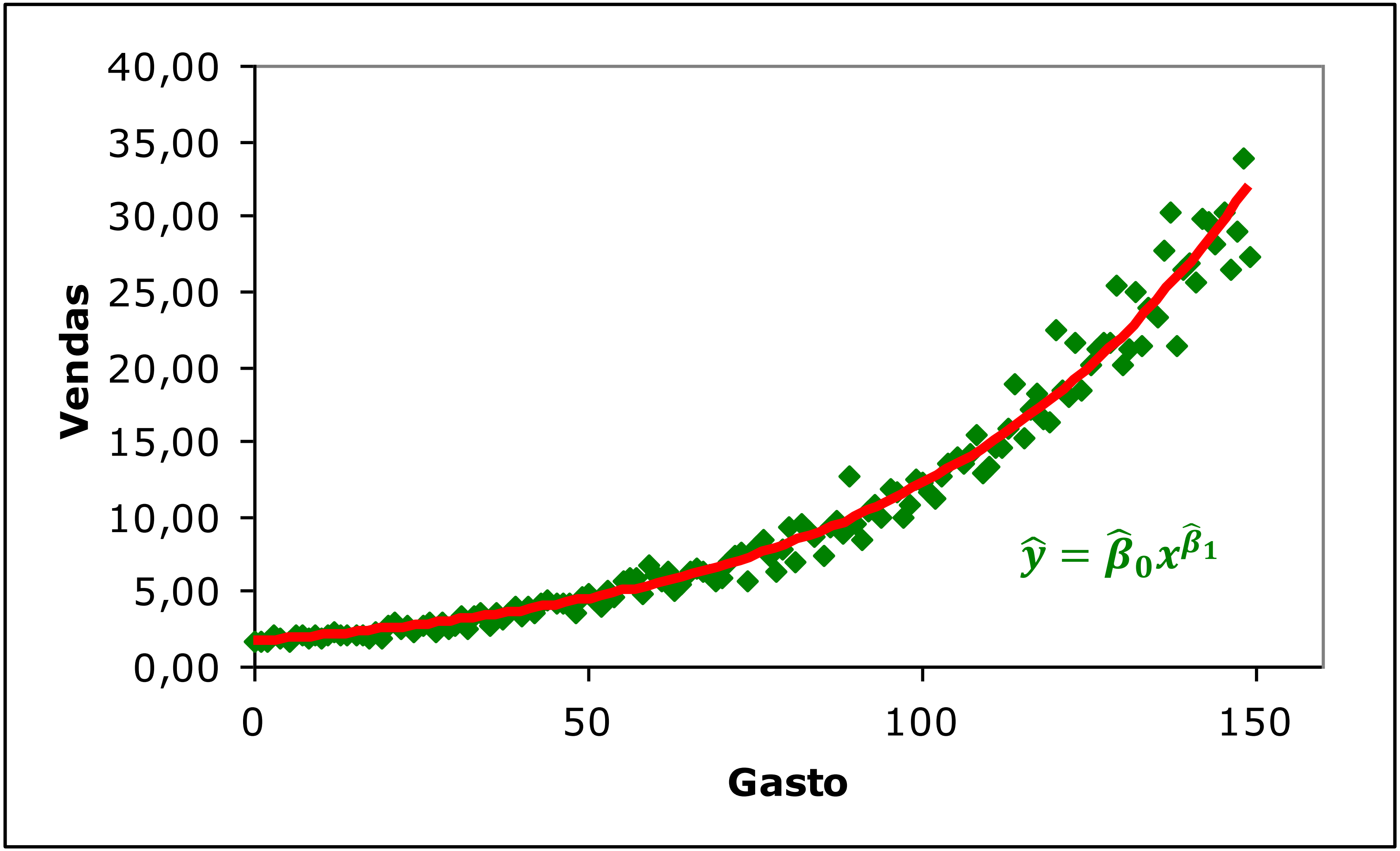
$$0 \leq R^2 \leq 1$$

Interpretação do Coeficiente de determinação: mede a fração da variação total de Y explicada pela regressão.

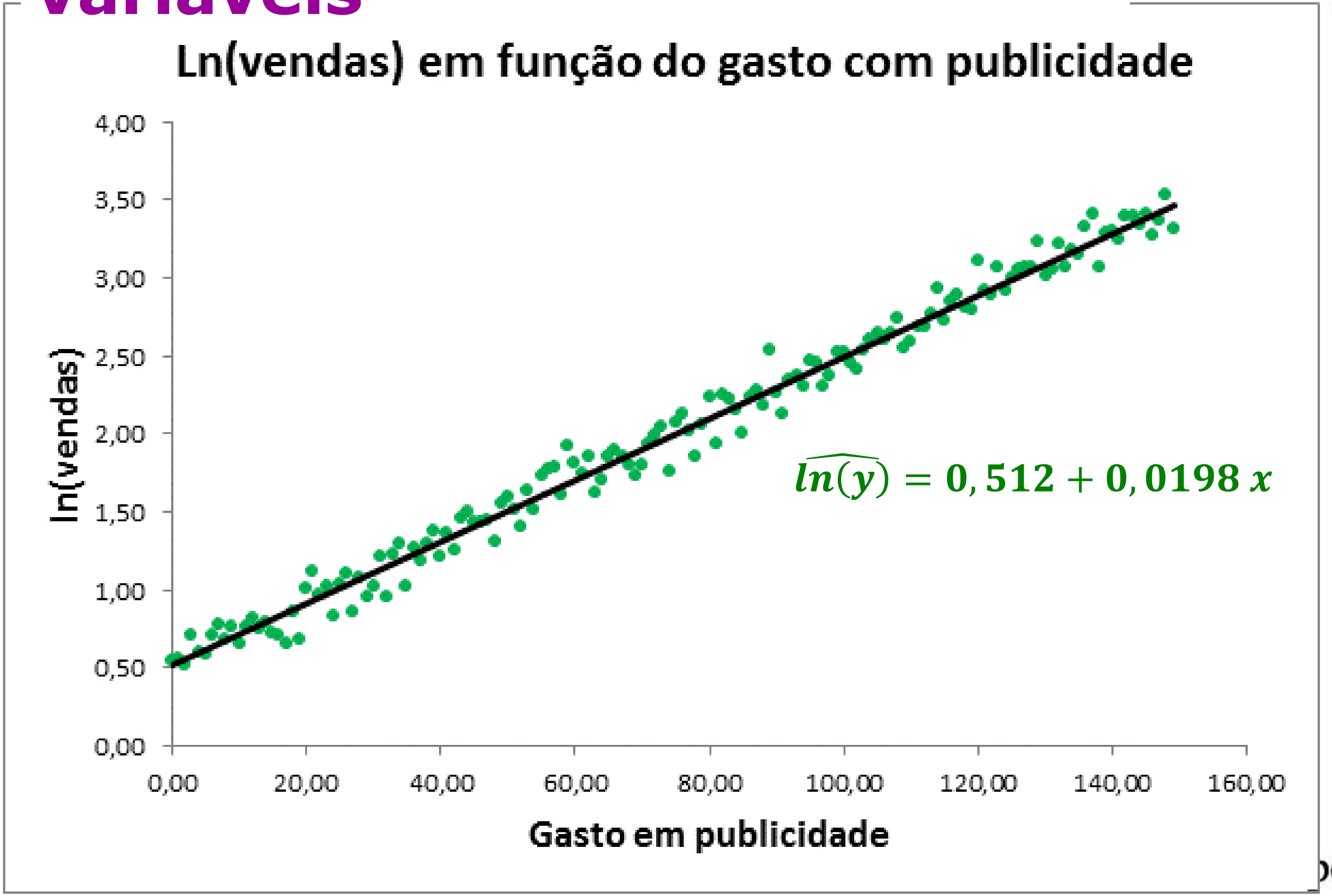
Insper | Instituto de Ensino e Pesquisa

ATENÇÃO

OUTRAS transformações nas variáveis



OUTRAS transformações nas variáveis



Modelos Linearizáveis

Modelo Padrão: $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$

exponencial

$$Y_i = \beta_0 e^{\beta_1 X_i} \varepsilon_i \;\Rightarrow\; \ln Y_i = \ln \beta_0 + \beta_1 X_i + \ln \varepsilon_i \;\Rightarrow\; Y_i' = \beta_0' + \beta_1 X_i + \varepsilon_i'$$

potencial

$$Y_i = \beta_0 X_i^{\beta_i} \varepsilon_i \;\Rightarrow\; \ln Y_i = \ln \beta_0 + \beta_1 \ln X_i + \ln \varepsilon_i \;\Rightarrow\; Y_i' = \beta_0' + \beta_1 X_i' + \varepsilon_i'$$

$$\boxed{Y'_i = \beta'_0 + \beta_1 x'_i + \varepsilon'_i} \quad \left\{ \begin{array}{l} \textbf{exponencial} \\ \textbf{potencial} \end{array} \right.$$

Cuidado com a interpretação dos parâmetros caso faça transformação na(s) variável(is).

Insper

Associação não é causalidade

Suponha que encontremos alta correlação entre duas variáveis A e B. Podem existir diversas explicações do porque elas variam conjuntamente, incluindo:

- Mudanças em outras variáveis causam mudanças tanto em A quanto em B.
- Mudanças em A causam mudanças em B.
- Mudanças em B causam mudanças em A.
- A relação observada é somente uma coincidência (**correlação espúria**). **CUIDADO!!**

Fonte: http://leg.ufpr.br/~silvia/CE003/node77.html