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

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

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)

Insper

Insper | Instituto de Ensino e Pesquisa

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

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

Um particular problema

Investimentos na saúde e saneamento básico têm alguma relação com a sobrevida de uma população?

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 têm alguma relação com sobrevida de uma população?

Variáveis selecionadas que podem auxiliar na análise:

Expectativa de vida

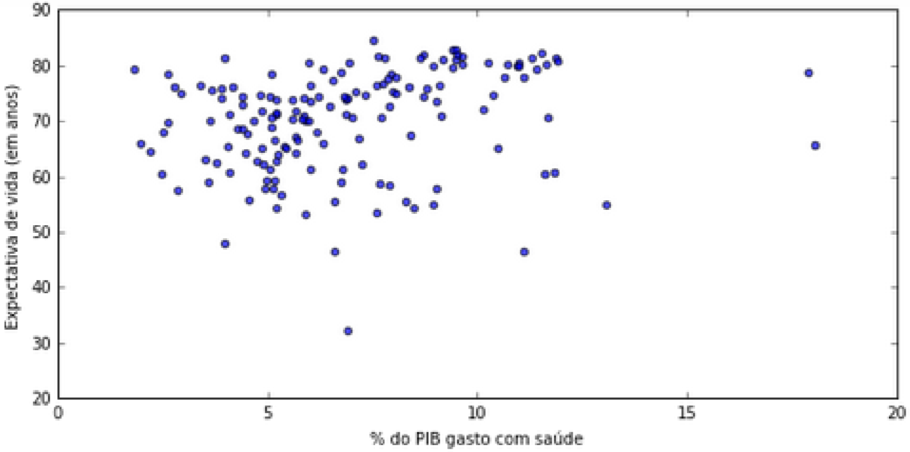
Gasto com Saúde per capita (em US$)

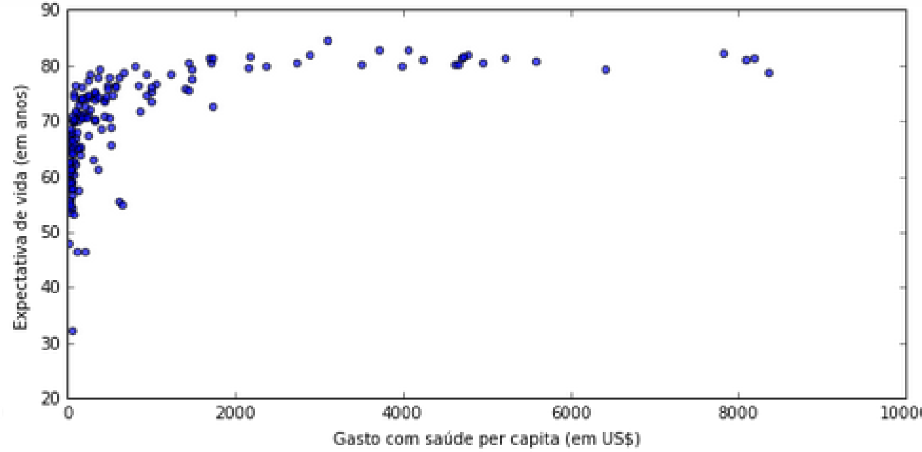
% do PIB investido na saúde

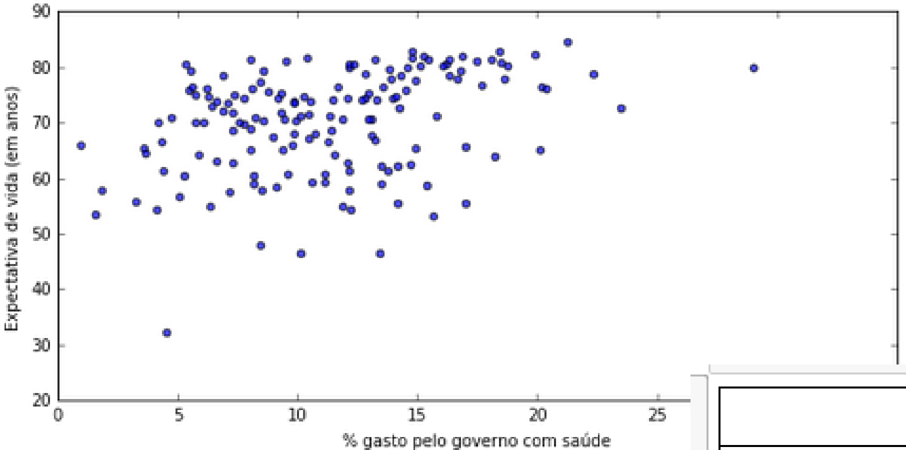
% gasto pelo governo com a saúde

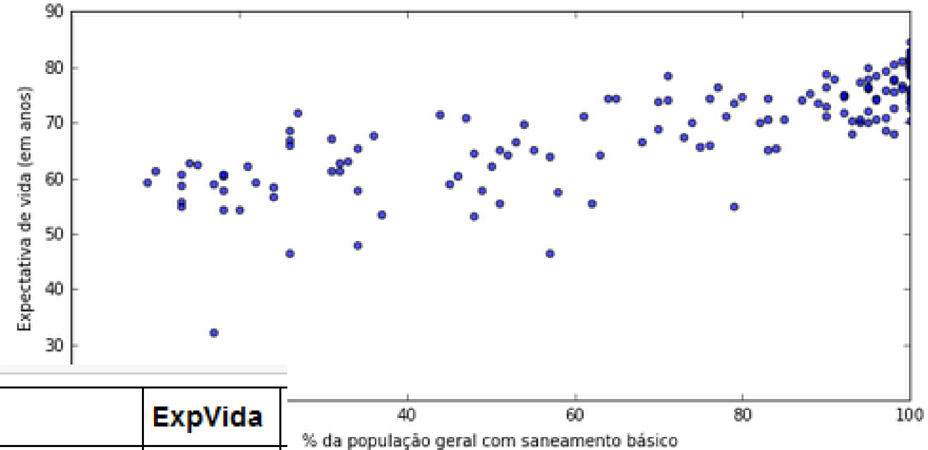
% da população com acesso ao saneamento

Análise Descritiva





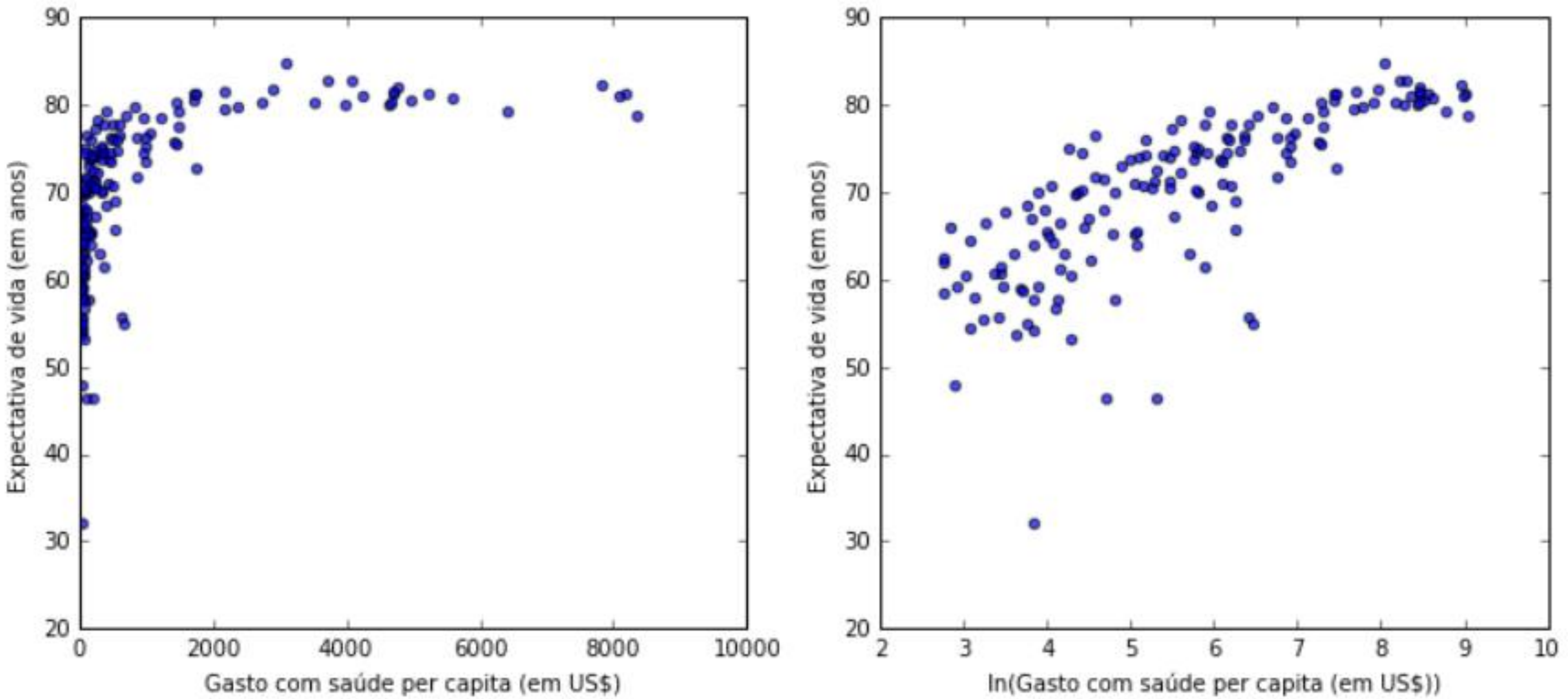




	ExpVida
ExpVida	1.000000
PercSaudePIB	0.236452
GastoSaudePerCap	0.553312
PercSaudeGov	0.361530
PropPopSanea	0.802367

Insper

Transformação na variável



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

Insper

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.

Insper

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 múltipla a ser estimado

Motivados pela análise descritiva (gráfico de dispersão e coeficiente de correlação), a escolha das variáveis para o modelo foram:

$ExpVida$: Expectativa de vida como variável resposta.

$GtSaude$: Gasto com saúde (per capita) considerando a transformação ln decorrente gráfico descrito no slide 9.

$Sanea$: % população com saneamento sem considerar transformação decorrente interpretação gráfica no slide 10.

Nota: Ambas variáveis explicativas têm forte correlação com a variável resposta.

Modelo de regressão múltipla:

$$ExpVida = \beta_0 + \beta_1 Sanea + \beta_2 \ln(GtSaude) + \varepsilon$$

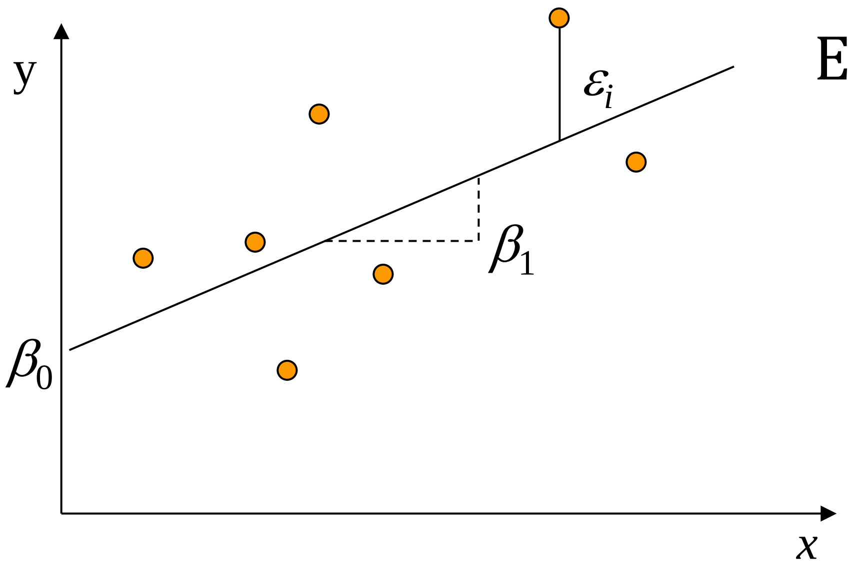
Insper

Insper Instituto de Ensino e Pesquisa

Modelo de regressão simples

Teoria

Modelo de Regressão Linear Simples



$$E(Y|x) = \beta_0 + \beta_1 x$$

Inclinação populacional

Intercepto populacional

Erro Aleatório

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

Variável Dependente

Variável Independente

Insper

Método dos Mínimos Quadrados

Os valores populacionais de β_0 e β_1 são desconhecidos.

Para estimá-los, é necessário minimizar o resíduo que é dado pela diferença entre o valor verdadeiro de y e seu valor estimado $\hat{y}$, ou seja,

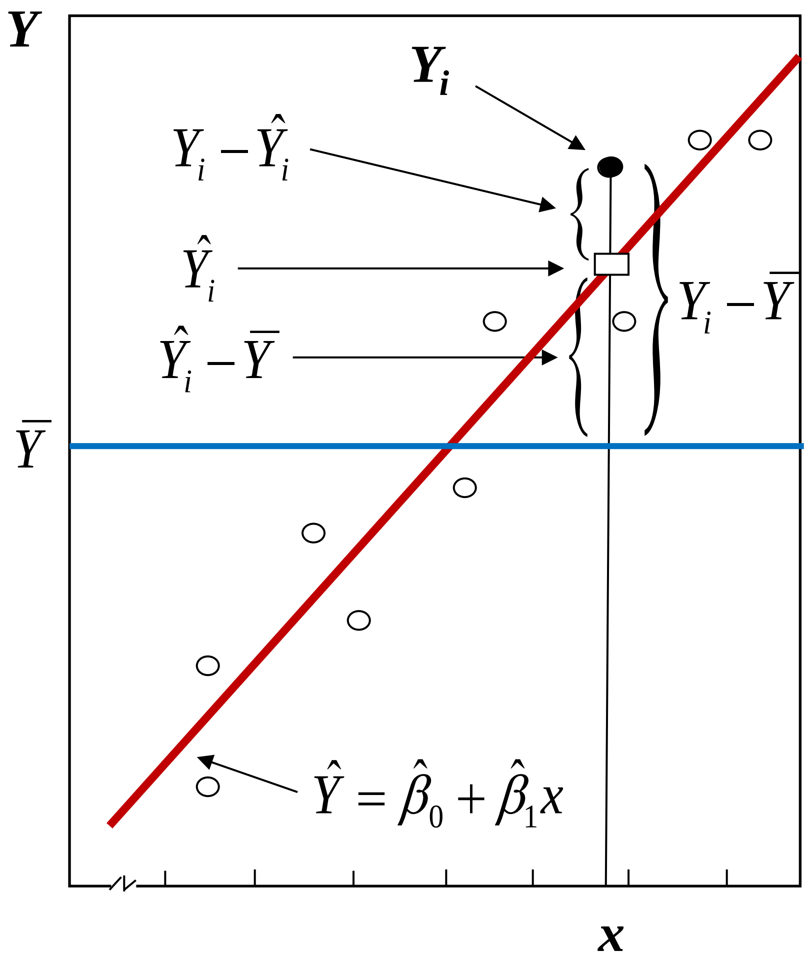
$$\hat{\varepsilon}_i = y_i - \hat{y}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i.$$

O método utilizado na estimação desses parâmetros é o método dos mínimos quadrados.

Logo, o método dos mínimos quadrados requer que consideremos a soma dos n resíduos quadrados, denotado por SQRes:

$$\text{SQRes} = \sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2 = \sum_{i=1}^{n}\left(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i\right)^2$$

Qualidade do ajuste



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

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

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

Coeficiente de determinação

$$0 \le R^2 \le 1$$

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

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,

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

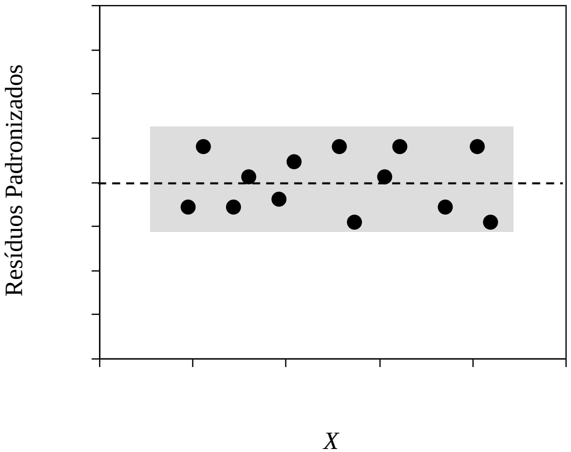
- Os **erros são independentes** entre si, ou seja,

$$\mathrm{Corr}(\varepsilon_i , \varepsilon_j)=0$$

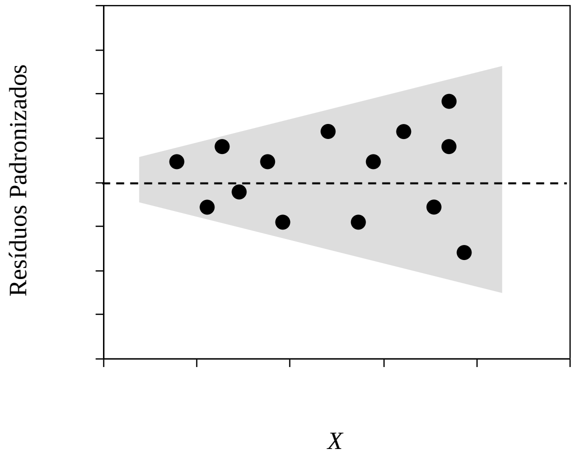
- **Modelo é linear nos parâmetros**.
- **Homocedasticidade**: $\mathrm{Var}(\varepsilon_i) = \sigma^2$ para qualquer $i = 1, \ldots, n$.

Análise de Resíduos

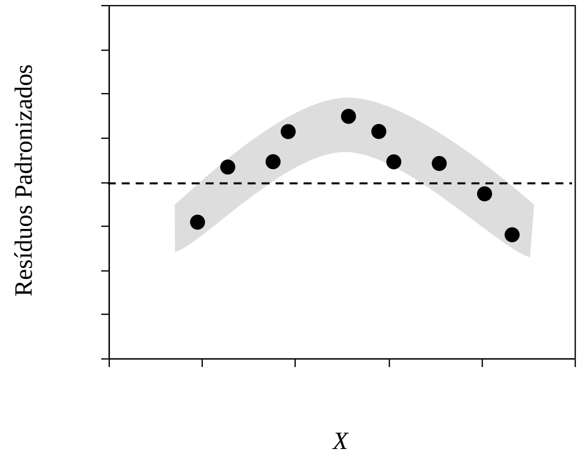
"ideal"



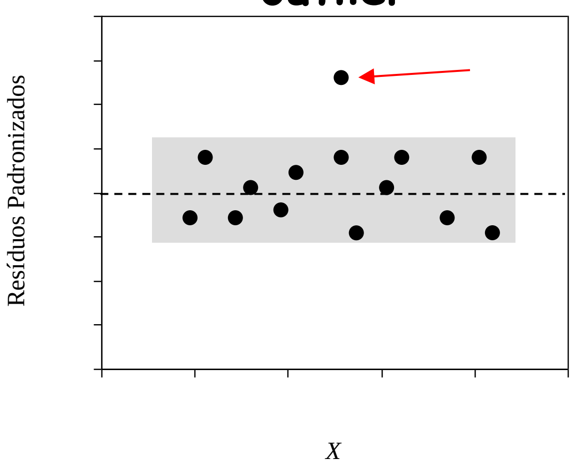
σ^2 não constante



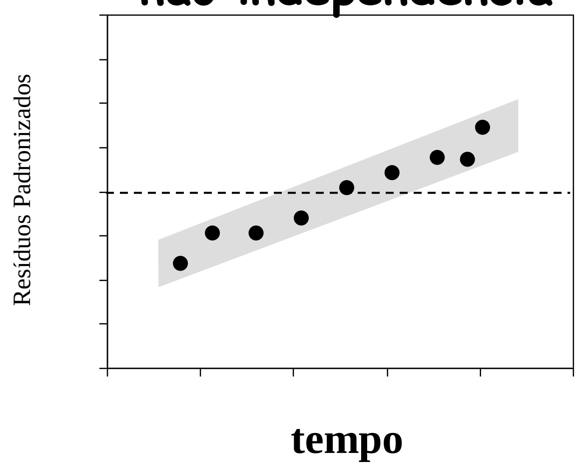
não linearidade



"outlier"



não independência



Interpretação das estimativas dos coeficientes de um modelo de regressão

Modelos lineares nos coeficientes e nas variáveis

Modelo de regressão linear simples – Lin-Lin

Reta estimada:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

Interpretação do coeficiente linear estimado:

O intercepto é o valor previsto (esperado ou médio) para y quando $x = 0$.

Quando não fizer sentido zerar a variável x, o valor $\hat{\beta}_0$, por si só, não será muito interessante.

Interpretação do coeficiente angular estimado:

De maneira geral, a cada variação Δx na variável explicativa x, $\hat{\beta}_1$ é a variação prevista (esperada ou média) na variável resposta.

$$\hat{\beta}_1 = \frac{\Delta \hat{y}}{\Delta x}$$

Insper

Modelo de regressão linear simples – Lin-Lin

Reta estimada:

$$\widehat{Salario} = -0{,}90 + 0{,}54\, Educ$$

Interpretação do coeficiente angular estimado:

A cada um ano a mais de educação formal, a variação média no salário é de 0,54 dólar/hora.

Wooldridge, J. M. **Introdução à econometria.** São Paulo: Pioneira Thomson Learning, 2006.

Insper

Interpretação das estimativas dos coeficientes de um modelo de regressão

Modelos lineares nos coeficientes, mas não lineares em algumas das variáveis

Modelos Linearizáveis

Modelo Padrão:

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

exponencial

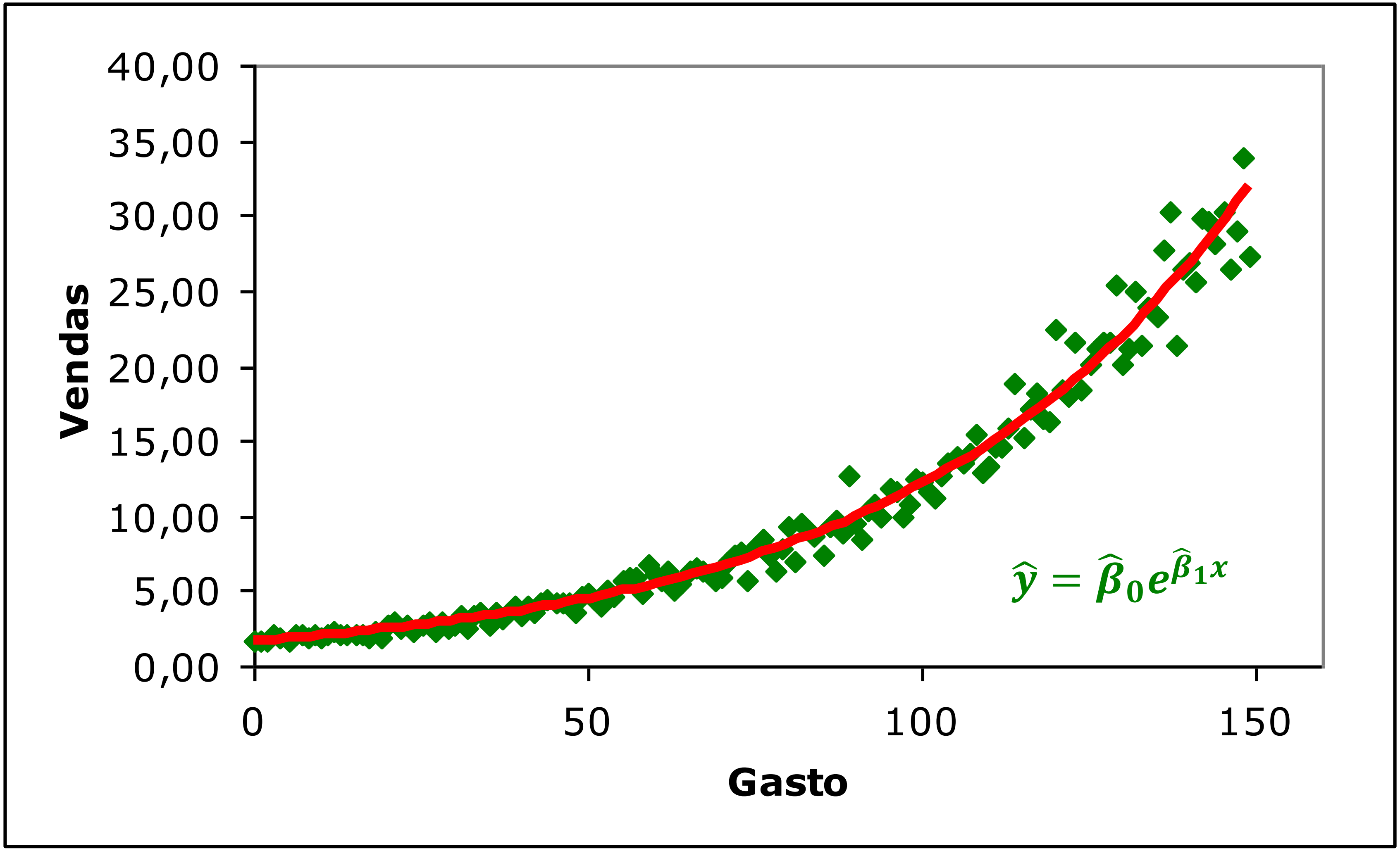
$$Y_i = \beta_0 e^{\beta_1 x_i} \varepsilon_i \rightarrow \ln Y_i = \ln \beta_0 + \beta_1 x_i + \ln \varepsilon_i \rightarrow Y_i' = \beta_0' + \beta_1 x_i + \varepsilon_i'$$

potencial

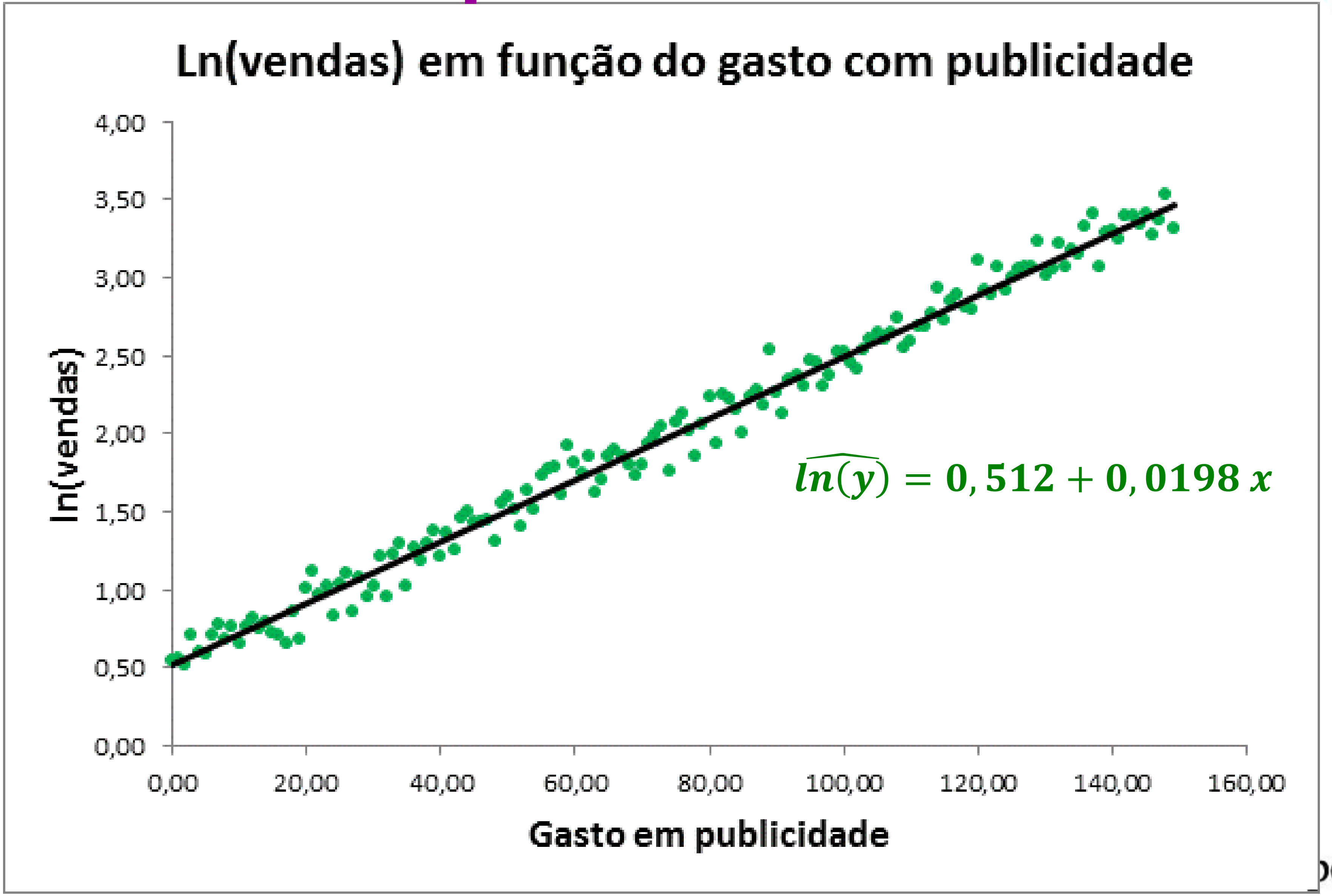
$$Y_i = \beta_0 x_i^{\beta_i} \varepsilon_i \rightarrow \ln Y_i = \ln \beta_0 + \beta_1 \ln x_i + \ln \varepsilon_i \rightarrow Y_i' = \beta_0' + \beta_1 x_i' + \varepsilon_i'$$

Caso tenha transformação na(s) variável(is), é necessário ter cuidado com a interpretação das estimativas dos coeficientes.

Um exemplo de transformação na variável resposta



Um exemplo de transformação na variável resposta



Transformações Logarítmicas

Transformações logarítmicas nos permitem modelar relações em termos "percentuais" (Na economia, essas relações são conhecidas como elasticidades).

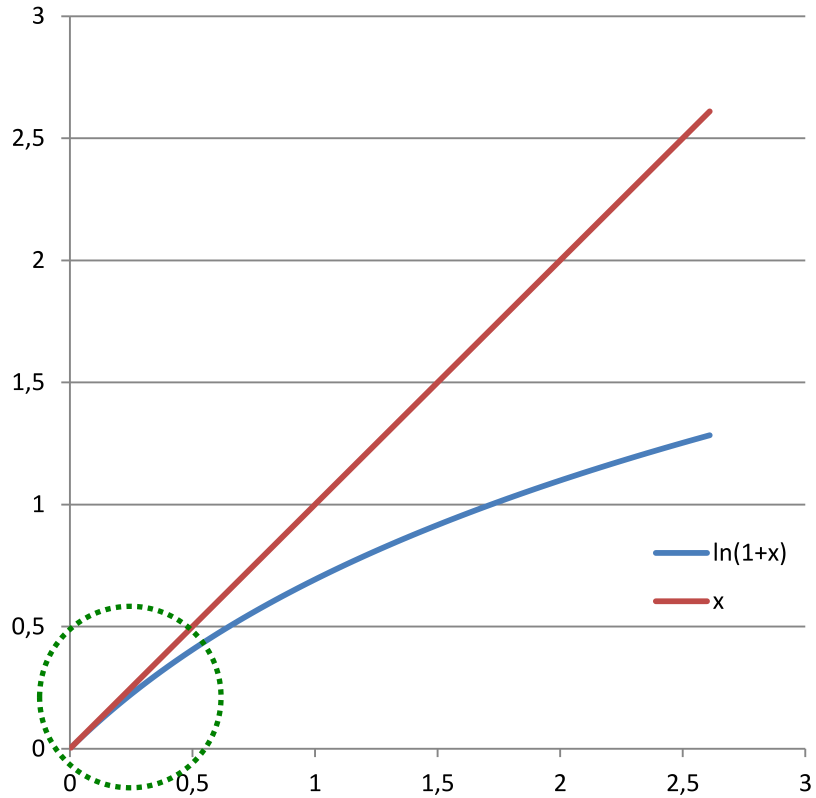
Resultado:

$$\ln(1+x) \cong x \; quando \; x \to 0$$

Propriedade usada nas variáveis transformadas:

$$\ln(x+\Delta x) - \ln(x) =$$

$$= \ln\left(1+\frac{\Delta x}{x}\right) \cong \frac{\Delta x}{x}$$



Insper

Modelo de regressão linear simples – Lin-Log

Reta estimada:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 \ln(x_i)$$

Interpretação do coeficiente angular estimado:

De maneira geral, a cada variação percentual $\%\Delta x$ na variável explicativa x, $\hat{\beta}_1$ tem interpretação de variação prevista na variável resposta quando dividido 100:

$$\frac{\hat{\beta}_1}{100} = \frac{\Delta \hat{y}}{\%\Delta x}$$

Modelo de regressão linear simples – Lin-Log

Reta estimada:

$$\widehat{Nota} = 557{,}8 + 36{,}4\ \ln(Renda)$$

Interpretação do coeficientes angular estimado:

- ✓ A cada aumento de 1% na Renda, há um aumento previsto de 0,36 pontos na nota da prova.

Modelo de regressão linear simples – Log-Lin

Reta estimada:

$$\widehat{\ln(y_i)} = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

Interpretação do coeficiente angular estimado:

De maneira geral, a cada variação Δx na variável explicativa x, $\hat{\beta}_1$ tem interpretação de variação percentual prevista na variável resposta quando multiplicado por 100:

$$100\hat{\beta}_1 = \frac{\%\Delta\hat{y}}{\Delta x}$$

Modelo de regressão linear simples – Log-Lin

Reta estimada:

$$\ln(\widehat{Salario}) = 0{,}584 + 0{,}083\,Educ$$

Interpretação do coeficientes angular estimado:

✓ A cada um ano a mais de educação formal, o salário aumenta, em média, 8,3%.

Wooldridge, J. M. **Introdução à econometria.** São Paulo: Pioneira Thomson Learning, 2006.

Insper

Modelo de regressão linear simples – Log-Log

Reta estimada:

$$\widehat{\ln(y_i)} = \hat{\beta}_0 + \hat{\beta}_1 \ln(x_i)$$

Interpretação do coeficiente angular estimado:

De maneira geral, a cada variação percentual Δx na variável explicativa x, $\hat{\beta}_1$ tem interpretação de variação percentual prevista na variável resposta :

$$\hat{\beta}_1 = \frac{\%\Delta\hat{y}}{\%\Delta x}$$

Insper

Modelo de regressão linear simples – Log-Log

Reta estimada:

$$\ln(\widehat{Salario}) = 4{,}822 + 0{,}257\ \ln(Vendas)$$

Interpretação do coeficientes angular estimado:

✓ A cada aumento de 1% nas vendas da empresa, a variação prevista no salário dos diretores é de 0,257% - interpretação usual de elasticidade.

Wooldridge, J. M. **Introdução à econometria.** São Paulo: Pioneira Thomson Learning, 2006.

Insper

Resumo das formas funcionais envolvendo transformações logarítmicas

- Há três casos de modelos (Lin-Log; Log-Lin e Log-Log), podendo a transformação log ser apenas em x, apenas em y ou em ambas.
- Os coeficientes podem ser estimadores via MQO (mínimos quadrados ordinários).
- Os testes de hipóteses em β_i’s são os mesmos do que os utilizados em modelos de regressão Lin-Lin.
- Cuidado: a interpretação da estimativa do coeficiente angular difere com o caso de transformação.
- A escolha da variável transformada deve ser auxiliada por bom senso e principalmente análise gráfica.

Insper

ATENÇÃO: 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