Minicurso de Ciência de Dados

Instrutor: Cesar Augusto Taconeli

Universidade Federal do Paraná

10 de Fevereiro de 2020

Aula 1 – Regressão linear e polinomial

Sumário

- Análise de regressão
- Regressão linear simples
  - Exemplo 1 - Duração de erupções de um vulcão
  - Exemplo 2 - Desempenho em exame de Matemática
- Regressão linear múltipla
  - Exemplo 3 - Faturamento e publicidade de empresas
- Regressão polinomial
  - Exemplo 4 - Valor de imóveis e condição sócio-econômica

Análise de regressão

- O termo regressão (*regression*) deve-se ao estatístico inglês Francis Galton (século XIX);
- Em seus estudos, Galton levantou dados sobre alturas de casais e respectivos descendentes;
- Os dados revelaram relação crescente entre as alturas de pais e filhos;
- No entanto, Galton observou que casais muito altos geravam filhos também altos, mas em geral de menor estatura (mais próximos a uma altura média). O mesmo ocorria para casais que se destacavam por uma baixa estatura;
- Galton denominou este fenômeno como **regressão à média**, ou simplesmente **regressão**.

Regressão linear

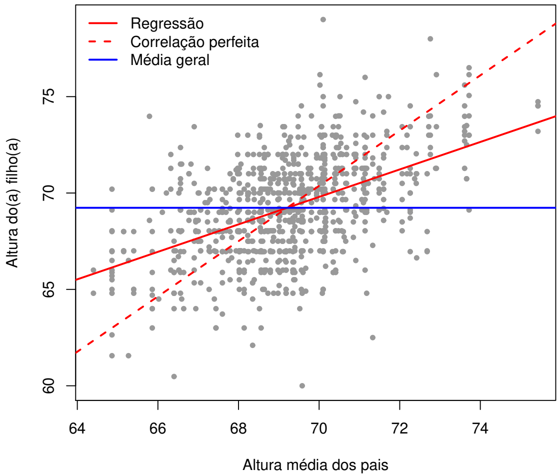

Figura 1: Ilustração - origem do termo regressão

Análise de regressão

- Modelos de regressão permitem descrever a relação (não determinística) entre uma variável de interesse (resposta) e uma ou mais covariáveis (preditoras)

- Uma análise de regressão pode ter diferentes objetivos, que em geral estão associados a duas finalidades principais:

    1. Modelos exploratórios: identificar e quantificar as relações entre a resposta e as covariáveis;

    2. Modelos preditivos: utilizar valores observados das covariáveis para predizer resultados não observados da resposta.

Análise de regressão

- Exemplos de problemas que envolvem modelos exploratórios:

  1. Há relação entre o tempo de uso do celular e o rendimento acadêmico dos alunos?
  2. Há relação entre o índice de massa corporal e características comportamentais (frequência de atividades físicas, horas de sono, número de refeições diárias...);
  3. Em quanto se espera que o rendimento anual de indivíduos de certa população aumente para cada ano a mais de escolaridade?
  4. A relação entre os desempenhos (notas) em exames de habilidades matemáticas e interpessoais é linear e crescente?

- Exemplos de problemas que envolvem modelos preditivos:

1. Qual o índice de rendimento acadêmico previsto, para um candidato recém aprovado em um vestibular, com base em suas notas no vestibular?
2. Qual o tempo de sobrevida previsto para um paciente diagnosticado com câncer no pâncreas com base em sua idade, sexo e outras variáveis referentes ao seu estado clínico?
3. Qual o faturamento previsto, para o próximo ano, de franquias de uma rede de *fast-food* com base nos faturamentos do ano atual, nos indicadores econômicos mais recentes e nos tamanhos das populações atendidas?
4. Qual a dose de um inseticida que mata 50% dos insetos?

Modelos de regressão

- Seja y a variável resposta e $\boldsymbol{x} = (x_1, x_2, ..., x_p)'$ o vetor de covariáveis.

- Uma formulação geral (mas não única) de modelos de regressão é a seguinte:

$$y = f(\boldsymbol{x}; \boldsymbol{\beta}) + \epsilon,$$

em que $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, ..., \beta_p)'$ é um vetor de parâmetros (constantes) e ϵ é o erro aleatório.

- Um modelo de regressão linear fica definido pela seguinte função:

$$f(\boldsymbol{x}; \boldsymbol{\beta}) = \boldsymbol{x'}\boldsymbol{\beta} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p$$

Regressão linear

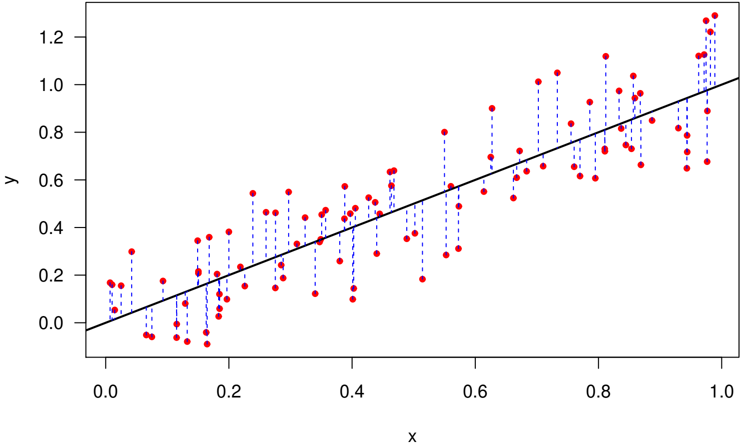

Figura 2: Ilustração - regressão linear simples

Regressão linear

- Os parâmetros $(\beta's)$ configuram a relação entre as variáveis;
- Na prática os parâmetros são desconhecidos, e devem ser estimados com base nos dados disponíveis (amostra);
- Vamos considerar a regressão linear com duas variáveis (a resposta, y, e uma covariável, x);
- Seja $\{(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)\}$ as observações de x e y em n indivíduos.

Exemplo 1 - Duração das erupções de um vulcão

- Para fins de ilustração, vamos considerar os dados sobre erupções de um vulcão (Old Faithful geyser in Yellowstone National Park, Wyoming, USA);

- Objetivo: modelar a duração da atual erupção (y) em função do tempo decorrido desde a erupção anterior (x);

- A relação entre as duas variáveis é aparentemente linear, induzindo o seguinte modelo:

$$y = \beta_0 + \beta_1 x + \epsilon.$$

Exemplo 1 - Duração de erupções de um vulcão

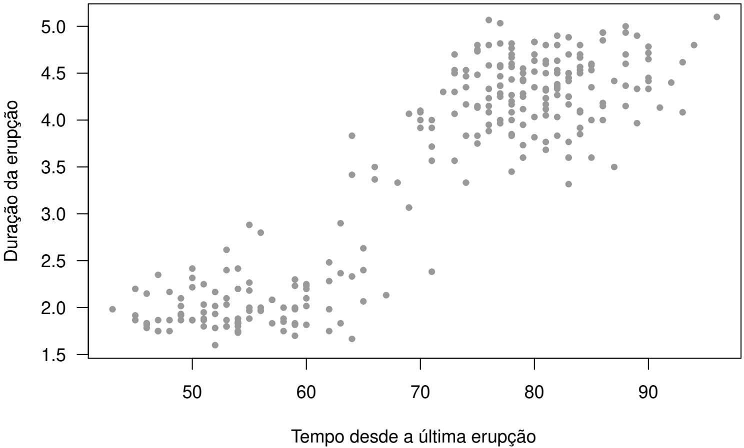

Figura 3: Duração da erupção vs tempo desde a erupção anterior

Método de mínimos quadrados

- A estimação dos parâmetros visa encontrar β_0 e β_1 tais que a reta de regressão mais se aproxime dos dados;
- Neste caso, deve-se encontrar os valores dos $\beta's$ tais que os erros sejam mínimos;
- Embora outras funções dos erros possam ser utilizadas, o mais usual é minimizar a soma de quadrados dos erros:

$$\text{SQE}(\beta_0, \beta_1) = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

Estimação por mínimos quadrados

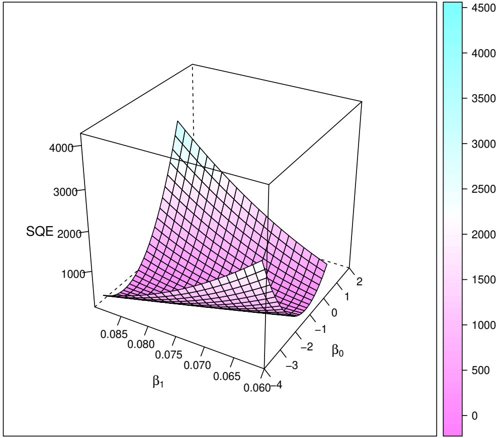

Figura 4: Gráfico da função soma de quadrados dos erros

Estimação por mínimos quadrados

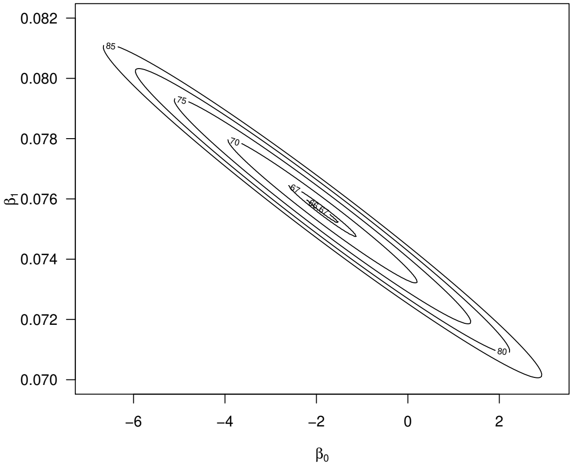

Figura 5: Curvas de nível da soma de quadrados dos erros

Método de mínimos quadrados

- Os estimadores de mínimos quadrados para β_0 e β_1, que vamos denotar por $\hat{\beta}_0$ e $\hat{\beta}_1$, são obtidos pela solução do seguinte sistema:

$$\frac{\partial \text{SQE}(\beta_0, \beta_1)}{\partial \beta_0} = 0 \quad \text{e} \quad \frac{\partial \text{SQE}(\beta_0, \beta_1)}{\partial \beta_1} = 0$$

- Para o caso da regressão linear simples temos:

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \quad \text{e} \quad \hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2},$$

em que $\bar{x}$ e $\bar{y}$ são as médias amostrais de x e y.

Exemplo 1 - Duração de erupções de um vulcão

- Para os dados das erupções do vulcão, obtemos pelo método de mínimos quadrados $\hat{\beta}_0 = -1.8740$ e $\hat{\beta}_1 = 0.0756$;
- Reta ajustada por mínimos qudrados:

$$\hat{y} = -1.8740 + 0.0756x$$

- Assim, espera-se que a duração de uma erupção aumente em 0.0756 minutos para cada minuto a mais desde a erupção mais recente;
- Se o tempo desde a última erupção é de uma hora (60 minutos), a duração da erupção pode ser prevista pelo modelo:

$$\hat{y} = -1.8740 + 0.0756 \times 60 = 2.662 \text{ minutos.}$$

Exemplo 1 - Duração de erupções de um vulcão

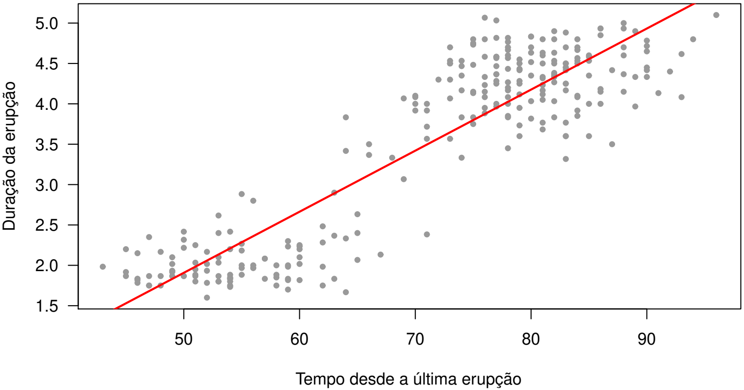

Figura 6: Duração da erupção vs tempo desde a erupção anterior com a reta de regressão ajustada

Exemplo 2 - Desempenho em teste de Matemática

- Nesta aplicação vamos considerar dados sobre desempenho de estudantes num teste de Matemática em 420 distritos do estado da Califórnia.

- As variáveis consideradas são as seguintes:

  - *Escore* - escore médio dos estudantes do distrito no teste de Matemática (y);

  - *Renda* - renda média dos habitantes do distrito (x).

- Diferentemente do caso anterior, aqui a relação entre as variáveis não parece ser linear.

Exemplo 2 - Desempenho em teste de Matemática

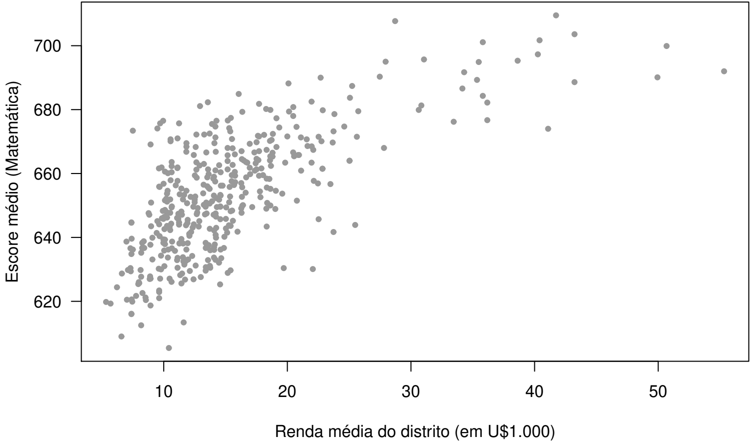

Figura 7: Escore vs renda média para os distritos do estado da Califórnia

Exemplo 2 - Desempenho em teste de Matemática

- Vamos ajustar quatro modelos distintos para esses dados:

  1. **Regressão linear:** $y = \beta_0 + \beta_1 x + \epsilon$;
  2. **Regressão quadrática:** $y = \beta_0 + \beta_1 x + \beta_2 x^2 + \epsilon$;
  3. **Regressão logarítmica:** $y = \beta_0 + \beta_1 \log(x) + \epsilon$;
  4. **Modelo não linear:** $y = \frac{\beta_0 x}{\beta_1 + x} + \epsilon$;

- Os quatro modelos foram ajustados usando o método de mínimos quadrados.

Exemplo 2 - Desempenho em teste de Matemática

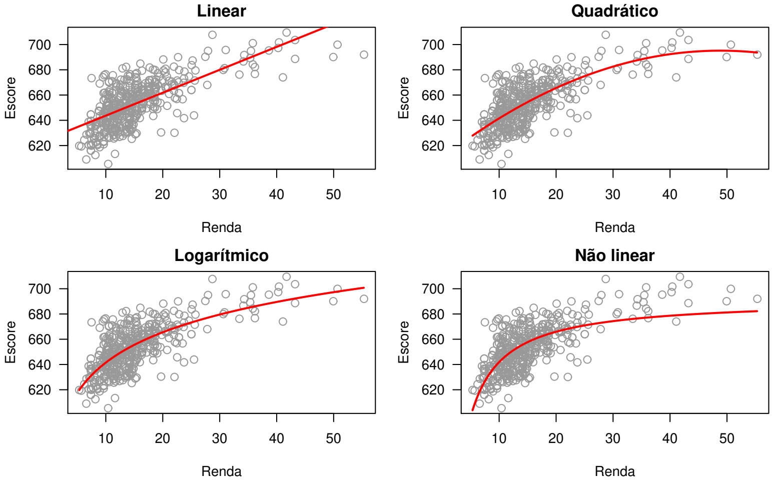

Figura 8: Diferentes ajustes para os dados de desempenho em Matemática

Exemplo 2 - Desempenho em teste de Matemática

- Claramente a regressão linear (reta de mínimos quadrados) produziu pior ajuste;
- Aparentemente a regressão logarítmica se ajusta melhor aos dados;
- Uma forma alternativa de avaliar a qualidade dos ajustes é por meio de um gráfico para os **resíduos**:

$$\textbf{Resíduo}_i = y_i - \hat{y}_i, \quad i = 1, 2, ..., n.$$

- Num gráfico de resíduos *vs* valores ajustados, espera-se, para um modelo bem ajustado, que os resíduos estejam dispersos aleatoriamente em torno de zero.

Exemplo 2 - Desempenho em teste de Matemática

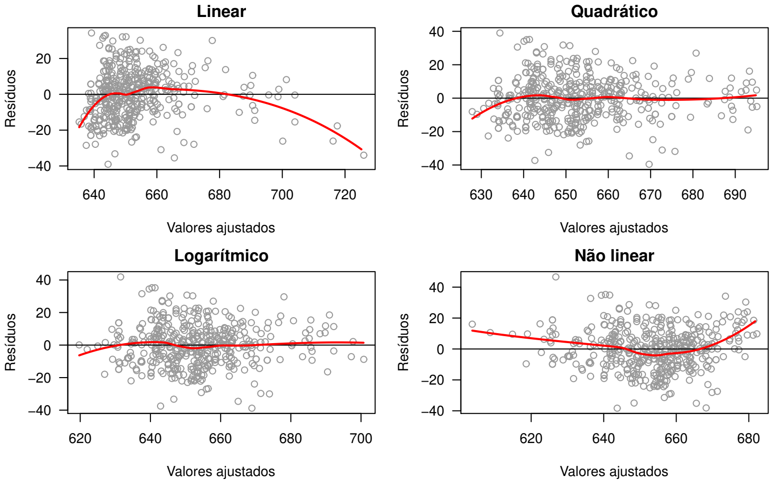

Figura 9: Análise dos resíduos dos modelos ajustados para os dados de desempenho em Matemática

- Uma medida útil para comparação de ajustes de modelos de regressão linear é o **coeficiente de determinação,** denotado por R^2:

$$R^2 = 1 - \frac{\text{SQRes}}{\text{SQTotal}},$$

em que SQRes é a soma de quadrados dos resíduos e SQTotal é a soma de quadrados total dos dados, corrigida pela média:

$$\text{SQRes} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2; \text{SQTotal} = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

- Desta forma, R^2 mede a proporção da variação dos dados explicada pelo modelo de regressão.

Exemplo 2 - Desempenho em teste de Matemática

Tabela 1: Coeficientes de determinação

Modelo	R^2
Linear	0.489
Quadrático	0.524
Logarítmico	0.529
Não linear	0.485

- O modelo logarítmico produziu maior valor de R^2, mesmo tendo um parâmetro a menos que o modelo quadrático.

Regressão linear múltipla

- O modelo de regressão linear múltipla é definido por $p \geq 2$ covariáveis:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p$$

- Para um conjunto (amostra) de n indivíduos, o modelo pode ser representado matricialmente por:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon},$$

em que

$$\boldsymbol{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \boldsymbol{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nk} \end{bmatrix}, \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix}, \boldsymbol{\epsilon} = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

- Assim, os estimadores de mínimos quadrados para $\beta_0, \beta_1, ..., \beta_p$ devem satisfazer:

$$\frac{\partial S(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \begin{bmatrix} \frac{\partial S(\boldsymbol{\beta})}{\partial \beta_0} \\ \frac{\partial S(\boldsymbol{\beta})}{\partial \beta_1} \\ \frac{\partial S(\boldsymbol{\beta})}{\partial \beta_2} \\ \vdots \\ \frac{\partial S(\boldsymbol{\beta})}{\partial \beta_p} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

sendo $S = S(\beta_0, \beta_1, ..., \beta_p) = \sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x_{i1} + ... + \beta_k x_{ip}))^2$

a soma de quadrados dos erros.

* Na forma matricial:

$$S(\boldsymbol{\beta}) = \sum_{i=1}^{n} \epsilon_i^2 = \boldsymbol{\epsilon'\epsilon} = (\boldsymbol{y} - \boldsymbol{X\beta})'(\boldsymbol{y} - \boldsymbol{X\beta}),$$

de maneira que o vetor $\hat{\boldsymbol{\beta}}$ tal que:

$$\left.\frac{\partial S}{\partial \boldsymbol{\beta}}\right|_{\hat{\boldsymbol{\beta}}} = \mathbf{0}$$

é o estimador de mínimos quadrados de $\boldsymbol{\beta}$, dado por:

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X'X})^{-1}\boldsymbol{X'y}.$$

Exemplo 3 - Faturamento e publicidade de empresas

- Vamos analisar dados sobre vendas e gastos em publicidade de 200 empresas. As variáveis são as seguintes:
  - `sales` (y): Total em vendas (em milhares de dólares);
  - `youtube` (x_1): Gastos em publicidade no *youtube*;
  - `facebook` (x_2): Gastos em publicidade no *facebook*;
  - `newspaper` (x_3): Gastos em publicidade em mídias impressas.

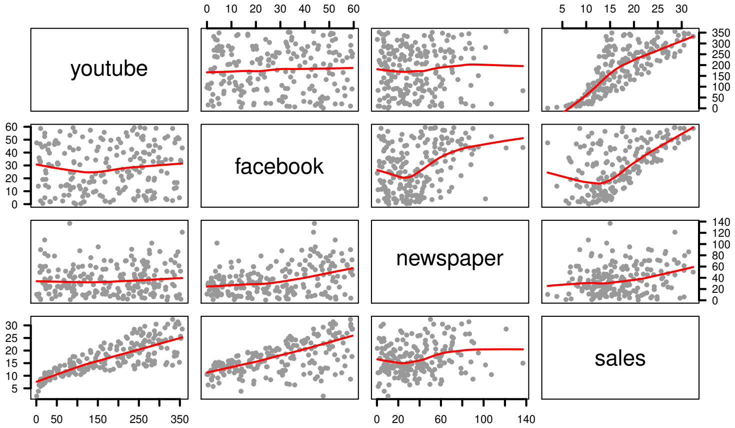

Figura 10: Faturamento em vendas e publicidade em mídias *online* e impressa

Exemplo 3 - Faturamento e publicidade de empresas

- Um extrato da base de dados é apresentado na sequência, na forma matricial, para fins de ilustração:

$$\boldsymbol{y} = \begin{bmatrix} 26.52 \\ 12.48 \\ 11.16 \\ \vdots \\ 8.64 \end{bmatrix}, \boldsymbol{X} = \begin{bmatrix} 1 & 276.12 & 45.36 & 83.04 \\ 1 & 53.40 & 47.16 & 54.12 \\ 1 & 20.64 & 55.08 & 83.16 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 278.52 & 10.32 & 10.44 \end{bmatrix}$$

- Calculando-se as estimativas de mínimos quadrados, obtemos o seguinte modelo ajustado:

$$\hat{y} = 3.5266 + 0.0457x_1 + 0.1885x_2 - 0.0010x_3$$

Exemplo 3 - Faturamento e publicidade de empresas

- Assim, estima-se:
  - Um aumento de 0.0457 no faturamento para cada unidade monetária a mais investida no *youtube* (mantendo-se fixos os investimentos nas outras mídias);
  - Um aumento de 0.1885 no faturamento para cada unidade monetária a mais investida no *facebook* (mantendo-se fixos os investimentos nas outras mídias);
  - Uma "redução"de 0.0010 no faturamento para cada unidade monetária a mais investida em mídias impressas (mantendo-se fixos os investimentos nas outras mídias).

Exemplo 3 - Faturamento e publicidade de empresas

- podemos predizer o faturamento para uma configuração qualquer de investimentos em publicidade. Por exemplo:
  - `youtube` 150 u.m.;
  - `facebook` 40 u.m.;
  - `newspaper` 30 u.m.
- O faturamento predito é:

$$\hat{y} = 3.5266 + 0.0457 \times 150 + 0.1885 \times 40 - 0.0010 \times 30 = 17.8916$$

- O coeficiente de determinação produzido pelo ajuste é $R^2 = 0.8972$ (aproximadamente 90% da variação dos faturamentos é explicada pelos investimentos nas três mídias consideradas).

Regressão polinomial

- Um modelo de regressão polinomial visa descrever a relação não linear entre a resposta e uma ou mais covariáveis incluindo potências das covariáveis originais ao modelo;

- Um modelo de regressão polinomial de ordem k com uma covariável é definido por:

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + ... + \beta_k x^k + \epsilon. \tag{1}$$

* Particularmente, o modelo polinomial de ordem 2 (quadrático) é definido por:

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \epsilon. \tag{2}$$

Regressão polinomial

- O uso de modelos polinomiais exige parcimônia, evitando-se o problema de *overfitting* (excesso de ajuste);
- A ordem do modelo, caso desconhecida a priori, deve ser a menor possível que ajuste satisfatoriamente os dados;
- Caso a relação entre as variáveis seja não linear e conhecida (mas não polinomial), deve-se optar pelo modelo correto

Regressão polinomial

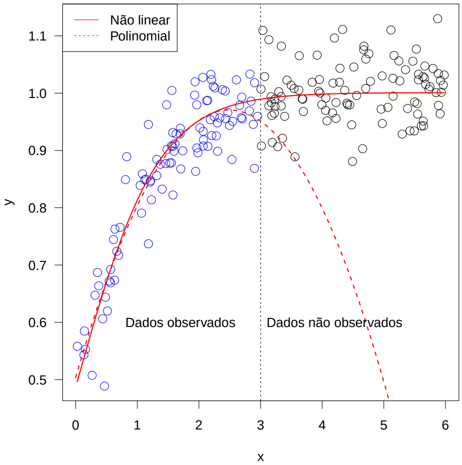

Figura 11: Comparação - modelo polinomial *vs* modelo não linear.

Exemplo 4 - Valor de imóveis e condição sócio-econômica

- Vamos analisar dados sobre valores de imóveis em sub-regiões da cidade de Boston-EUA.

  - `medv` (y): Valor mediano dos imóveis na região (em milhares de dólares);

  - `lstat` (x): Percentual dos habitantes pertencentes à classe social mais baixa.

Exemplo 4 - Valor de imóveis e condição sócio-econômica

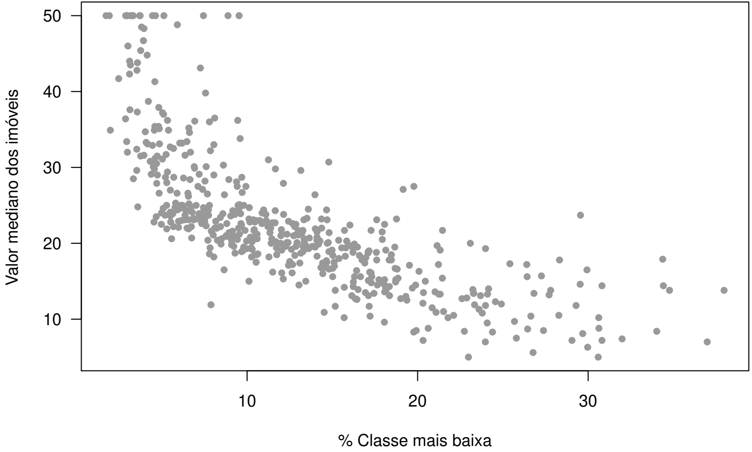

Figura 12: Valor mediano dos imóveis segundo percentual de habitantes na classe mais baixa.

Exemplo 4 - Valor de imóveis e condição sócio-econômica

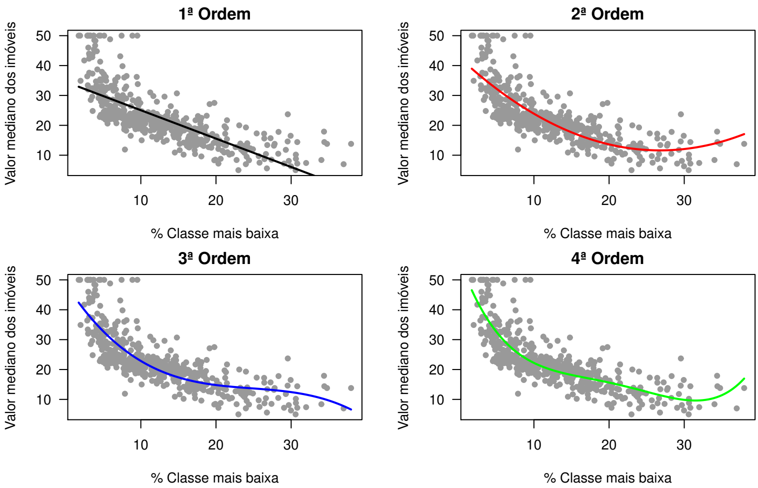

Figura 13: Modelos polinomiais ajustados.

Exemplo 4 - Valor de imóveis e condição sócio-econômica

- O modelo de 1ª ordem (reta) claramente não ajusta bem os dados.
- À medida que se aumenta a ordem do polinômio, os modelos ajustam melhor os dados, mas tornam-se mais 'ruidosos';
- Vamos analisar os coeficientes de determinação produzidos por modelos polinomiais de diferentes ordens.

Exemplo 4 - Valor de imóveis e condição sócio-econômica

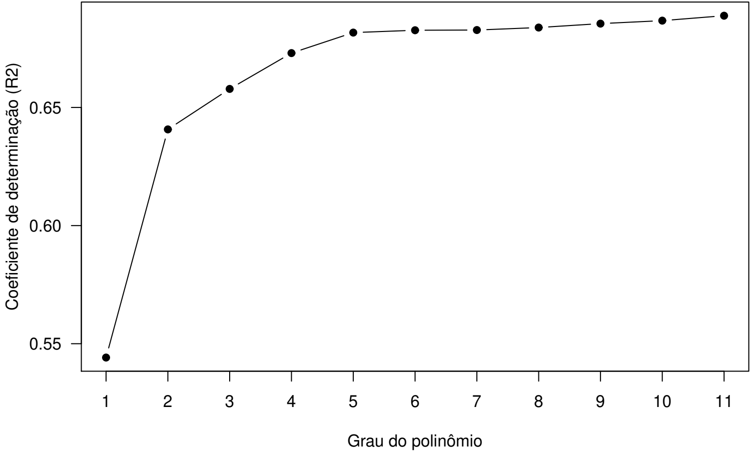

Figura 14: R^2 para polinômios com diferentes graus

Exemplo 4 - Valor de imóveis e condição sócio-econômica

- Agora, vamos analisar a capacidade preditiva dos modelos. Para isso a base foi dividida aleatoriamente em duas:
  - Base de ajuste: 400 observações;
  - Base de predição: as demais 106 observações.

- Modelos polinomiais de diferentes graus (1 a 11) foram ajustados usando a base de ajuste;

- Para cada modelo, usando a base de predição foi calculado e erro médio quadrático de predição (EQMP):

$$\text{EQMP} = \frac{1}{106}\sum_{i=1}^{106}(y_i - \hat{y}_i)^2 \tag{3}$$

Exemplo 4 - Valor de imóveis e condição sócio-econômica

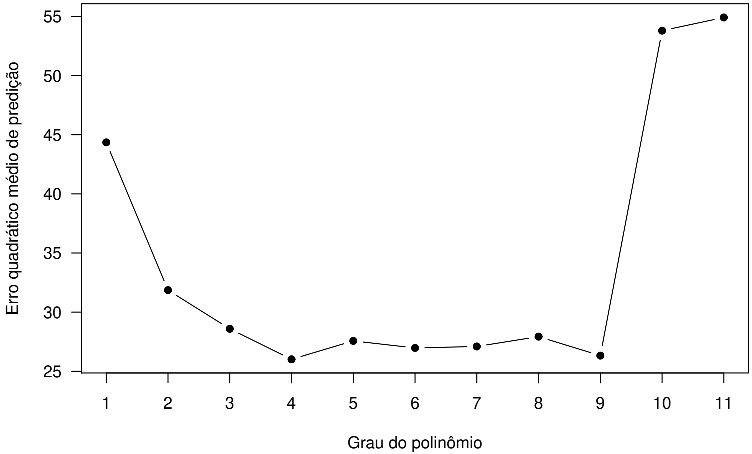

Figura 15: EQMP para polinômios com diferentes graus