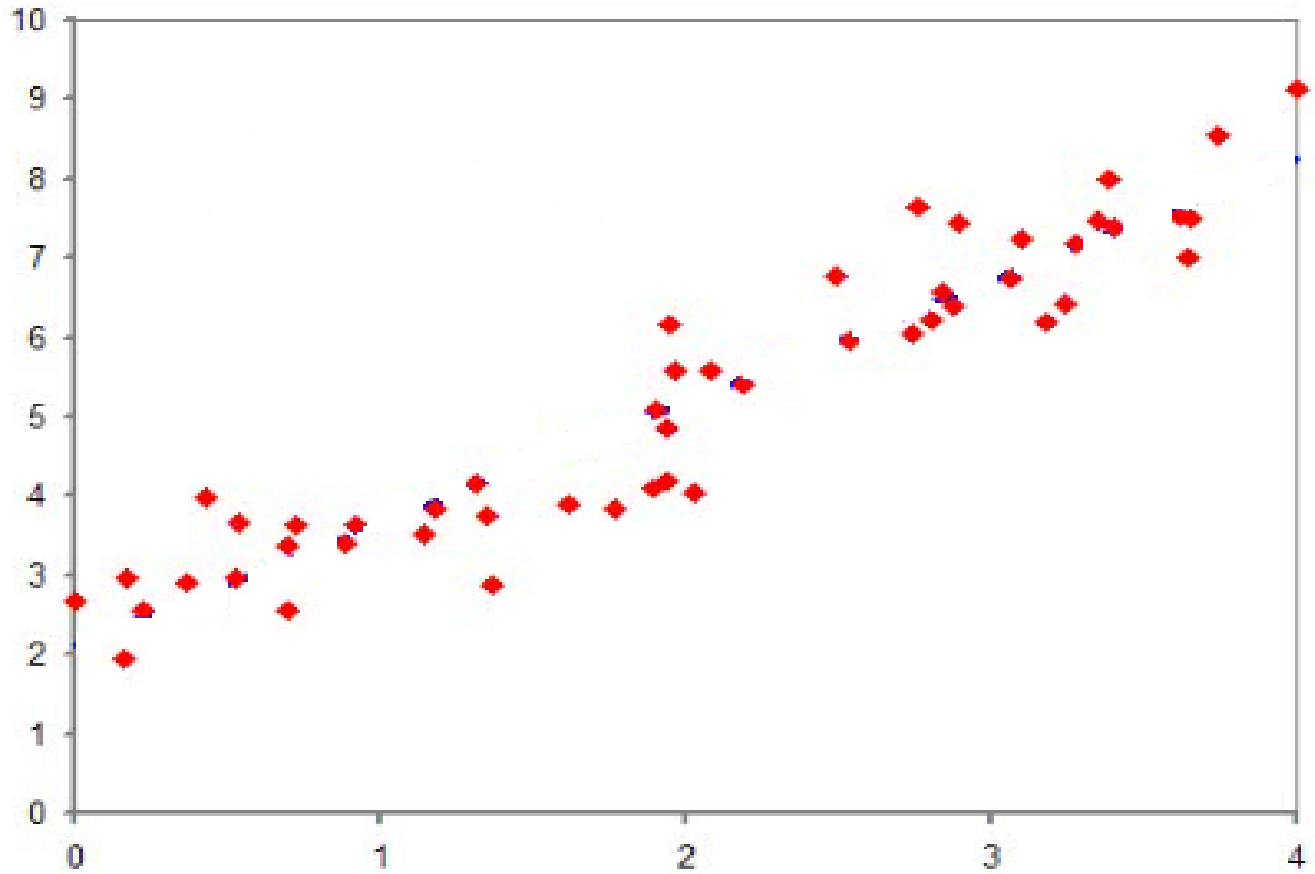
Regressão Linear Simples e Múltipla

Universidade Federal da Paraíba

Sumário

Regressão Linear Simples: Construção do Modelo

- Dados observados:



Regressão Linear Simples: Construção do Modelo

- Dados observados: Nuvem de pontos

$$\mathcal{O} = \{(x_1, y_1), (x_2, y_2), \ldots, (x_N, y_N)\}$$

- A ideia da regressão linear consiste em determinar os parâmetros a e b que minimizam a "distância" entre a reta $f(x) = ax + b$ e o conjunto de pontos $\mathcal{O}$.

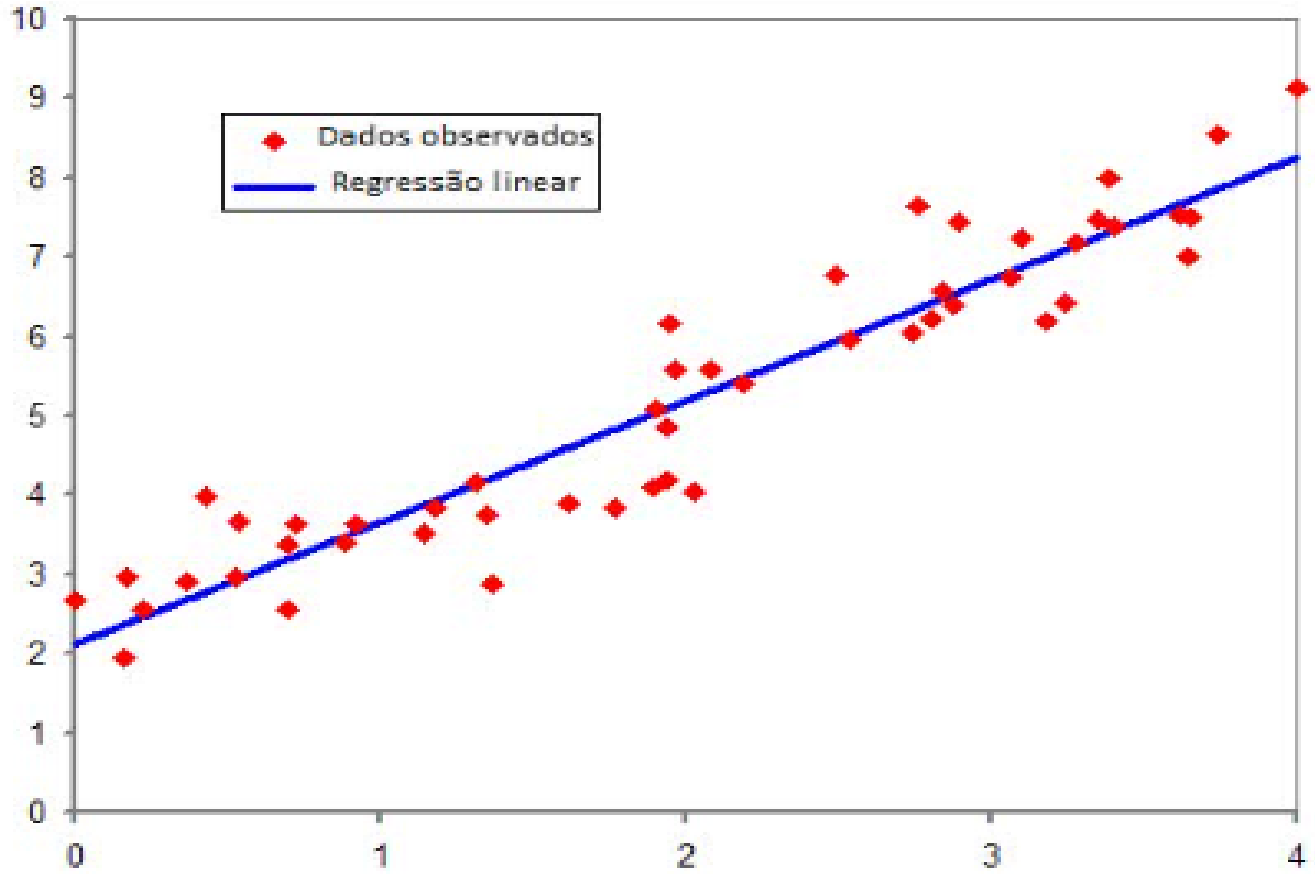
- Em termos matemáticos, a regressão linear pode ser escrita como o seguinte problema de minimização:

$$(a, b) = \min_{\alpha,\beta \in \mathbb{R}} d(f(x), \mathcal{O}),$$

onde a função d representa uma distância que será definida posteriormente.

Regressão Linear Simples: Construção do Modelo

- Representação de uma regressão linear



Regressão Linear Simples: Construção do Modelo

- Para construir o modelo, utilizaremos o método dos mínimos quadrados, que utiliza a métrica Euclidiana para mensurar a distância entre a reta do modelo de regressão linear e o conjunto de pontos. Matematicamente, temos o seguinte:

$$d(f(x), \mathcal{O}) = \sum_{i=1}^{N} (y_i - f(x_i))^2 = \underbrace{\sum_{i=1}^{N} (y_i - a.x_i - b)^2}_{F(a,b)}$$

- A partir daqui, a distância passa a ser considerada como uma função que depende de a (coeficiente angular) e b (coeficiente linear). Neste caso, para minimizar F, basta determinar os seus pontos críticos, ou seja, basta resolver

$$\nabla F(a, b) = 0 \quad \Rightarrow \quad (\hat{a}, \hat{b}) \leftarrow \text{Parâmetros Ótimos}$$

VALIDAÇÃO DO MODELO

Regressão Linear Simples: Validação do Modelo

- A regressão linear nem sempre configura uma boa representação para um conjunto de dados.
- A validação da regressão consiste em verificar se o modelo construído acima é aceitável como descrição dos dados.
- O processo de validação consiste em analisar:
  - a qualidade do ajuste da regressão;
  - a aleatoriedade dos resíduos da regressão;
  - o desempenho preditivo do modelo quando aplicado a dados não utilizados na estimativa do modelo.

Regressão Linear Simples: Validação do Modelo

- Características para validação do modelo:
  (1) Qualidade do ajuste;
  (2) Normalidade dos resíduos;
  (3) Ausência de outliers;
  (4) Independência dos resíduos.
  (5) Homocedasticidade;

Validação do Modelo: Normalidade dos Resíduos

- Principais métodos que podem ser utilizados para avaliar a normalidade de uma dada amostra:
  1. Análise Gráfica;
  2. Curtose (mede o achatamento da curva);
  3. Assimetria (mede a assimetria das caudas da distribuição);
  4. Teste de Shapiro-Wilk (S. S. Shapiro and M. B. Wilk. An analysis of variance test for normality (complete samples). Biometrika 52 (3–4), 591–611, 1965.);
  5. Teste de Anderson-Darling (T. W. Anderson and D. A. Darling. Asymptotic theory of certain "goodness of fit"criteria based on stochastic processes. Annals of Mathematical Statistics 23 (2), 193–212, 1952.).

Validação do Modelo: Normalidade dos Resíduos

Teste de Shapiro-Wilk:

- Estatística do teste: $W = \dfrac{\left(\sum_{i=1}^{N} a_i x_{(i)}\right)^2}{\sum_{i=1}^{N}(x_i - \bar{x})^2}$
- Teste de hipóteses (Nível de Significância = 5%):

$$\begin{cases} H_0 : \text{A distribuição dos resíduos é normal} \\ H_1 : \text{A distribuição dos resíduos não é normal} \end{cases}$$

- Utiliza-se o valor de W na Tabela de Shapiro para calcular o valor de p. Se $p > 0,05$, a hipótese nula deve ser aceita.

Observação:

O teste de Shapiro possui algumas limitações. Por exemplo, ele não é capaz de detectar normalidade em amostras pequenas.

Validação do Modelo: Normalidade dos Resíduos

Teste de Anderson-Darling

- Estatística do teste:

$$A^2 = -N - \sum_{i=1}^{N} \frac{2i-1}{N} [\ln(F(X_i)) + \ln(1 - F(X_{n+1-i}))]$$

- Teste de hipóteses (Nível de Significância = 5%):

$$\begin{cases} H_0 : \text{A distribuição dos resíduos é normal} \\ H_1 : \text{A distribuição dos resíduos não é normal} \end{cases}$$

- O valor da estatística é comparado com valores críticos específicos para determinar o valor de p. Se $p > 0,05$, a hipótese nula deve ser aceita.

Validação do Modelo: Identificação de Outliers

- Outliers são observações que se desviam significativamente das outras observações do conjunto de dados. Eles podem surgir devido a erros de medição, variabilidade natural nos dados ou condições experimentais específicas.
- Alguns problemas que podem ocorrer pela presença de outliers nos resíduos:
  - Eles podem influenciar nos coeficientes estimados.
  - Podem afetar a precisão das das inferências estatísticas, como intervalos de confiança e testes de hipóteses.
  - Outliers podem indicar problemas com os dados, como erros de entrada, medições imprecisas ou variáveis não observadas que influenciam a resposta.
  - Detectar e tratar outliers pode ajudar a desenvolver modelos mais robustos.

Validação do Modelo: Identificação de Outliers

- Principais métodos utilizados para detecção de outliers:
  1. Análise Gráfica: Boxplot, Scatterplot, Gráfico Q-Q;
  2. Métodos Estatísticos Simples:
     – Z-Score (valores mais distantes que 3 z-scores)
     – IQR (observações além de 1,5 vezes o IQR acima do terceiro quartil ou abaixo do primeiro quartil);
  3. Testes Específicos para Outliers: Grubbs, Dixon, Rosner;
  4. Métodos Baseados em Regressão: Resíduos Padronizados, Distância de Cook;

Validação do Modelo: Identificação de Outliers

Distância de Cook:

É uma medida utilizada para identificar observações influentes em uma análise de regressão linear. A Distância de Cook combina informações sobre a magnitude dos resíduos com a posição da observação em relação aos valores ajustados para determinar sua influência no modelo.

- Uma regra prática comum é considerar observações com uma Distância de Cook maior que $\frac{4}{N}$ (sendo N o número de observações) como possivelmente influentes.
- Observações identificadas como influentes devem ser investigadas para entender por que são influentes.

Validação do Modelo: Independência dos Resíduos

- A detecção de autocorrelação nos resíduos é crucial para garantir a validade das inferências feitas a partir de modelos de regressão.
- A presença de autocorrelação nos resíduos pode tornar os estimadores dos coeficientes de regressão enviesados e ineficientes.
- Resíduos não independentes podem indicar que o modelo não possui variáveis importantes, tem uma especificação funcional inadequada ou que há um padrão temporal ou espacial nos dados que não está sendo modelado adequadamente.
- Testes de significância (como os testes t e F) assumem que os resíduos são independentes. A autocorrelação pode inflacionar ou deflacionar os valores dos testes, levando a falsas rejeições ou aceitações da hipótese nula.

Validação do Modelo: Independência dos Resíduos

- Algumas técnicas e testes que podem ser usados para detectar a presença de autocorrelação nos resíduos:
  1. Análise Gráfica;
  2. Teste de Durbin-Watson (J. Durbin and G. S. Watson. Testing for serial correlation in least squares regression. Biometrika 58 (1), 1–19, 1971.);
  3. Teste de Breusch-Godfrey (L. G. Godfrey. Testing against general autoregressive and moving average error models when the regressors include lagged dependent variables. Econometrica 46, 1293–1301, 1978.);
  4. Modelos ARIMA (Auto-Regressive Integrated Moving Average)
  5. Métodos Não-paramétricos.

Validação do Modelo: Independência dos Resíduos

Teste de Durbin-Watson:

- É usado para detectar a presença de autocorrelação nos resíduos de um modelo de regressão.
- Estatística do teste: $d = \frac{\sum_{t=2}^{N}(R_t - R_{t-1})^2}{\sum_{t=1}^{N}(R_t)^2}$
- Teste de hipóteses: a hipótese nula é que não há autocorrelação entre os resíduos e o nível de significância é de 5%.
- Interpretação da Estatística Durbin-Watson:
  - $d = 2$ indica que não existe autocorrelação;
  - $d = 0$ indica autocorrelação positiva extrema;
  - $d = 4$ indica autocorrelação negativa extrema.

Validação do Modelo: Homocedasticidade

- A homocedasticidade ocorre quando os resíduos têm a mesma dispersão para todos os valores preditos pelo modelo de regressão.

- A homocedasticidade é um dos principais pressupostos da regressão linear. Sua importância para a validação do modelo pode ser destacada pelos seguintes motivos:
  - Validade das Inferências Estatísticas: Tanto os testes t e F quanto os intervalos de confiança dos coeficientes de regressão são baseados na suposição da homocedasticidade;
  - Eficácia dos Estimadores: Em presença de homocedasticidade, os estimadores de OLS são os melhores estimadores lineares não viesados;
  - Consistência das Previsões: A homocedasticidade assegura que a incerteza nas previsões do modelo seja constante.

Validação do Modelo: Homocedasticidade

- Principais formas de verificação da homocedasticidade:
  - Análise Gráfica;
  - Teste de Breusch-Pagan (T. S. Breusch and A. R. Pagan. A Simple Test for Heteroskedasticity and Random Coefficient Variation. Econometrica 47 (5), 1287–1294, 1979.);
  - Teste de White (H. White. A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity. Econometrica 48 (4), 817–838, 1980.);

Validação do Modelo: Homocedasticidade

Teste de Breusch-Pagan:

- É um teste qui-quadrado. A estatística do teste é distribuída $N\chi^2$, com k graus de liberdade.
- A estatística do teste é obtida a partir de uma regressão dos quadrados dos resíduos nas variáveis independentes.
- Teste de hipóteses: a hipótese nula é que ocorre a homocedasticidade e o nível de significância é de 5%.

Validação do Modelo: Homocedasticidade

Teste de White:

- Também segue uma distribuição qui-quadrado.
- Teste de hipóteses: a hipótese nula é que não há autocorrelação entre os resíduos e o nível de significância é de 5%.
- O teste de White é um teste mais geral que não assume uma forma específica para a variância dos resíduos. Isso o torna mais flexível e aplicável em uma ampla gama de situações onde a estrutura da heterocedasticidade pode ser complexa ou desconhecida.
- O teste de White pode ser aplicado em modelos de regressão múltipla, onde a variância dos resíduos pode depender de múltiplas variáveis explicativas e suas interações.