CE310 - Modelos de Regressão Linear

Diagnóstico do ajuste

Cesar Augusto Taconeli

25 de abril, 2025

Conteúdo

Introdução

Introdução

- Um modelo de regressão baseia-se em várias especificações (ou suposições).

Introdução

- Um modelo de regressão baseia-se em várias especificações (ou suposições).

- A avaliação dessas suposições é necessária para a validade do modelo ajustado e das consequentes inferências.

Introdução

- Um modelo de regressão baseia-se em várias especificações (ou suposições).

- A avaliação dessas suposições é necessária para a validade do modelo ajustado e das consequentes inferências.

- Após o ajuste do modelo, devemos avaliar a validade dessas suposições, bem como checar outros possíveis problemas de ajuste.

Introdução

- Um modelo de regressão baseia-se em várias especificações (ou suposições).

- A avaliação dessas suposições é necessária para a validade do modelo ajustado e das consequentes inferências.

- Após o ajuste do modelo, devemos avaliar a validade dessas suposições, bem como checar outros possíveis problemas de ajuste.

- Esta etapa da análise é comumente denominada *diagnóstico do ajuste* da regressão linear.

Introdução

- Os potenciais problemas quanto à má especificação de um modelo de regressão linear são:

- Os potenciais problemas quanto à má especificação de um modelo de regressão linear são:

1. $E(y|\boldsymbol{x}) \neq \boldsymbol{x}'\boldsymbol{\beta}$ (não linearidade);

Introdução

- Os potenciais problemas quanto à má especificação de um modelo de regressão linear são:

1. $E(y|\boldsymbol{x}) \neq \boldsymbol{x'\beta}$ (não linearidade);
2. Os erros não têm variância constante (σ^2) ou são correlacionados;

Introdução

- Os potenciais problemas quanto à má especificação de um modelo de regressão linear são:

1. $E(y|\boldsymbol{x}) \neq \boldsymbol{x}'\boldsymbol{\beta}$ (não linearidade);
2. Os erros não têm variância constante (σ^2) ou são correlacionados;
3. Os erros não têm distribuição normal;

Introdução

- Os potenciais problemas quanto à má especificação de um modelo de regressão linear são:

1. $E(y|\boldsymbol{x}) \neq \boldsymbol{x}'\boldsymbol{\beta}$ (não linearidade);
2. Os erros não têm variância constante (σ^2) ou são correlacionados;
3. Os erros não têm distribuição normal;
4. Presença de observações atípicas e mal ajustadas.

Exemplo- Gastos com apostas

Exemplo- Gastos com apostas

- Para fins de ilustração vamos considerar a base de dados `teengamb` da biblioteca `faraway` do `R`, que contém dados de 47 apostadores jovens. As variáveis são as seguintes:

Exemplo- Gastos com apostas

- Para fins de ilustração vamos considerar a base de dados **`teengamb`** da biblioteca **`faraway`** do `R`, que contém dados de 47 apostadores jovens. As variáveis são as seguintes:
- **`sex`**: 0=masculino, 1=feminino;

Exemplo- Gastos com apostas

- Para fins de ilustração vamos considerar a base de dados `teengamb` da biblioteca `faraway` do `R`, que contém dados de 47 apostadores jovens. As variáveis são as seguintes:
- `sex`: 0=masculino, 1=feminino;
- `status`: escore de status socioeconômico baseado na ocupação profissional dos pais;

Exemplo- Gastos com apostas

- Para fins de ilustração vamos considerar a base de dados `teengamb` da biblioteca `faraway` do `R`, que contém dados de 47 apostadores jovens. As variáveis são as seguintes:
- `sex`: 0=masculino, 1=feminino;
- `status`: escore de status socioeconômico baseado na ocupação profissional dos pais;
- `income`: renda semanal em pesos;

Exemplo- Gastos com apostas

- Para fins de ilustração vamos considerar a base de dados `teengamb` da biblioteca `faraway` do `R`, que contém dados de 47 apostadores jovens. As variáveis são as seguintes:
- `sex`: 0=masculino, 1=feminino;
- `status`: escore de status socioeconômico baseado na ocupação profissional dos pais;
- `income`: renda semanal em pesos;
- `verbal`: escore de proficiência verbal;

Exemplo- Gastos com apostas

- Para fins de ilustração vamos considerar a base de dados `teengamb` da biblioteca `faraway` do `R`, que contém dados de 47 apostadores jovens. As variáveis são as seguintes:
- `sex`: 0=masculino, 1=feminino;
- `status`: escore de status socioeconômico baseado na ocupação profissional dos pais;
- `income`: renda semanal em pesos;
- `verbal`: escore de proficiência verbal;
- `gamble`: gastos em apostas em pesos por ano (variável resposta).

Exemplo- Gastos com apostas

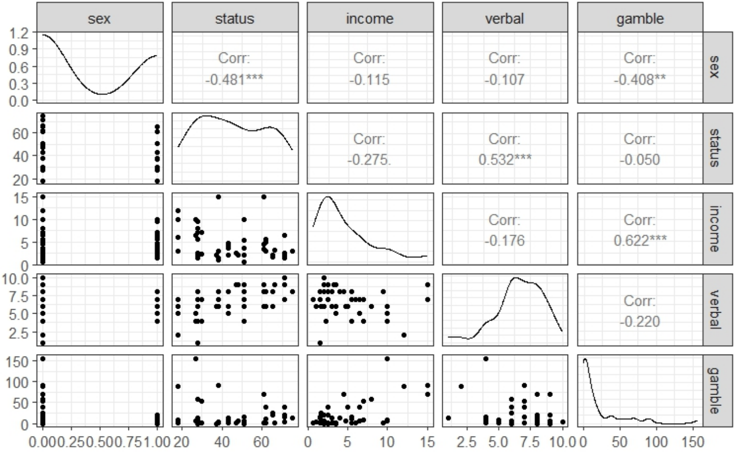

Figura 1: Gráficos de dispersão para os dados de 47 apostadores jovens

Exemplo- Gastos com apostas

- Nesta aplicação, consideramos o seguinte modelo de regressão linear:

$$\texttt{gamble} = \beta_0 + \beta_1\texttt{sex} + \beta_2\texttt{status} + \beta_3\texttt{income} + \beta_4\texttt{verbal} + \epsilon, \quad \epsilon \sim \mathrm{N}(0, \sigma^2)$$

Exemplo- Gastos com apostas

- Nesta aplicação, consideramos o seguinte modelo de regressão linear:

$$\texttt{gamble} = \beta_0 + \beta_1\texttt{sex} + \beta_2\texttt{status} + \beta_3\texttt{income} + \beta_4\texttt{verbal} + \epsilon, \quad \epsilon \sim \mathrm{N}(0, \sigma^2)$$

- Com base nos dados amostrais, o seguinte modelo foi ajustado por mínimos quadrados:

$$\widehat{\texttt{gamble}} = 22.55 - 22.11 \times \texttt{sex} + 0.05 \times \texttt{status} + 4.96 \times \texttt{income} - 2.95 \times \texttt{verbal}$$

Exemplo- Gastos com apostas

- Nesta aplicação, consideramos o seguinte modelo de regressão linear:

$$\texttt{gamble} = \beta_0 + \beta_1\texttt{sex} + \beta_2\texttt{status} + \beta_3\texttt{income} + \beta_4\texttt{verbal} + \epsilon, \quad \epsilon \sim \mathrm{N}(0, \sigma^2)$$

- Com base nos dados amostrais, o seguinte modelo foi ajustado por mínimos quadrados:

$$\widehat{\texttt{gamble}} = 22.55 - 22.11 \times \texttt{sex} + 0.05 \times \texttt{status} + 4.96 \times \texttt{income} - 2.95 \times \texttt{verbal}$$

- As análises subsequentes estão apresentadas nos scripts `R`.

Análise de resíduos

Análise de resíduos

- Os **resíduos ordinários** (ou simplesmente resíduos) de um modelo de regressão linear são definidos por:

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

Análise de resíduos

- Os **resíduos ordinários** (ou simplesmente resíduos) de um modelo de regressão linear são definidos por:

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

- O vetor de resíduos, $\mathrm{r}' = (r_1, r_2, ..., r_n)$, pode ser expresso na seguinte forma:

$$\mathrm{r} = (\mathrm{I} - \mathrm{H})\mathrm{y},$$

Análise de resíduos

- Os **resíduos ordinários** (ou simplesmente resíduos) de um modelo de regressão linear são definidos por:

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

- O vetor de resíduos, $\mathrm{r}' = (r_1, r_2, ..., r_n)$, pode ser expresso na seguinte forma:

$$\mathrm{r} = (\mathrm{I} - \mathrm{H})\mathrm{y},$$

em que $\mathrm{H} = \mathrm{X}(\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}'$ e I é a matriz identidade $n \times n$.

Resíduos em regressão linear

- Propriedades dos resíduos:

Resíduos em regressão linear

- Propriedades dos resíduos:

1. $E(\mathrm{r}) = \mathbf{0}$;

Resíduos em regressão linear

- Propriedades dos resíduos:

1. $E(\mathrm{r}) = \mathbf{0}$;

2. $\mathrm{Var}(\mathrm{r}) = \sigma^2(\mathrm{I} - \mathrm{H})$;

Resíduos em regressão linear

- Propriedades dos resíduos:

1. $E(\mathrm{r}) = \mathbf{0}$;

2. $\mathrm{Var}(\mathrm{r}) = \sigma^2(\mathrm{I} - \mathrm{H})$;

3. Podemos descrever a distribuição dos resíduos, de forma resumida, por:

$$r_i \sim \mathrm{Normal}(0, \sigma^2(1 - h_{ii}));$$

$$\mathrm{Cov}(r_i, r_i') = -\sigma^2(h_{ii'}), \quad i, i' = 1, 2, ..., n; i \neq i'.$$

Resíduos escalonados

- Resíduos escalonados são úteis para a identificação de valores extremos (outliers).

Resíduos escalonados

- Resíduos escalonados são úteis para a identificação de valores extremos (outliers).

- Uma primeira versão de resíduos escalonados são os **resíduos padronizados**, definidos por:

$$e_i = \frac{r_i}{\sqrt{\text{QM}_{\text{Res}}}}, \quad i = 1, 2, ..., n.$$

Resíduos escalonados

- Resíduos escalonados são úteis para a identificação de valores extremos (outliers).

- Uma primeira versão de resíduos escalonados são os **resíduos padronizados**, definidos por:

$$e_i = \frac{r_i}{\sqrt{\text{QM}_{\text{Res}}}}, \quad i = 1, 2, ..., n.$$

- Neste caso, QM_{Res} serve como estimativa para σ^2.

Resíduos escalonados

- Resíduos escalonados são úteis para a identificação de valores extremos (outliers).

- Uma primeira versão de resíduos escalonados são os **resíduos padronizados**, definidos por:

$$e_i = \frac{r_i}{\sqrt{\mathrm{QM}_{\mathrm{Res}}}}, \quad i = 1, 2, ..., n.$$

- Neste caso, $\mathrm{QM}_{\mathrm{Res}}$ serve como estimativa para σ^2.

- Observações com $|e_i| > 3$ são atípicas e devem ser investigadas.

Resíduos studentizados

- Os **resíduos studentizados** têm como vantagem adicional incorporar as variâncias dos resíduos no escalonamento, sendo definidos por:

$$t_i = \frac{r_i}{\sqrt{\mathrm{QM}_{\mathrm{Res}}(1 - h_{ii})}}, \quad i = 1, 2, ..., n.$$

Resíduos studentizados

- Os **resíduos studentizados** têm como vantagem adicional incorporar as variâncias dos resíduos no escalonamento, sendo definidos por:

$$t_i = \frac{r_i}{\sqrt{\mathrm{QM}_{\mathrm{Res}}(1 - h_{ii})}}, \quad i = 1, 2, ..., n.$$

- Por sua construção, os resíduos studentizados têm variância igual a um se o modelo especificado se ajustar aos dados.

Resíduos studentizados

- Os **resíduos studentizados** têm como vantagem adicional incorporar as variâncias dos resíduos no escalonamento, sendo definidos por:

$$t_i = \frac{r_i}{\sqrt{\mathrm{QM}_{\mathrm{Res}}(1 - h_{ii})}}, \quad i = 1, 2, ..., n.$$

- Por sua construção, os resíduos studentizados têm variância igual a um se o modelo especificado se ajustar aos dados.

- Resíduos studentizados são recomendados por facilitar a identificação de outliers e observações influentes.

Resíduos studentizados externamente

- **Resíduos studentizados externamente** fazem uso da estratégia *leave one out* na estimação de σ^2:

$$t_{(i)} = \frac{r_i}{\sqrt{\text{QM}_{\text{Res}_{(i)}}(1-h_{ii})}}, \quad i = 1, 2, ..., n,$$

em que $\text{QM}_{\text{Res}_{(i)}}$ é a estimativa de σ^2 gerada pelo modelo ajustado com $n-1$ observações (exceto a $i-$ésima).

Resíduos studentizados externamente

- **Resíduos studentizados externamente** fazem uso da estratégia *leave one out* na estimação de σ^2:

$$t_{(i)} = \frac{r_i}{\sqrt{\text{QM}_{\text{Res}_{(i)}}(1 - h_{ii})}}, \quad i = 1, 2, ..., n,$$

em que $\text{QM}_{\text{Res}_{(i)}}$ é a estimativa de σ^2 gerada pelo modelo ajustado com $n-1$ observações (exceto a $i-$ésima).

- Pode-se mostrar que o ajuste dos n modelos não é necessário para o cálculo de $\text{QM}_{\text{Res}_{(i)}}$, uma vez que:

$$\text{QM}_{\text{Res}_{(i)}} = \frac{(n-p)\text{QM}_{\text{Res}} - r_i^2/(1 - h_{ii})}{n - p - 1}.$$

Resíduos parciais

- **Resíduos parciais** permitem avaliar a relação entre a resposta e uma particular variável explicativa **ajustado o efeito das demais variáveis**.

Resíduos parciais

- **Resíduos parciais** permitem avaliar a relação entre a resposta e uma particular variável explicativa **ajustado o efeito das demais variáveis**.

- Suponha que o modelo ajustado contenha as variáveis $x_1, x_2, ..., x_k$. O resíduo parcial associado à variável x_j é definido por:

$$r_i^*(y|x_j) = r_i + \hat{\beta}_j x_{ij}, \quad i = 1, 2, ..., n.$$

Resíduos parciais

- **Resíduos parciais** permitem avaliar a relação entre a resposta e uma particular variável explicativa **ajustado o efeito das demais variáveis**.

- Suponha que o modelo ajustado contenha as variáveis $x_1, x_2, ..., x_k$. O resíduo parcial associado à variável x_j é definido por:

$$r_i^*(y|x_j) = r_i + \hat{\beta}_j x_{ij}, \quad i = 1, 2, ..., n.$$

- Observe que o resíduo parcial "desconta" do resíduo original o efeito de x_j.

Análise de resíduos

- Diversos gráficos podem ser construídos para checar o ajuste de modelos de regressão com base nos resíduos, dentre os quais:

Análise de resíduos

- Diversos gráficos podem ser construídos para checar o ajuste de modelos de regressão com base nos resíduos, dentre os quais:

1. Resíduos *vs* valores ajustados:
    - Padrões não lineares indicam relações não lineares não ajustadas;
    - Avaliar se os erros têm variância constante;
    - Identificar outliers.

Análise de resíduos

- Diversos gráficos podem ser construídos para checar o ajuste de modelos de regressão com base nos resíduos, dentre os quais:

1. Resíduos *vs* valores ajustados:
   - Padrões não lineares indicam relações não lineares não ajustadas;
   - Avaliar se os erros têm variância constante;
   - Identificar outliers.

2. Gráfico quantil-quantil:
   - Checar se os erros têm distribuição normal;
   - Identificar outliers.

Análise de resíduos

3. Resíduos *vs* ordem de coleta:
   - Analisar possível correlação nos dados induzida pela ordem de coleta (caso se aplique);

Análise de resíduos

3. Resíduos *vs* ordem de coleta:
    - Analisar possível correlação nos dados induzida pela ordem de coleta (caso se aplique);

4. Resíduos *vs* variável incluída no modelo
    - Verificar tendência não linear, indicativo de que o efeito da variável na resposta não é bem ajustado pelo modelo.
    - Avaliar variância não constante.

Análise de resíduos

3. Resíduos *vs* ordem de coleta:
   - Analisar possível correlação nos dados induzida pela ordem de coleta (caso se aplique);

4. Resíduos *vs* variável incluída no modelo
   - Verificar tendência não linear, indicativo de que o efeito da variável na resposta não é bem ajustado pelo modelo.
   - Avaliar variância não constante.

5. Resíduos parciais *vs* correspondente variável explicativa
   - Analisar a relação entre a resposta e a variável sob investigação ajustado o efeito das demais variáveis.

Análise de resíduos

3. Resíduos *vs* ordem de coleta:
   - Analisar possível correlação nos dados induzida pela ordem de coleta (caso se aplique);
4. Resíduos *vs* variável incluída no modelo
   - Verificar tendência não linear, indicativo de que o efeito da variável na resposta não é bem ajustado pelo modelo.
   - Avaliar variância não constante.
5. Resíduos parciais *vs* correspondente variável explicativa
   - Analisar a relação entre a resposta e a variável sob investigação ajustado o efeito das demais variáveis.
6. Resíduos *vs* variáveis não incluídas no modelo
   - Objetivos similares ao gráfico de resíduos parciais.

Padrões em gráficos de resíduos

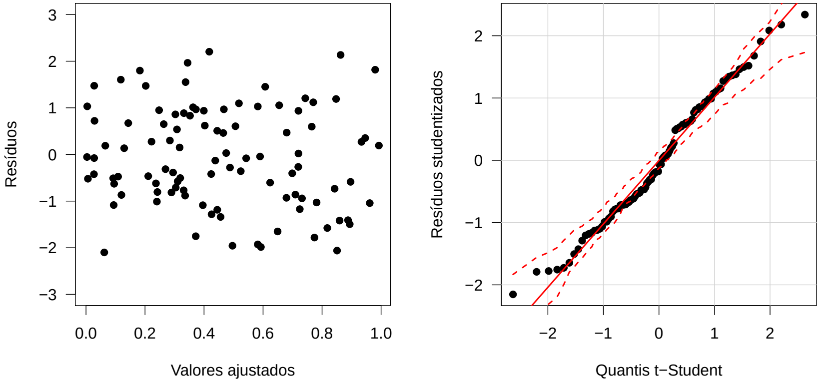

Figura 2: Ajuste satisfatório

Padrões em gráficos de resíduos

- O ajuste satisfatório da regressão linear é verificado por:

Padrões em gráficos de resíduos

- O ajuste satisfatório da regressão linear é verificado por:
  - Os resíduos estão dispersos aleatoriamente, centrados em zero;

Padrões em gráficos de resíduos

- O ajuste satisfatório da regressão linear é verificado por:
  - Os resíduos estão dispersos aleatoriamente, centrados em zero;
  - A dispersão dos resíduos é aproximadamente constante e igual a 1 (aproximadamente 95% dos resíduos entre -2 e 2; praticamente todos entre -3 e 3);

Padrões em gráficos de resíduos

- O ajuste satisfatório da regressão linear é verificado por:
  - Os resíduos estão dispersos aleatoriamente, centrados em zero;
  - A dispersão dos resíduos é aproximadamente constante e igual a 1 (aproximadamente 95% dos resíduos entre -2 e 2; praticamente todos entre -3 e 3);
  - Os resíduos têm distribuição bastante aderente à distribuição t-Student de referência (também seria apropriado verificar aderência à distribuição normal).

Padrões em gráficos de resíduos

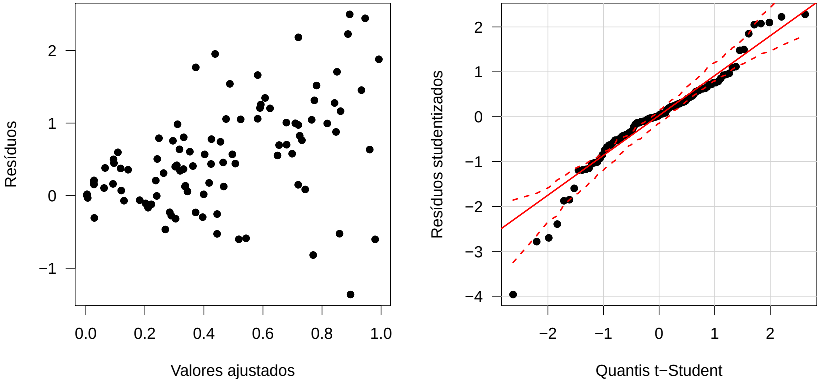

Figura 3: Variância não constante

Padrões em gráficos de resíduos

- Alternativas:

Padrões em gráficos de resíduos

- Alternativas:
    - Mínimos quadrados ponderados;

Padrões em gráficos de resíduos

- Alternativas:
  - Mínimos quadrados ponderados;
  - Transformação na variável resposta;

Padrões em gráficos de resíduos

- Alternativas:
    - Mínimos quadrados ponderados;
    - Transformação na variável resposta;
    - Modelos de regressão generalizados (caso particular, regressão para contagens).

Padrões em gráficos de resíduos

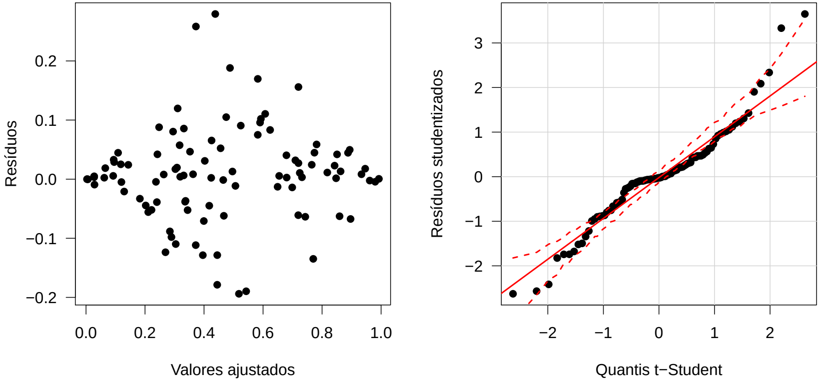

Figura 4: Variância não constante (2)

Padrões em gráficos de resíduos

- Alternativas:

Padrões em gráficos de resíduos

- Alternativas:
    - Mínimos quadrados ponderados;

Padrões em gráficos de resíduos

- Alternativas:
    - Mínimos quadrados ponderados;
    - Transformação na variável resposta;

Padrões em gráficos de resíduos

- Alternativas:
  - Mínimos quadrados ponderados;
  - Transformação na variável resposta;
  - Modelos de regressão generalizados (caso particular, regressão para proporções).

Padrões em gráficos de resíduos

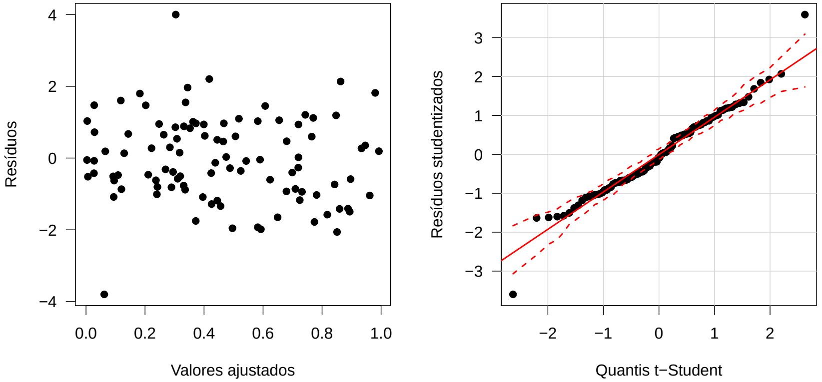

Figura 5: Presença de outliers

Padrões em gráficos de resíduos

- Alternativas:

Padrões em gráficos de resíduos

- Alternativas:
  - Investigação dos outliers (verificação; correção; remoção...);

Padrões em gráficos de resíduos

- Alternativas:
  - Investigação dos outliers (verificação; correção; remoção...);
  - Regressão quantílica;

Padrões em gráficos de resíduos

- Alternativas:
  - Investigação dos outliers (verificação; correção; remoção...);
  - Regressão quantílica;
  - Métodos de regressão robustos a outliers.

Padrões em gráficos de resíduos

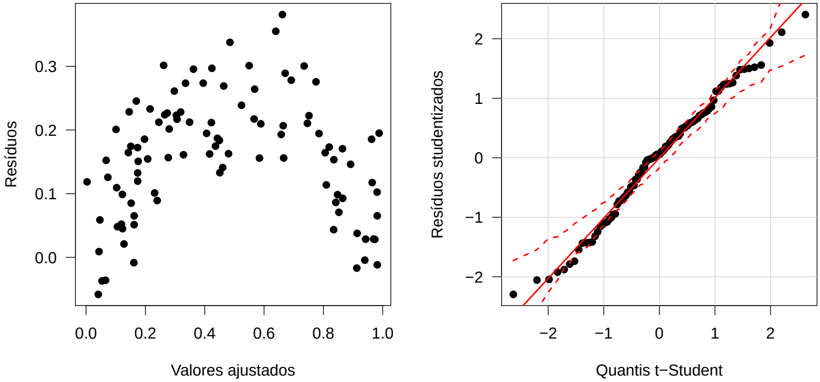

Figura 6: Não linearidade

Padrões em gráficos de resíduos

- Alternativas:

Padrões em gráficos de resíduos

- Alternativas:
    - Transformação nas variáveis;

Padrões em gráficos de resíduos

- Alternativas:
  - Transformação nas variáveis;
  - Regressão polinomial;

Padrões em gráficos de resíduos

- Alternativas:
    - Transformação nas variáveis;
    - Regressão polinomial;
    - Regressão não linear;

Padrões em gráficos de resíduos

- Alternativas:
  - Transformação nas variáveis;
  - Regressão polinomial;
  - Regressão não linear;
  - Regressão não paramétrica.

Padrões em gráficos de resíduos

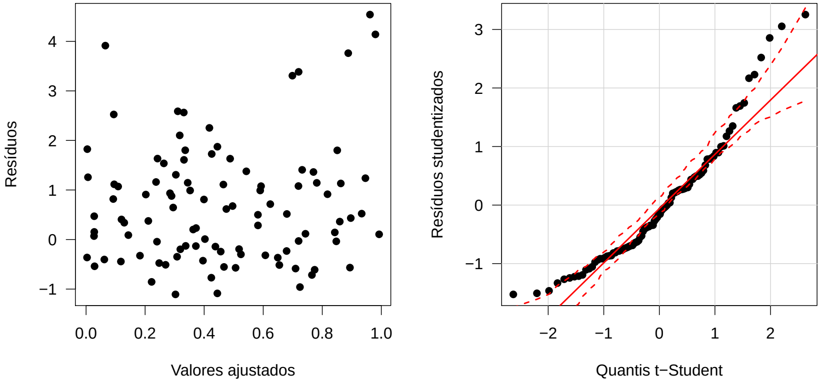

Figura 7: Erros com distribuição assimétrica

Padrões em gráficos de resíduos

- Alternativas:

Padrões em gráficos de resíduos

- Alternativas:
  - Transformação nas variáveis;

Padrões em gráficos de resíduos

- Alternativas:
  - Transformação nas variáveis;
  - Modelos de regressão generalizados;

Padrões em gráficos de resíduos

- Alternativas:
  - Transformação nas variáveis;
  - Modelos de regressão generalizados;
  - Modelos de análise de sobrevivência e confiabilidade.

Padrões em gráficos de resíduos

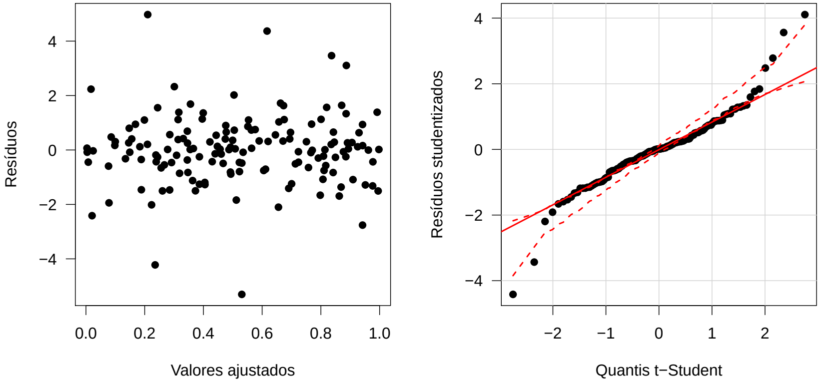

Figura 8: Erros com distribuição simétrica - caudas pesadas

Padrões em gráficos de resíduos

- Alternativas:

- Alternativas:
  - Transformação nas variáveis;

Padrões em gráficos de resíduos

- Alternativas:
    - Transformação nas variáveis;
    - Modelos de regressão generalizados.

Padrões em gráficos de resíduos

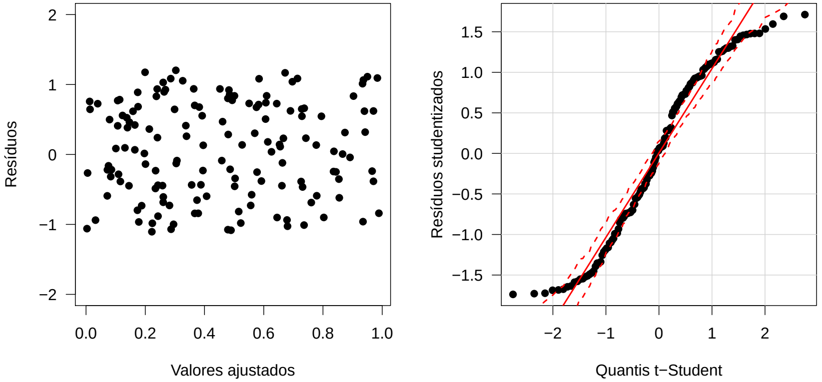

Figura 9: Erros com distribuição simétrica - caudas leves

Padrões em gráficos de resíduos

- Alternativas:

Padrões em gráficos de resíduos

- Alternativas:
    - Transformação nas variáveis;

Padrões em gráficos de resíduos

- Alternativas:
    - Transformação nas variáveis;
    - Modelos de regressão generalizados.

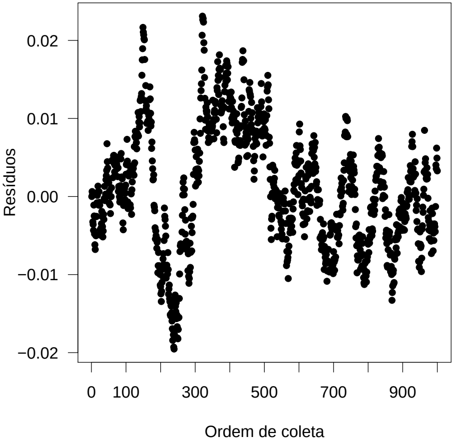

Figura 10: Erros auto-correlacionados

Padrões em gráficos de resíduos

- Alternativas:

Padrões em gráficos de resíduos

- Alternativas:
  - Mínimos quadrados generalizados;

Padrões em gráficos de resíduos

- Alternativas:
  - Mínimos quadrados generalizados;
  - Regressão para dados longitudinais ou espaciais;

Padrões em gráficos de resíduos

- Alternativas:
  - Mínimos quadrados generalizados;
  - Regressão para dados longitudinais ou espaciais;
  - Modelos de séries temporais.

Padrões em gráficos de resíduos

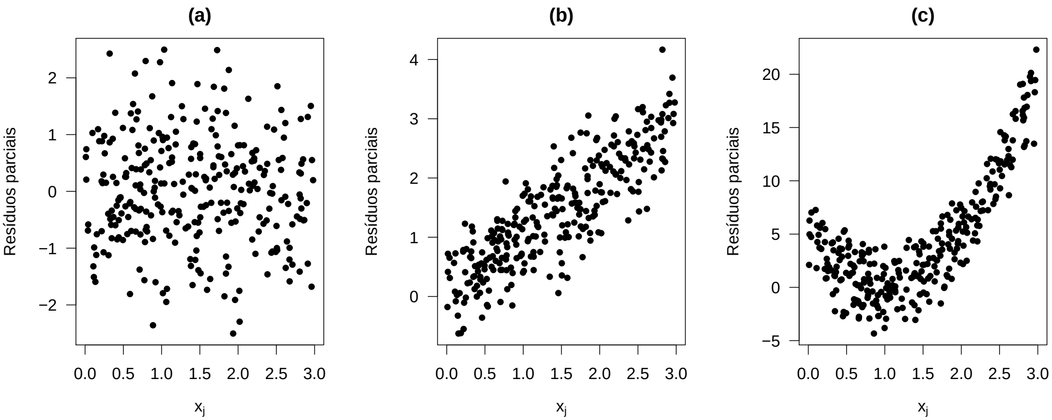

Figura 11: Gráficos de resíduos parciais: (a) Não efeito da variável (ajustado pelo efeito das demais); (b) Efeito linear; (c) Efeito não linear

Exemplo- Valores de imóveis

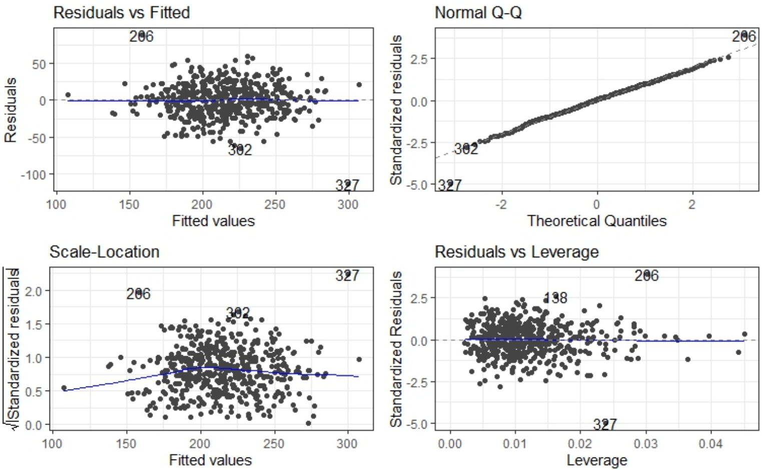

Figura 12: Análise de resíduos para os dados de valores de vendas de imóveis

Exemplo- Valores de imóveis

- No gráfico no canto superior direito, percebemos que os resíduos estão dispersos aleatoriamente em torno da média, sem exibir padrão não aleatório ou variância não constante.

Exemplo- Valores de imóveis

- No gráfico no canto superior direito, percebemos que os resíduos estão dispersos aleatoriamente em torno da média, sem exibir padrão não aleatório ou variância não constante.

- No canto superior direito, o gráfico quantil-quantil apresenta boa aderência dos resíduos à distribuição normal. Duas observações (206 e 327) são destacadas como possíveis pontos atípicos.

Exemplo- Valores de imóveis

- No gráfico no canto superior direito, percebemos que os resíduos estão dispersos aleatoriamente em torno da média, sem exibir padrão não aleatório ou variância não constante.

- No canto superior direito, o gráfico quantil-quantil apresenta boa aderência dos resíduos à distribuição normal. Duas observações (206 e 327) são destacadas como possíveis pontos atípicos.

- No canto inferior esquerdo, tendência crescente nos pontos indicaria variância não constante dos resíduos, o que não é verificado.

Exemplo- Valores de imóveis

- No gráfico no canto superior direito, percebemos que os resíduos estão dispersos aleatoriamente em torno da média, sem exibir padrão não aleatório ou variância não constante.

- No canto superior direito, o gráfico quantil-quantil apresenta boa aderência dos resíduos à distribuição normal. Duas observações (206 e 327) são destacadas como possíveis pontos atípicos.

- No canto inferior esquerdo, tendência crescente nos pontos indicaria variância não constante dos resíduos, o que não é verificado.

- O gráfico do canto inferior direito apresenta os resíduos versus valores de alavancagem, para diagnóstico de influência. Estudaremos isso mais adiante.

Testes de hipóteses

Testes de hipóteses

- Testes de hipóteses também podem ser aplicados para identificar padrões nos resíduos. Alguns exemplos:

Testes de hipóteses

- Testes de hipóteses também podem ser aplicados para identificar padrões nos resíduos. Alguns exemplos:

- **Testes de normalidade**:

Testes de hipóteses

- Testes de hipóteses também podem ser aplicados para identificar padrões nos resíduos. Alguns exemplos:

- **Testes de normalidade**:
    - D'Agostino's K-squared test;

Testes de hipóteses

- Testes de hipóteses também podem ser aplicados para identificar padrões nos resíduos. Alguns exemplos:

- **Testes de normalidade**:
  - D'Agostino's K-squared test;
  - Jarque–Bera test;

Testes de hipóteses

- Testes de hipóteses também podem ser aplicados para identificar padrões nos resíduos. Alguns exemplos:

- **Testes de normalidade**:
  - D'Agostino's K-squared test;
  - Jarque–Bera test;
  - Anderson–Darling test;

Testes de hipóteses

- Testes de hipóteses também podem ser aplicados para identificar padrões nos resíduos. Alguns exemplos:

- **Testes de normalidade**:
  - D'Agostino's K-squared test;
  - Jarque–Bera test;
  - Anderson–Darling test;
  - Cramér–von Mises;

Testes de hipóteses

- Testes de hipóteses também podem ser aplicados para identificar padrões nos resíduos. Alguns exemplos:

- **Testes de normalidade**:
  - D'Agostino's K-squared test;
  - Jarque–Bera test;
  - Anderson–Darling test;
  - Cramér–von Mises;
  - Lilliefors test;

Testes de hipóteses

- Testes de hipóteses também podem ser aplicados para identificar padrões nos resíduos. Alguns exemplos:

- **Testes de normalidade**:
  - D'Agostino's K-squared test;
  - Jarque–Bera test;
  - Anderson–Darling test;
  - Cramér–von Mises;
  - Lilliefors test;
  - Shapiro–Wilk test, dentre outros.

Testes de hipóteses

- Testes de hipóteses também podem ser aplicados para identificar padrões nos resíduos. Alguns exemplos:

- **Testes de normalidade**:
  - D'Agostino's K-squared test;
  - Jarque–Bera test;
  - Anderson–Darling test;
  - Cramér–von Mises;
  - Lilliefors test;
  - Shapiro–Wilk test, dentre outros.

- Nesses testes, a hipótese nula é a de normalidade, de forma que a rejeição confirma a não normalidade dos resíduos.

- **Teste de variância homogênea:**

Testes de hipóteses

- **Teste de variância homogênea**:
    - Teste de Bartlett;

Testes de hipóteses

- **Teste de variância homogênea**:
  - Teste de Bartlett;
  - Teste escore (ou de Breusch-Pagan).

- **Teste de variância homogênea**:
    - Teste de Bartlett;
    - Teste escore (ou de Breusch-Pagan).

- Em ambos os casos, a hipótese nula é a de variância homogênea, de tal forma que a rejeição indica a violação do pressuposto de variância constante.

Testes de hipóteses

- Para a hipótese de independência dos erros, um teste bastante útil é o de Durbin-Watson.

Testes de hipóteses

- Para a hipótese de independência dos erros, um teste bastante útil é o de Durbin-Watson.

- Esse teste é particularmente útil quando os dados são coletados ao longo do tempo, para investigar correlação temporal.

Testes de hipóteses

- Para a hipótese de independência dos erros, um teste bastante útil é o de Durbin-Watson.

- Esse teste é particularmente útil quando os dados são coletados ao longo do tempo, para investigar correlação temporal.

- A hipótese nula do teste, neste caso, é a de independência dos erros.

Testes de hipóteses

- Para a hipótese de independência dos erros, um teste bastante útil é o de Durbin-Watson.

- Esse teste é particularmente útil quando os dados são coletados ao longo do tempo, para investigar correlação temporal.

- A hipótese nula do teste, neste caso, é a de independência dos erros.

- No caso em que os dados são espacializados (coletados em diferentes locações do espaço), testes de dependência espacial são apropriados.

Testes de hipóteses

- O uso dos testes em substituição à análise gráfica é **altamente desaconselhável**, porque:

Testes de hipóteses

- O uso dos testes em substituição à análise gráfica é **altamente desaconselhável**, porque:

1. Testes de hipóteses não fornecem informações necessárias para avaliar adequadamente o desajuste e identificar medidas corretivas;

Testes de hipóteses

- O uso dos testes em substituição à análise gráfica é **altamente desaconselhável**, porque:

1. Testes de hipóteses não fornecem informações necessárias para avaliar adequadamente o desajuste e identificar medidas corretivas;

2. Desvios moderados (e aceitáveis) das suposições dos modelos podem produzir evidências significativas de desajuste para grandes amostras;

Testes de hipóteses

- O uso dos testes em substituição à análise gráfica é **altamente desaconselhável**, porque:

1. Testes de hipóteses não fornecem informações necessárias para avaliar adequadamente o desajuste e identificar medidas corretivas;

2. Desvios moderados (e aceitáveis) das suposições dos modelos podem produzir evidências significativas de desajuste para grandes amostras;

3. Para amostras pequenas, os testes podem não ter poder suficiente para indicar desvios consideráveis (e não aceitáveis) das suposições assumidas.

Observações atípicas

Observações atípicas

- Neste ponto vamos tratar de observações que apresentam comportamento atípico numa análise de regressão:

Observações atípicas

- Neste ponto vamos tratar de observações que apresentam comportamento atípico numa análise de regressão:

1. **Outliers:** Observações que não são bem ajustadas pelo modelo;

Observações atípicas

- Neste ponto vamos tratar de observações que apresentam comportamento atípico numa análise de regressão:

1. **Outliers:** Observações que não são bem ajustadas pelo modelo;

2. **Observações influentes:** Observações que afetam alguma característica do ajuste de maneira substancial;

Observações atípicas

- Neste ponto vamos tratar de observações que apresentam comportamento atípico numa análise de regressão:

1. **Outliers:** Observações que não são bem ajustadas pelo modelo;
2. **Observações influentes:** Observações que afetam alguma característica do ajuste de maneira substancial;
3. **Ponto de alavanca:** É um ponto extremo no espaço das variáveis explicativas.

Observações atípicas

- Neste ponto vamos tratar de observações que apresentam comportamento atípico numa análise de regressão:

1. **Outliers:** Observações que não são bem ajustadas pelo modelo;
2. **Observações influentes:** Observações que afetam alguma característica do ajuste de maneira substancial;
3. **Ponto de alavanca:** É um ponto extremo no espaço das variáveis explicativas.

- Uma mesma observação pode apresentar duas ou até mesmo as três características relacionadas simultaneamente.

Identificando observações não usuais

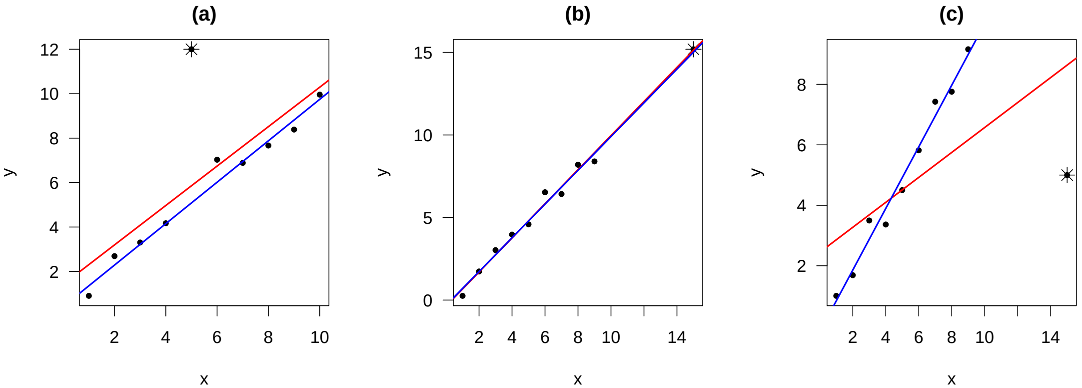

Figura 13: Observações atípicas - as retas em vermelho são ajustadas com todos os pontos e as azuis excluindo as observações atípicas.

Identificando observações não usuais

- As observações atípicas apresentadas na Figura 13 podem ser classificadas como:

Identificando observações não usuais

- As observações atípicas apresentadas na Figura 13 podem ser classificadas como:

(a) Outlier (a): trata-se de uma observação com valor extremo de y para o seu particular valor de x. No entanto, não pode ser classificado como ponto de alavanca ou influente;

Identificando observações não usuais

- As observações atípicas apresentadas na Figura 13 podem ser classificadas como:

(a) Outlier (a): trata-se de uma observação com valor extremo de y para o seu particular valor de x. No entanto, não pode ser classificado como ponto de alavanca ou influente;

(b) Ponto de alavanca (b): trata-se de uma observação com valor extremo de x. No entanto não é um ponto mal ajustado pelo modelo, nem tem grande influência no ajuste;

Identificando observações não usuais

- As observações atípicas apresentadas na Figura 13 podem ser classificadas como:

a) Outlier (a): trata-se de uma observação com valor extremo de y para o seu particular valor de x. No entanto, não pode ser classificado como ponto de alavanca ou influente;

b) Ponto de alavanca (b): trata-se de uma observação com valor extremo de x. No entanto não é um ponto mal ajustado pelo modelo, nem tem grande influência no ajuste;

c) A observação em (c) reúne as três características atípicas: é um ponto extremo quanto a x, claramente influente e mal ajustado pela reta de regressão (extremo quanto a y).

Outliers

- A maneira mais eficaz de identificar outliers é através da análise dos resíduos escalonados (por exemplo, os resíduos studentizados).

Outliers

- A maneira mais eficaz de identificar outliers é através da análise dos resíduos escalonados (por exemplo, os resíduos studentizados).
- Resíduos escalonados com valor absoluto maior que 3 são indicadores de outliers.

Outliers

- A maneira mais eficaz de identificar outliers é através da análise dos resíduos escalonados (por exemplo, os resíduos studentizados).

- Resíduos escalonados com valor absoluto maior que 3 são indicadores de outliers.

- Importante ter em mente que a existência de um "grande número de outliers" deve ser resultado da má especificação do modelo, e não propriamente indicador de observações atípicas.

Outliers

- A maneira mais eficaz de identificar outliers é através da análise dos resíduos escalonados (por exemplo, os resíduos studentizados).

- Resíduos escalonados com valor absoluto maior que 3 são indicadores de outliers.

- Importante ter em mente que a existência de um "grande número de outliers" deve ser resultado da má especificação do modelo, e não propriamente indicador de observações atípicas.

- Outliers devem ser cuidadosamente avaliados, investigando-se a causa e os possíveis efeitos no ajuste do modelo.

Outliers

- Dependendo da origem do outlier, a observação pode (e deve) ser excluída da análise.

Outliers

- Dependendo da origem do outlier, a observação pode (e deve) ser excluída da análise.

- Algumas causas que justificam a exclusão da observação são a coleta ou o registro incorreto do dado (se possível, ele deverá ser corrigido) e problemas nos instrumentos de medida, dentre outros.

Outliers

- Dependendo da origem do outlier, a observação pode (e deve) ser excluída da análise.

- Algumas causas que justificam a exclusão da observação são a coleta ou o registro incorreto do dado (se possível, ele deverá ser corrigido) e problemas nos instrumentos de medida, dentre outros.

- Em outros casos, não há uma justificativa de ordem operacional para excluir o outlier (a observação é atípica mas sua ocorrência é plausível).

Outliers

- Dependendo da origem do outlier, a observação pode (e deve) ser excluída da análise.

- Algumas causas que justificam a exclusão da observação são a coleta ou o registro incorreto do dado (se possível, ele deverá ser corrigido) e problemas nos instrumentos de medida, dentre outros.

- Em outros casos, não há uma justificativa de ordem operacional para excluir o outlier (a observação é atípica mas sua ocorrência é plausível).

- Nesses casos **não se deve eliminar a observação da análise** simplesmente com o objetivo de obter um melhor ajuste.

Outliers

- Um procedimento recomendável para a análise de regressão na presença de outliers é checar o efeito desses dados nos principais resultados do ajuste.

Outliers

- Um procedimento recomendável para a análise de regressão na presença de outliers é checar o efeito desses dados nos principais resultados do ajuste.

- Para isso, pode-se ajustar um novo modelo excluindo os outliers da base e comparar os resultados produzidos aos obtidos com o uso da base completa.

Outliers

- Um procedimento recomendável para a análise de regressão na presença de outliers é checar o efeito desses dados nos principais resultados do ajuste.

- Para isso, pode-se ajustar um novo modelo excluindo os outliers da base e comparar os resultados produzidos aos obtidos com o uso da base completa.

- Alterações substanciais nas estimativas, como trocas de sinais ou mudanças nas significâncias dos parâmetros devem ser relatadas, complementando a análise.

Pontos de alavanca

- Pontos de alavanca correspondem a observações com valores atípicos (extremos) no espaço das variáveis explicativas.

Pontos de alavanca

- Pontos de alavanca correspondem a observações com valores atípicos (extremos) no espaço das variáveis explicativas.

- Pontos remotos no espaço das variáveis explicativas são potencialmente (mas não necessariamente) pontos influentes, podendo alterar de maneira substancial as estimativas e correspondentes erros padrões, dentre outros.

Pontos de alavanca

- Pontos de alavanca correspondem a observações com valores atípicos (extremos) no espaço das variáveis explicativas.

- Pontos remotos no espaço das variáveis explicativas são potencialmente (mas não necessariamente) pontos influentes, podendo alterar de maneira substancial as estimativas e correspondentes erros padrões, dentre outros.

- A forma mais eficiente de detectar pontos de alavanca é através da matriz de projeção (ou matriz chapéu):

$$\mathrm{H} = \mathrm{X}(\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}'.$$

Pontos de alavanca

- Já vimos que $\hat{\mathrm{y}} = \mathrm{Hy}$. Desta forma:

$$\hat{y}_i = h_{i1}y_1 + h_{i2}y_2 + ... + h_{ii}y_i + ... + h_{in}y_n, \quad i = 1, 2, ..., n.$$

Pontos de alavanca

- Já vimos que $\hat{\mathrm{y}} = \mathrm{H}\mathrm{y}$. Desta forma:

$$\hat{y}_i = h_{i1}y_1 + h_{i2}y_2 + ... + h_{ii}y_i + ... + h_{in}y_n, \quad i = 1, 2, ..., n.$$

- Assim, h_{ii} pode ser interpretado como o peso exercido por y_i em seu próprio ajuste $(\hat{y}_i)$.

Pontos de alavanca

- Já vimos que $\hat{\mathrm{y}} = \mathrm{Hy}$. Desta forma:

$$\hat{y}_i = h_{i1}y_1 + h_{i2}y_2 + ... + h_{ii}y_i + ... + h_{in}y_n, \quad i = 1, 2, ..., n.$$

- Assim, h_{ii} pode ser interpretado como o peso exercido por y_i em seu próprio ajuste ($\hat{y}_i$).

- Observações com valores extremos para h_{ii} são pontos de alavancagem.

Pontos de alavanca

- Adicionalmente, pode-se mostrar que os elementos h_{ii} estão relacionados à distância de Mahalanobis da i-ésima observação ao centroide $\bar{\mathrm{x}}$.

Pontos de alavanca

- Adicionalmente, pode-se mostrar que os elementos h_{ii} estão relacionados à distância de Mahalanobis da i-ésima observação ao centroide $\bar{\mathrm{x}}$.

- A distância de Mahalanobis entre $\boldsymbol{x}_i$ e $\bar{\boldsymbol{x}}$ é dada por:

$$D(\mathrm{x_i}, \bar{\mathrm{x}}) = (\mathrm{x_i} - \bar{\mathrm{x}})'\mathrm{S}^{-1}(\mathrm{x_i} - \bar{\mathrm{x}}),$$

em que S é a matriz de covariâncias amostral de x.

Pontos de alavanca

- Adicionalmente, pode-se mostrar que os elementos h_{ii} estão relacionados à distância de Mahalanobis da i-ésima observação ao centroide $\bar{\mathrm{x}}$.

- A distância de Mahalanobis entre $\boldsymbol{x}_i$ e $\bar{\boldsymbol{x}}$ é dada por:

$$D(\mathrm{x_i}, \bar{\mathrm{x}}) = (\mathrm{x_i} - \bar{\mathrm{x}})'\mathrm{S}^{-1}(\mathrm{x_i} - \bar{\mathrm{x}}),$$

em que S é a matriz de covariâncias amostral de x.

- Assim, quanto mais afastada estiver x_i do centroide de x, maior o valor de h_{ii}.

Pontos de alavanca

- Outra propriedade importante de H é que seu traço é igual a p, sendo p o rank de X.

Pontos de alavanca

- Outra propriedade importante de H é que seu traço é igual a p, sendo p o rank de X.

- Assim, se cada observação contribuir igualmente para o seu próprio ajuste, teremos um h_{ii} médio, para cada observação, igual a p/n.

Pontos de alavanca

- Outra propriedade importante de H é que seu traço é igual a p, sendo p o rank de X.

- Assim, se cada observação contribuir igualmente para o seu próprio ajuste, teremos um h_{ii} médio, para cada observação, igual a p/n.

- É usual classificar uma observação i como sendo ponto de alavanca caso o correspondente h_{ii} seja maior que $2p/n$.

Pontos de alavanca

- Outra propriedade importante de H é que seu traço é igual a p, sendo p o rank de X.

- Assim, se cada observação contribuir igualmente para o seu próprio ajuste, teremos um h_{ii} médio, para cada observação, igual a p/n.

- É usual classificar uma observação i como sendo ponto de alavanca caso o correspondente h_{ii} seja maior que $2p/n$.

Nota: Observações com elevado h_{ii} e elevado resíduo studentizado são potenciais pontos influentes.

Observações influentes

- Observações influentes são aquelas que, quando removidas da base de dados, produzem expressiva mudança no ajuste do modelo.

Observações influentes

- Observações influentes são aquelas que, quando removidas da base de dados, produzem expressiva mudança no ajuste do modelo.

- As estratégias usadas para detecção de observações influentes fazem uso da estratégia *leave one out*.

Observações influentes

- Observações influentes são aquelas que, quando removidas da base de dados, produzem expressiva mudança no ajuste do modelo.

- As estratégias usadas para detecção de observações influentes fazem uso da estratégia *leave one out*.

- Neste caso, algum resultado (estimativas de parâmetros, predições,...) do modelo ajustado é avaliado em dois momentos: usando toda a base e mediante exclusão de cada observação da base (uma por vez).

Observações influentes

- Observações influentes são aquelas que, quando removidas da base de dados, produzem expressiva mudança no ajuste do modelo.

- As estratégias usadas para detecção de observações influentes fazem uso da estratégia *leave one out*.

- Neste caso, algum resultado (estimativas de parâmetros, predições,...) do modelo ajustado é avaliado em dois momentos: usando toda a base e mediante exclusão de cada observação da base (uma por vez).

- Na prática, não há necessidade de proceder os ajustes de todos os n modelos, mediante exclusão de cada observação, pois há resultados que permitem calcular as medidas de interesse usando apenas o ajuste para a base completa.

Observações influentes

- Uma das principais medidas de influência é a **distância de Cook**, baseada na diferença das estimativas de mínimos quadrados obtidas com as n observações ($\hat{\boldsymbol{\beta}}$) para as estimativas obtidas mediante exclusão da base da observação i ($\hat{\beta}_{(i)}$):

$$D_i = \frac{(\hat{\beta}_{(i)} - \hat{\beta})'\mathrm{X}'\mathrm{X}(\hat{\beta}_{(i)} - \hat{\boldsymbol{\beta}})}{p\mathrm{QM}_{Res}}, \quad i = 1, 2, ..., n.$$

Observações influentes

- Uma das principais medidas de influência é a **distância de Cook**, baseada na diferença das estimativas de mínimos quadrados obtidas com as n observações ($\hat{\boldsymbol{\beta}}$) para as estimativas obtidas mediante exclusão da base da observação i ($\hat{\beta}_{(i)}$):

$$D_i = \frac{(\hat{\beta}_{(i)} - \hat{\beta})' \mathrm{X}' \mathrm{X} (\hat{\beta}_{(i)} - \hat{\boldsymbol{\beta}})}{p \mathrm{QM}_{Res}}, \quad i = 1, 2, ..., n.$$

- Uma regra usual é classificar como influentes observações tais que $D_i > 1$.

Observações influentes

- Uma das principais medidas de influência é a **distância de Cook**, baseada na diferença das estimativas de mínimos quadrados obtidas com as n observações ($\hat{\boldsymbol{\beta}}$) para as estimativas obtidas mediante exclusão da base da observação i ($\hat{\beta}_{(i)}$):

$$D_i = \frac{(\hat{\beta}_{(i)} - \hat{\beta})'\mathrm{X}'\mathrm{X}(\hat{\beta}_{(i)} - \hat{\boldsymbol{\beta}})}{p\mathrm{QM}_{Res}}, \quad i = 1, 2, ..., n.$$

- Uma regra usual é classificar como influentes observações tais que $D_i > 1$.

- Uma forma equivalente de calcular D_i é dada por:

$$D_i = \frac{r_i^2}{p}\frac{h_{ii}}{1 - h_{ii}}, \quad i = 1, 2, ..., n.$$

Observações influentes

- **DFBetas:** Medem a alteração na estimativa de um particular β_j resultante da deleção da i-ésima observação:

$$\text{DFBetas}_{j,i} = \frac{\hat{\beta}_j - \hat{\beta}_{j(i)}}{\sqrt{\text{QM}_{Res_{(i)}} C_{jj}}}, \quad i = 1, 2, ..., n,$$

em que $\hat{\beta}_{j(i)}$ e $\text{QM}_{\text{Res}_{(i)}}$ são calculados mediante exclusão da i-ésima observação e C_{jj} é o j-ésimo elemento da diagonal de $(\text{X}'\text{X})^{-1}$.

Observações influentes

- **DFBetas:** Medem a alteração na estimativa de um particular β_j resultante da deleção da i-ésima observação:

$$\text{DFBetas}_{j,i} = \frac{\hat{\beta}_j - \hat{\beta}_{j(i)}}{\sqrt{\text{QM}_{Res_{(i)}} C_{jj}}}, \quad i = 1, 2, ..., n,$$

em que $\hat{\beta}_{j(i)}$ e $\text{QM}_{\text{Res}_{(i)}}$ são calculados mediante exclusão da i-ésima observação e C_{jj} é o j-ésimo elemento da diagonal de $(\text{X}'\text{X})^{-1}$.

- Recomenda-se investigar observações para as quais $|\text{DFBetas}_{j,i}| > 2/\sqrt{n}$.

Observações influentes

- **DFFITS:** Mede a alteração na predição ou valor ajustado de uma observação resultante de sua deleção:

$$\text{DFFITS}_i = \frac{\hat{y}_i - \hat{y}_{(i)}}{\sqrt{\text{QM}_{\text{Res}_{(i)}} h_{ii}}}.$$

Observações influentes

- **DFFITS:** Mede a alteração na predição ou valor ajustado de uma observação resultante de sua deleção:

$$\text{DFFITS}_i = \frac{\hat{y}_i - \hat{y}_{(i)}}{\sqrt{\text{QM}_{\text{Res}_{(i)}} h_{ii}}}.$$

- Recomenda-se investigar observações para as quais $|\text{DFFITS}_i| > 2/\sqrt{p/n}$.

Observações influentes

- Uma vez detectada uma ou mais observações influentes, é necessário avaliar adequadamente o impacto dessas observações nos principais resultados da análise.

Observações influentes

- Uma vez detectada uma ou mais observações influentes, é necessário avaliar adequadamente o impacto dessas observações nos principais resultados da análise.

- Quanto a desconsiderar definitivamente tais observações, as mesmas orientações apresentadas quanto ao tratamento de outliers se aplicam.

Observações influentes

- Uma vez detectada uma ou mais observações influentes, é necessário avaliar adequadamente o impacto dessas observações nos principais resultados da análise.

- Quanto a desconsiderar definitivamente tais observações, as mesmas orientações apresentadas quanto ao tratamento de outliers se aplicam.

- Novamente, deve-se avaliar criteriosamente se a presença de múltiplos outliers e obsevações influentes não se deve à má especificação do modelo.

Observações influentes

- Uma vez detectada uma ou mais observações influentes, é necessário avaliar adequadamente o impacto dessas observações nos principais resultados da análise.

- Quanto a desconsiderar definitivamente tais observações, as mesmas orientações apresentadas quanto ao tratamento de outliers se aplicam.

- Novamente, deve-se avaliar criteriosamente se a presença de múltiplos outliers e obsevações influentes não se deve à má especificação do modelo.

- Uma alternativa para análise na presença de observações atípicas é usar métodos robustos, que atribuam menor peso a tais observações no ajuste do modelo.

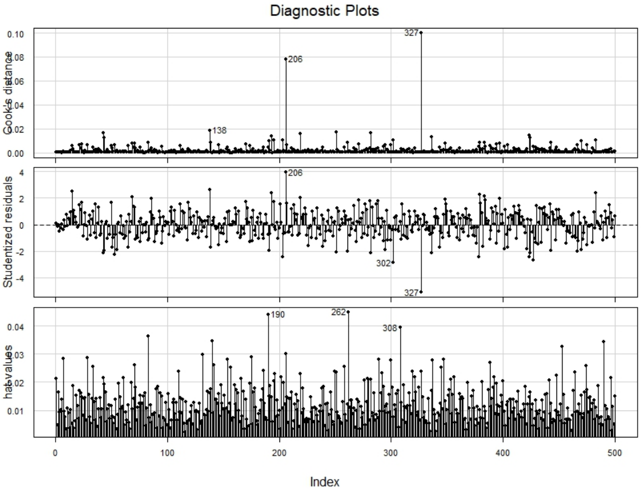

Figura 14: Diagnóstico de valores atípicos para os dados de vendas de imóveis

Exemplo- Valores de imóveis

- As observações de número 206 e 327 têm valores relativamente maiores para a distância de Cook, o que pode indicar elevada influência no ajuste do modelo.

Exemplo- Valores de imóveis

- As observações de número 206 e 327 têm valores relativamente maiores para a distância de Cook, o que pode indicar elevada influência no ajuste do modelo.

- As mesmas observações apresentam maiores resíduos, indicando que não são bem ajustadas pelo modelo.

Exemplo- Valores de imóveis

- As observações de número 206 e 327 têm valores relativamente maiores para a distância de Cook, o que pode indicar elevada influência no ajuste do modelo.

- As mesmas observações apresentam maiores resíduos, indicando que não são bem ajustadas pelo modelo.

- No gráfico dos valores da matriz H, não há pontos claramente destacados.

Exemplo- Valores de imóveis

- As observações de número 206 e 327 têm valores relativamente maiores para a distância de Cook, o que pode indicar elevada influência no ajuste do modelo.

- As mesmas observações apresentam maiores resíduos, indicando que não são bem ajustadas pelo modelo.

- No gráfico dos valores da matriz H, não há pontos claramente destacados.

- Vamos avaliar o impacto das observações nas linhas 206 e 327 no ajuste do modelo.

Exemplo- Valores de imóveis

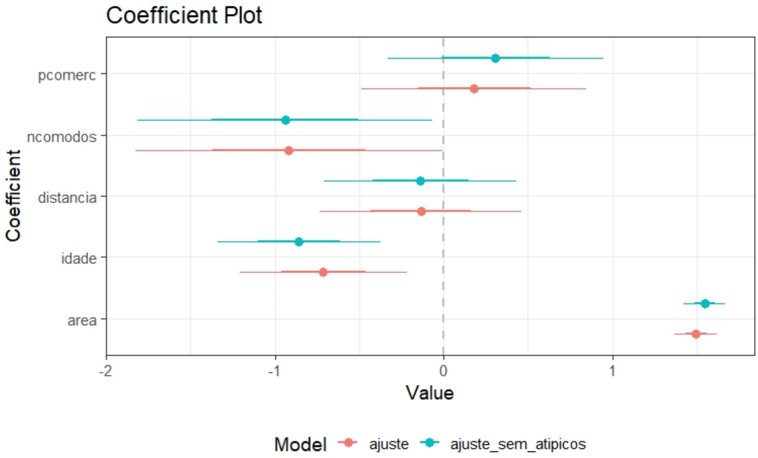

Figura 15: Estimativas e intervalos de confiança para os parâmetros de regressão ajustados com e sem as observações atípicas

Exemplo- Valores de imóveis

- Podemos notar que:

Exemplo- Valores de imóveis

- Podemos notar que:
  - As estimativas dos parâmetros são ligeiramente alteradas mediante exclusão do par de observações.

Exemplo- Valores de imóveis

- Podemos notar que:
  - As estimativas dos parâmetros são ligeiramente alteradas mediante exclusão do par de observações.
  - No entanto, não se pode notar alterações inferenciais mais importantes (como perda ou ganho de significância, troca do sinal do efeito) devido a essa exclusão.

Exemplo- Valores de imóveis

- Podemos notar que:
  - As estimativas dos parâmetros são ligeiramente alteradas mediante exclusão do par de observações.
  - No entanto, não se pode notar alterações inferenciais mais importantes (como perda ou ganho de significância, troca do sinal do efeito) devido a essa exclusão.
  - Desta forma, o impacto dessas observações no ajuste do modelo não aparenta ser significativo.

Exemplo- Valores de imóveis

- Podemos notar que:
  - As estimativas dos parâmetros são ligeiramente alteradas mediante exclusão do par de observações.
  - No entanto, não se pode notar alterações inferenciais mais importantes (como perda ou ganho de significância, troca do sinal do efeito) devido a essa exclusão.
  - Desta forma, o impacto dessas observações no ajuste do modelo não aparenta ser significativo.
  - Em geral, para amostras de tamanhos moderado a grande, o impacto de um pequeno número de observações atípicas torna-se negligenciável.

Multicolinearidade

Multicolinearidade

- A multicolinearidade se caracteriza por uma quase dependência linear entre as colunas de $\mathbf{X}$.

Multicolinearidade

- A multicolinearidade se caracteriza por uma quase dependência linear entre as colunas de $\mathbf{X}$.

- Se as colunas da matriz X (X_1, X_2, $\ldots$, X_p) forem exatamente colineares, ou seja, se houver um conjunto de constantes $c_1, c_2, ..., c_n$ nem todas nulas, tal que:

$$\sum_{j=1}^{p} c_j X_j = 0,$$

segue que $(X'X)$ é singular, não havendo solução única na estimação por mínimos quadrados.

Efeitos da multicolinearidade

- Nos casos em que as colunas da matriz X exibem uma quase dependência linear, como resultado tem-se baixa precisão (elevado erro) na estimação dos parâmetros do modelo.

Efeitos da multicolinearidade

- Nos casos em que as colunas da matriz X exibem uma quase dependência linear, como resultado tem-se baixa precisão (elevado erro) na estimação dos parâmetros do modelo.

- Para o modelo de regressão linear múltipla, a variância de $\hat{\beta}_j$, estimador de um particular parâmetro β_j do modelo, pode ser expressa por:

$$\text{Var}(\hat{\beta}_j) = \sigma^2 \left(\frac{1}{1 - R_j^2} \right) \frac{1}{\sum_{i=1}^{n} (x_{ij} - \bar{x}_j)^2},$$

em que R_j^2 é o coeficiente de determinação da regressão de x_j nas demais variáveis.

Efeitos da multicolinearidade

- Nos casos em que as colunas da matriz X exibem uma quase dependência linear, como resultado tem-se baixa precisão (elevado erro) na estimação dos parâmetros do modelo.

- Para o modelo de regressão linear múltipla, a variância de $\hat{\beta}_j$, estimador de um particular parâmetro β_j do modelo, pode ser expressa por:

$$\text{Var}(\hat{\beta}_j) = \sigma^2 \left(\frac{1}{1 - R_j^2} \right) \frac{1}{\sum_{i=1}^{n} (x_{ij} - \bar{x}_j)^2},$$

em que R_j^2 é o coeficiente de determinação da regressão de x_j nas demais variáveis.

- É fácil observar que $\text{Var}(\hat{\beta}_j) \to \infty$ quando $R_j^2 \to 1$.

Diagnóstico de multicolinearidade

- $\text{VIF}_j = 1/(1 - R_j^2)$ é chamado **fator de inflação da variância** e pode ser utilizado para diagnóstico de multicolinearidade.

Diagnóstico de multicolinearidade

- $\text{VIF}_j = 1/(1 - R_j^2)$ é chamado **fator de inflação da variância** e pode ser utilizado para diagnóstico de multicolinearidade.

- Se as colunas de X forem ortogonais (variáveis não correlacionadas), então $\text{VIF}_j = 1$ para todo j.

Diagnóstico de multicolinearidade

- $\text{VIF}_j = 1/(1 - R_j^2)$ é chamado **fator de inflação da variância** e pode ser utilizado para diagnóstico de multicolinearidade.

- Se as colunas de X forem ortogonais (variáveis não correlacionadas), então $\text{VIF}_j = 1$ para todo j.

- Quanto mais próximos de 1 os valores de VIF_j, menor a preocupação com a multicolinearidade e seus efeitos.

Diagnóstico de multicolinearidade

- $\text{VIF}_j = 1/(1 - R_j^2)$ é chamado **fator de inflação da variância** e pode ser utilizado para diagnóstico de multicolinearidade.

- Se as colunas de X forem ortogonais (variáveis não correlacionadas), então $\text{VIF}_j = 1$ para todo j.

- Quanto mais próximos de 1 os valores de VIF_j, menor a preocupação com a multicolinearidade e seus efeitos.

- Uma regra prática, mas não formal, para indicação de multicolinearidade é a identificação de qualquer $\text{VIF}_j > 10$.

Exemplo- Valores de imóveis

- Vamos calcular o valor do VIF para cada uma das cinco variáveis explicativas.

Exemplo- Valores de imóveis

- Vamos calcular o valor do VIF para cada uma das cinco variáveis explicativas.

- Para isso, vamos ajustar um modelo de regressão linear para cada variável explicativa em função das demais e calcular o valor de R^2.

Exemplo- Valores de imóveis

- Vamos calcular o valor do VIF para cada uma das cinco variáveis explicativas.

- Para isso, vamos ajustar um modelo de regressão linear para cada variável explicativa em função das demais e calcular o valor de R^2.

- Calculado R^2, temos condições de efetivamente calcular o valor de VIF.

Exemplo- Valores de imóveis

- Vamos calcular o valor do VIF para cada uma das cinco variáveis explicativas.

- Para isso, vamos ajustar um modelo de regressão linear para cada variável explicativa em função das demais e calcular o valor de R^2.

- Calculado R^2, temos condições de efetivamente calcular o valor de VIF.

- Observe que esta etapa da análise considera apenas os valores das variáveis explicativas, e não da resposta.

Exemplo- Valores de imóveis

- `area`:

$$\widehat{\texttt{area}} = 95.766 - 0.217 \times \texttt{idade} + 0.321 \times \texttt{distancia} + 3.845 \times \texttt{ncomodos} - 0.030 \times \texttt{pcomerc}$$

Exemplo- Valores de imóveis

- `area`:

$$\widehat{\texttt{area}} = 95.766 - 0.217 \times \texttt{idade} + 0.321 \times \texttt{distancia} + 3.845 \times \texttt{ncomodos} - 0.030 \times \texttt{pcomerc}$$

$$R^2_{\texttt{area}} = 0.2951; \qquad \text{VIF}_{\texttt{area}} = \frac{1}{1 - R^2_{\texttt{area}}} = 1.4186$$

Exemplo- Valores de imóveis

- `idade`:

$$\widehat{\texttt{idade}} = 13.128 - 0.014 \times \texttt{area} + 0.043 \times \texttt{distancia} - 0.082 \times \texttt{ncomodos} + 0.057 \times \texttt{pcomerc}$$

Exemplo- Valores de imóveis

- `idade`:

$$\widehat{\texttt{idade}} = 13.128 - 0.014 \times \texttt{area} + 0.043 \times \texttt{distancia} - 0.082 \times \texttt{ncomodos} + 0.057 \times \texttt{pcomerc}$$

$$R^2_{\texttt{idade}} = 0.0133; \qquad \text{VIF}_{\texttt{idade}} = \frac{1}{1 - R^2_{\texttt{idade}}} = 1.0134$$

Exemplo- Valores de imóveis

- `distancia`:

$$\widehat{\texttt{distancia}} = 9.832 + 0.014 \times \texttt{area} + 0.030 \times \texttt{idade} - 0.027 \times \texttt{ncomodos} - 0.109 \times \texttt{pcomerc}$$

Exemplo- Valores de imóveis

- `distancia`:

$$\widehat{\texttt{distancia}} = 9.832 + 0.014 \times \texttt{area} + 0.030 \times \texttt{idade} - 0.027 \times \texttt{ncomodos} - 0.109 \times \texttt{pcomerc}$$

$$R^2_{\texttt{distancia}} = 0.0155; \qquad \text{VIF}_{\texttt{distancia}} = \frac{1}{1 - R^2_{\texttt{distancia}}} = 1.0157$$

Exemplo- Valores de imóveis

- $\texttt{ncomodos}$:

$$\widehat{\texttt{ncomodos}} = -2.311 + 0.074 \times \texttt{area} - 0.024 \times \texttt{idade} - 0.011 \times \texttt{distancia} + 0.006 \times \texttt{pcomerc}$$

Exemplo- Valores de imóveis

- `ncomodos`:

$$\widehat{\texttt{ncomodos}} = -2.311 + 0.074 \times \texttt{area} - 0.024 \times \texttt{idade} - 0.011 \times \texttt{distancia} + 0.006 \times \texttt{pcomerc}$$

$$R^2_{\texttt{ncomodos}} = 0.2914; \qquad \text{VIF}_{\texttt{ncomodos}} = \frac{1}{1 - R^2_{\texttt{ncomodos}}} = 1.4112$$

Exemplo- Valores de imóveis

- `pcomerc`:

$$\widehat{\texttt{pcomerc}} = 10.718 - 0.001 \times \texttt{area} + 0.031 \times \texttt{idade} - 0.087 \times \texttt{distancia} + 0.012 \times \texttt{ncomodos}$$

Exemplo- Valores de imóveis

- $\texttt{pcomerc}$:

$$\widehat{\texttt{pcomerc}} = 10.718 - 0.001 \times \texttt{area} + 0.031 \times \texttt{idade} - 0.087 \times \texttt{distancia} + 0.012 \times \texttt{ncomodos}$$

$$R^2_{\texttt{pcomerc}} = 0.0113; \qquad \text{VIF}_{\texttt{pcomerc}} = \frac{1}{1 - R^2_{\texttt{pcomerc}}} = 1.0114$$

Exemplo- Valores de imóveis

- `pcomerc`:

$$\widehat{\texttt{pcomerc}} = 10.718 - 0.001 \times \texttt{area} + 0.031 \times \texttt{idade} - 0.087 \times \texttt{distancia} + 0.012 \times \texttt{ncomodos}$$

$$R^2_{\texttt{pcomerc}} = 0.0113; \qquad \text{VIF}_{\texttt{pcomerc}} = \frac{1}{1 - R^2_{\texttt{pcomerc}}} = 1.0114$$

- **Conclusão:** Os valores calculados de VIF são todos pequenos (próximos de 1) indicando não haver problemas de multicolinearidade.

Como lidar com a multicolinearidade

- Alguns procedimentos podem ser adotados para contornar o problema da multicolinearidade, dentre eles:

Como lidar com a multicolinearidade

- Alguns procedimentos podem ser adotados para contornar o problema da multicolinearidade, dentre eles:

1. Coleta de dados adicionais: coletar dados em regiões do espaço das variáveis não amostradas (ou amostradas com baixa frequência);

Como lidar com a multicolinearidade

- Alguns procedimentos podem ser adotados para contornar o problema da multicolinearidade, dentre eles:

1. Coleta de dados adicionais: coletar dados em regiões do espaço das variáveis não amostradas (ou amostradas com baixa frequência);

2. Reespecificação do modelo: por exemplo, se as variáveis x_1, x_2 e x_3 exibirem multicolinearidade, pode-se optar por:
   - Substituí-las por alguma função que preserve a informação original mas reduza a colinearidade (ex: $z = (x_1 + x_2 + x_3)/3$ ou $w = x_1 x_2 / x_3$ ou...);
   - Eliminar uma ou mais variáveis pode ser uma alternativa, embora isso possa reduzir o poder preditivo do modelo.

Como lidar com a multicolinearidade

3. Regressão ridge, lasso ou elastic-net- Esses métodos consistem em encontrar um estimador $\hat{\beta}^*$ que seja viciado para β mas com menor variância que $\hat{\boldsymbol{\beta}}$, o estimador de mínimos quadrados.

Como lidar com a multicolinearidade

3. Regressão ridge, lasso ou elastic-net- Esses métodos consistem em encontrar um estimador $\hat{\beta}^*$ que seja viciado para β mas com menor variância que $\hat{\boldsymbol{\beta}}$, o estimador de mínimos quadrados.

4. Regressão por componentes principais - O método de componentes principais permite identificar um conjunto de $q < p$ combinações lineares ortogonais das variáveis regressoras originais que expliquem a maior parcela possível da variação original presente em X.

Como lidar com a multicolinearidade

3. Regressão ridge, lasso ou elastic-net- Esses métodos consistem em encontrar um estimador $\hat{\beta}^*$ que seja viciado para β mas com menor variância que $\hat{\boldsymbol{\beta}}$, o estimador de mínimos quadrados.

4. Regressão por componentes principais - O método de componentes principais permite identificar um conjunto de $q < p$ combinações lineares ortogonais das variáveis regressoras originais que expliquem a maior parcela possível da variação original presente em X.

- Após identificadas as novas variáveis (componentes), as p variáveis originais podem ser substituídas pelos q componentes principais no ajuste do modelo de regressão.

Exercícios

Exercícios

- Resolva os exercícios da lista de exercícios relativa a este módulo, disponibilizada na página da disciplina.