CE310 - Modelos de Regressão Linear

Medidas corretivas

Cesar Augusto Taconeli

14 de maio, 2025

Introdução

Introdução

- Neste módulo da disciplina vamos discutir como remediar problemas de ajuste, relativos à especificação do modelo de regressão linear. Os problemas mais comuns são:

Introdução

- Neste módulo da disciplina vamos discutir como remediar problemas de ajuste, relativos à especificação do modelo de regressão linear. Os problemas mais comuns são:

1. Os erros não têm média zero, não têm variância constante ou são correlacionados;

Introdução

- Neste módulo da disciplina vamos discutir como remediar problemas de ajuste, relativos à especificação do modelo de regressão linear. Os problemas mais comuns são:

i. Os erros não têm média zero, não têm variância constante ou são correlacionados;

ii. Os erros não têm distribuição normal;

Introdução

- Neste módulo da disciplina vamos discutir como remediar problemas de ajuste, relativos à especificação do modelo de regressão linear. Os problemas mais comuns são:

i. Os erros não têm média zero, não têm variância constante ou são correlacionados;

ii. Os erros não têm distribuição normal;

iii. A função de regressão (preditor linear) não está corretamente especificada.

Introdução

- Importante ter em mente que em muitos casos os dados requerem técnicas de modelagem que vão além de uma regressão linear.

Introdução

- Importante ter em mente que em muitos casos os dados requerem técnicas de modelagem que vão além de uma regressão linear.

- Na análise de dados de contagens, por exemplo, a relação entre média e variância não é constante, e uma regressão com resposta Poisson pode ser apropriada.

Introdução

- Importante ter em mente que em muitos casos os dados requerem técnicas de modelagem que vão além de uma regressão linear.

- Na análise de dados de contagens, por exemplo, a relação entre média e variância não é constante, e uma regressão com resposta Poisson pode ser apropriada.

- Dados coletados sequencialmente ao longo do tempo podem ser modelados adequadamente incorporando a correlação temporal, por exemplo através de modelos de séries temporais.

Introdução

- Importante ter em mente que em muitos casos os dados requerem técnicas de modelagem que vão além de uma regressão linear.

- Na análise de dados de contagens, por exemplo, a relação entre média e variância não é constante, e uma regressão com resposta Poisson pode ser apropriada.

- Dados coletados sequencialmente ao longo do tempo podem ser modelados adequadamente incorporando a correlação temporal, por exemplo através de modelos de séries temporais.

- Algumas variáveis apresentam relação não linear que só podem ser bem descritas por modelos não lineares, e assim por diante.

Transformação de variáveis- caso de não normalidade e variância não constante

Transformação de variáveis- não normalidade e variância não constante

- Já discutimos, anteriormente, o uso de transformações para linearizar a relação entre variáveis.

Transformação de variáveis- não normalidade e variância não constante

- Já discutimos, anteriormente, o uso de transformações para linearizar a relação entre variáveis.

- Em determinadas situações, transformar a variável resposta pode estabilizar a variância ou mesmo induzir normalidade.

Transformação de variáveis- não normalidade e variância não constante

- Já discutimos, anteriormente, o uso de transformações para linearizar a relação entre variáveis.

- Em determinadas situações, transformar a variável resposta pode estabilizar a variância ou mesmo induzir normalidade.

- Algumas transformações adequadas para estabilizar a variância dos erros são apresentadas na Tabela 1.

Transformação de variáveis- não normalidade e variância não constante

Tabela 1: Transformações recomendadas para estabilizar a variância

Relação entre σ^2 e μ	Transformação indicada
$\sigma^2 \propto$ cte	$y' = y$ (sem transformação)
$\sigma^2 \propto \mu$	$y' = \sqrt{y}$ (raiz quadrada - dados de contagens - Poisson)
$\sigma^2 \propto \mu(1-\mu)$	$y' = sen^{-1} y$ (arco-seno - dados de proporções - binomial)
$\sigma^2 \propto \mu^2$	$y' = ln(y)$ (log)
$\sigma^2 \propto \mu^3$	$y' = y^{-1/2}$ (raiz inversa)
$\sigma^2 \propto \mu^4$	$y' = y^{-1}$ (inversa)

Método de Box-Cox

- O método de Box-Cox é um procedimento analítico usado para identificar uma transformação para y que induza normalidade e/ou variância constante.

Método de Box-Cox

- O método de Box-Cox é um procedimento analítico usado para identificar uma transformação para y que induza normalidade e/ou variância constante.

- Para este método são consideradas as transformações do tipo potência, ou seja, $y^* = y^\lambda$, sendo λ um parâmetro a ser especificado.

Método de Box-Cox

- O método de Box-Cox é um procedimento analítico usado para identificar uma transformação para y que induza normalidade e/ou variância constante.

- Para este método são consideradas as transformações do tipo potência, ou seja, $y^* = y^\lambda$, sendo λ um parâmetro a ser especificado.

- Para a estimação de λ o usual é utilizar o método de máxima verossimilhança.

Método de Box-Cox

- A família de transformações do tipo potência proposta por Box e Cox é definida por:

$$y^{(\lambda)} = \frac{y^\lambda - 1}{\lambda \dot{y}^{\lambda-1}}, \quad \text{se } \lambda \neq 0,$$

onde $\dot{y} = \ln^{-1}\left[(1/n)\sum_{i=1}^{n} \ln y_i\right]$ é a média geométrica das observações.

Método de Box-Cox

- A família de transformações do tipo potência proposta por Box e Cox é definida por:

$$y^{(\lambda)} = \frac{y^{\lambda} - 1}{\lambda \dot{y}^{\lambda-1}}, \quad \text{se } \lambda \neq 0,$$

onde $\dot{y} = \ln^{-1}\left[(1/n)\sum_{i=1}^{n} \ln y_i\right]$ é a média geométrica das observações.

- Nesta especificação temos que $y^{(\lambda)} \to \log(y)$ quando $\lambda \to 0$, de forma que tomamos $y^{(\lambda)} = \log(y)$ para $\lambda = 0$.

Método de Box-Cox

- A família de transformações do tipo potência proposta por Box e Cox é definida por:

$$y^{(\lambda)} = \frac{y^{\lambda} - 1}{\lambda \dot{y}^{\lambda-1}}, \quad \text{se } \lambda \neq 0,$$

onde $\dot{y} = \ln^{-1}\left[(1/n)\sum_{i=1}^{n} \ln y_i\right]$ é a média geométrica das observações.

- Nesta especificação temos que $y^{(\lambda)} \to \log(y)$ quando $\lambda \to 0$, de forma que tomamos $y^{(\lambda)} = \log(y)$ para $\lambda = 0$.

- A divisão por $\lambda \dot{y}^{\lambda-1}$ tem por objetivo eliminar o efeito de escala, de forma que as somas de quadrados de resíduos para diferentes valores de λ sejam comparáveis.

Método de Box-Cox

- O valor escolhido para λ, denotado por $\hat{\lambda}$, será aquele que maximizar a log-verossimilhança:

$$L(\lambda) = -\frac{n}{2} \log \left[SQ_{\text{Res}}(\lambda) \right],$$

em que $SQ_{\text{Res}}(\lambda)$ a soma de quadrados de resíduos da regressão de $y^{(\lambda)}$ em função das covariáveis.

Método de Box-Cox

- O valor escolhido para λ, denotado por $\hat{\lambda}$, será aquele que maximizar a log-verossimilhança:

$$L(\lambda) = -\frac{n}{2}\log\left[SQ_{\text{Res}}(\lambda)\right],$$

em que $SQ_{\text{Res}}(\lambda)$ a soma de quadrados de resíduos da regressão de $y^{(\lambda)}$ em função das covariáveis.

- Assim, o valor escolhido para o parâmetro λ é aquele que minimiza a soma de quadrados de resíduos.

Método de Box-Cox

- Baseado na teoria da verossimilhança, um intervalo de confiança 100(1-α)% para λ é composto por todo $\lambda = \lambda_0$ tal que:

$$L(\hat{\lambda}) - L(\lambda_0) \leq \frac{1}{2}\chi^2_{1-\alpha,1},$$

em que $\chi^2_{\alpha,1}$ é o quantil $1 - \alpha$ da distribuição chi-quadrado com um grau de liberdade.

Método de Box-Cox

- Baseado na teoria da verossimilhança, um intervalo de confiança 100(1-α)% para λ é composto por todo $\lambda = \lambda_0$ tal que:

$$L(\hat{\lambda}) - L(\lambda_0) \leq \frac{1}{2}\chi^2_{1-\alpha,1},$$

em que $\chi^2_{\alpha,1}$ é o quantil $1-\alpha$ da distribuição chi-quadrado com um grau de liberdade.

- Obtido o intervalo de confiança, pode-se optar por algum valor alternativo pertencente ao intervalo (ao invés de $\hat{\lambda}$), sobretudo se isso proporcionar interpretações mais simples.

Método de Box-Cox

Tabela 2: Transformações de Box-Cox (casos particulares)

λ	Transformação
-2	Inversa quadrática
-1	Inversa
0	Logarítmica
1/2	Raiz quadrada
1	Não transformada
2	Quadrática
3	Cúbica

Método de Box-Cox

- Uma vez encontrada uma transformação apropriada aos dados, a análise deve ser conduzida com base nos dados transformados.

Método de Box-Cox

- Uma vez encontrada uma transformação apropriada aos dados, a análise deve ser conduzida com base nos dados transformados.

- Nem todos os resultados produzidos pelos dados transformados são facilmente convertidos para a escala original.

Método de Box-Cox

- Uma vez encontrada uma transformação apropriada aos dados, a análise deve ser conduzida com base nos dados transformados.

- Nem todos os resultados produzidos pelos dados transformados são facilmente convertidos para a escala original.

- As predições na escala original são facilmente obtidas aplicando a transformação inversa (ex: se $y^{(\lambda)} = \log(y)$ e $\hat{y}^{(\lambda)} = \widehat{\log(y)} = k$, então $\hat{y} = e^k$).

Exemplo- Níveis de ozônio

Exemplo- Níveis de ozônio

- Nesta aplicação vamos considerar a base de dados `ozone` da biblioteca `faraway` do `R`.

Exemplo- Níveis de ozônio

- Nesta aplicação vamos considerar a base de dados `ozone` da biblioteca `faraway` do `R`.

- Os dados se referem a 330 registros diários de variáveis atmosféricas em Los Angeles. As variáveis consideradas são as seguintes:

Exemplo- Níveis de ozônio

- Nesta aplicação vamos considerar a base de dados `ozone` da biblioteca `faraway` do `R`.

- Os dados se referem a 330 registros diários de variáveis atmosféricas em Los Angeles. As variáveis consideradas são as seguintes:
    - `O3`: Concentração de ozônio (variável resposta);

Exemplo- Níveis de ozônio

- Nesta aplicação vamos considerar a base de dados `ozone` da biblioteca `faraway` do `R`.

- Os dados se referem a 330 registros diários de variáveis atmosféricas em Los Angeles. As variáveis consideradas são as seguintes:
  - `O3`: Concentração de ozônio (variável resposta);
  - `temp`: Temperatura;

Exemplo- Níveis de ozônio

- Nesta aplicação vamos considerar a base de dados `ozone` da biblioteca `faraway` do `R`.

- Os dados se referem a 330 registros diários de variáveis atmosféricas em Los Angeles. As variáveis consideradas são as seguintes:
    - `O3`: Concentração de ozônio (variável resposta);
    - `temp`: Temperatura;
    - `humidity`: umidade;

Exemplo- Níveis de ozônio

- Nesta aplicação vamos considerar a base de dados `ozone` da biblioteca `faraway` do `R`.

- Os dados se referem a 330 registros diários de variáveis atmosféricas em Los Angeles. As variáveis consideradas são as seguintes:
    - `O3`: Concentração de ozônio (variável resposta);
    - `temp`: Temperatura;
    - `humidity`: umidade;
    - `ibh`: inversion base height.

Exemplo- Níveis de ozônio

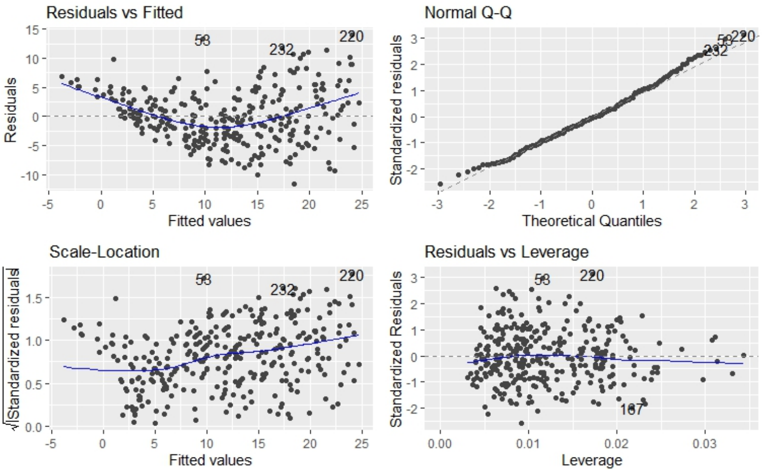

Figura 1: Análise de resíduos para os dados de níveis de ozônio

Exemplo- Níveis de ozônio

- Os gráficos de resíduos apontam desvios das suposições de variância constante e normalidade.

Exemplo- Níveis de ozônio

- Os gráficos de resíduos apontam desvios das suposições de variância constante e normalidade.

- Vamos tentar remediar esses desvios mediante transformação da variável resposta.

Exemplo- Níveis de ozônio

- Os gráficos de resíduos apontam desvios das suposições de variância constante e normalidade.

- Vamos tentar remediar esses desvios mediante transformação da variável resposta.

- Na sequência apresentamos o gráfico do perfil da função de verossimilhança para o parâmetro λ do método de Box-Cox.

Exemplo- Níveis de ozônio

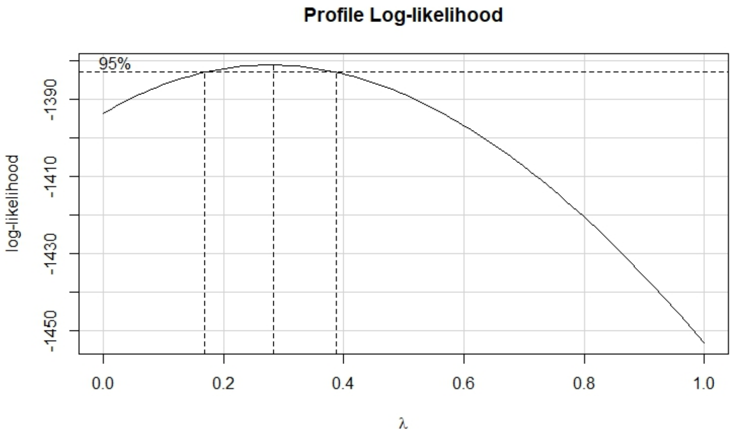

Figura 2: Gráfico do perfil de verossimilhança para o método de Box-Cox

Exemplo- Níveis de ozônio

- Observamos que o gráfico do perfil de verossimilhança indica a necessidade de transformação, dado que $\lambda = 1$ não pertence ao intervalo de confiança (95%) para λ.

Exemplo- Níveis de ozônio

- Observamos que o gráfico do perfil de verossimilhança indica a necessidade de transformação, dado que $\lambda = 1$ não pertence ao intervalo de confiança (95%) para λ.

- A função de verossimilhança assume seu máximo nas proximidades de $\lambda = 1/3$. Vamos adotar esse valor na transformação.

Exemplo- Níveis de ozônio

- Observamos que o gráfico do perfil de verossimilhança indica a necessidade de transformação, dado que $\lambda = 1$ não pertence ao intervalo de confiança (95%) para λ.

- A função de verossimilhança assume seu máximo nas proximidades de $\lambda = 1/3$. Vamos adotar esse valor na transformação.

- Nesta caso, cada valor de `O3` será transformado para $\texttt{O3}^{1/3}$, ou seja, $\sqrt[3]{\texttt{O3}}$. O modelo ajustado com base na variável transformada é dado por:

$$\widehat{\sqrt[3]{\texttt{O3}}} = 0.7625 + 0.02116 \times \texttt{temp} + 0.00488 \times \texttt{humidity} - 0.000077 \times \texttt{ibh}$$

Exemplo- Níveis de ozônio

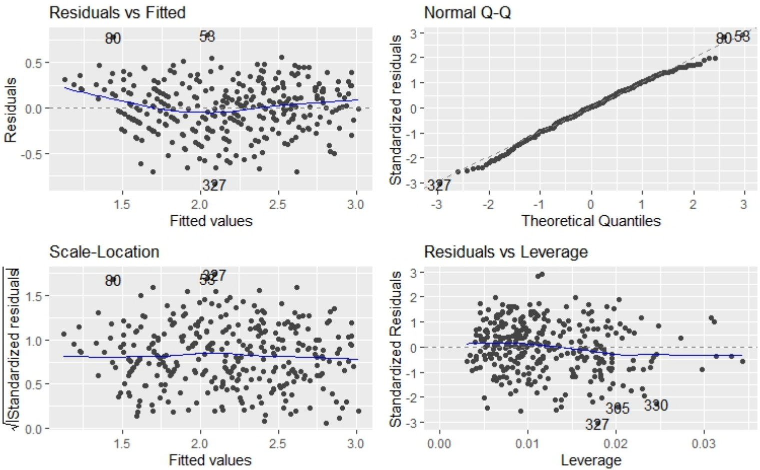

Figura 3: Análise de resíduos para os dados de níveis de ozônio com transformação na resposta

Exemplo- Níveis de ozônio

- O modelo ajustado pode ser expresso, de maneira equivalente, por:

$$\widehat{\texttt{O3}} = (0.7625 + 0.02116 \times \texttt{temp} + 0.00488 \times \texttt{humidity} - 0.000077 \times \texttt{ibh})^3$$

Exemplo- Níveis de ozônio

- O modelo ajustado pode ser expresso, de maneira equivalente, por:

$$\widehat{\texttt{O3}} = (0.7625 + 0.02116 \times \texttt{temp} + 0.00488 \times \texttt{humidity} - 0.000077 \times \texttt{ibh})^3$$

- Os gráficos de resíduos para os dados transformados apontam que os desvios dos pressupostos anteriormente verificados agora estão atenuados.

Exemplo- Níveis de ozônio

- O modelo ajustado pode ser expresso, de maneira equivalente, por:

$$\widehat{\texttt{O3}} = (0.7625 + 0.02116 \times \texttt{temp} + 0.00488 \times \texttt{humidity} - 0.000077 \times \texttt{ibh})^3$$

- Os gráficos de resíduos para os dados transformados apontam que os desvios dos pressupostos anteriormente verificados agora estão atenuados.

- Na aula prática vamos utilizar testes de hipóteses para melhor análise dos pressupostos.

Transformação de variáveis- caso de não linearidade

Transformação de variáveis- caso de não linearidade

- Em alguns casos em a relação entre as variáveis é não linear mas pode ser linearizada mediante alguma transformação adequada.

Transformação de variáveis- caso de não linearidade

- Em alguns casos em a relação entre as variáveis é não linear mas pode ser linearizada mediante alguma transformação adequada.

- Os modelos de regressão resultantes são denominados *modelos intrinsicamente lineares*.

Transformação de variáveis- caso de não linearidade

- Em alguns casos em a relação entre as variáveis é não linear mas pode ser linearizada mediante alguma transformação adequada.

- Os modelos de regressão resultantes são denominados *modelos intrinsicamente lineares*.

- Usar transformações pode remediar o não atendimento de outros pressupostos do modelo (como variância não constante ou ausência de normalidade).

Transformação de variáveis- caso de não linearidade

- Em alguns casos em a relação entre as variáveis é não linear mas pode ser linearizada mediante alguma transformação adequada.

- Os modelos de regressão resultantes são denominados *modelos intrinsicamente lineares*.

- Usar transformações pode remediar o não atendimento de outros pressupostos do modelo (como variância não constante ou ausência de normalidade).

- Neste ponto vamos nos ater à aplicação de transformações com o objetivo de linearizar a relação entre as variáveis.

Transformação de variáveis- caso de não linearidade

- Suponha a seguinte relação não linear entre um par de variáveis x e y:

$$y = \beta_0 e^{\beta_1 x} \epsilon$$

Transformação de variáveis- caso de não linearidade

- Suponha a seguinte relação não linear entre um par de variáveis x e y:

$$y = \beta_0 e^{\beta_1 x} \epsilon$$

- Este modelo pode ser linearizado mediante transformação logarítmica:

$$\log(y) = \log(\beta_0) + \beta_1 x + \log(\epsilon)$$

ou

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

Transformação de variáveis- caso de não linearidade

- Suponha a seguinte relação não linear entre um par de variáveis x e y:

$$y = \beta_0 e^{\beta_1 x} \epsilon$$

- Este modelo pode ser linearizado mediante transformação logarítmica:

$$\log(y) = \log(\beta_0) + \beta_1 x + \log(\epsilon)$$

ou

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

- Neste caso assumimos que ϵ' representa os erros independentes, com distribuição $\mathrm{N}(0, \sigma^2)$.

Transformação de variáveis- caso de não linearidade

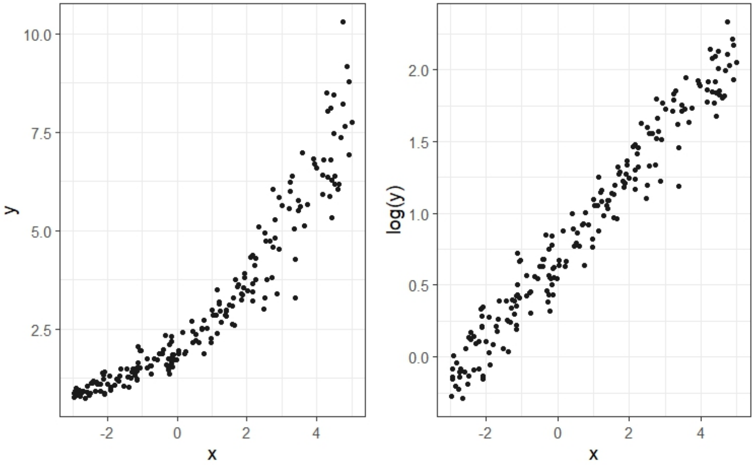

Figura 4: Ilustração de transformação induzindo linearidade

Transformação de variáveis- caso de não linearidade

- De maneira semelhante, se x e y apresentam relação logarítmica, a linearidade pode ser induzida substituindo x por $\log(x)$ na regressão:

$$y = \beta_0 + \beta_1 \log(x) + \epsilon$$

Transformação de variáveis- caso de não linearidade

- De maneira semelhante, se x e y apresentam relação logarítmica, a linearidade pode ser induzida substituindo x por $\log(x)$ na regressão:

$$y = \beta_0 + \beta_1 \log(x) + \epsilon$$

- Outra transformação usualmente considerada para uma (ou ambas) as variáveis é a recíproca, que no caso da transformação de y produz:

$$\frac{1}{y} = \beta_0 + \beta_1 x + \epsilon$$

Transformação de variáveis- caso de não linearidade

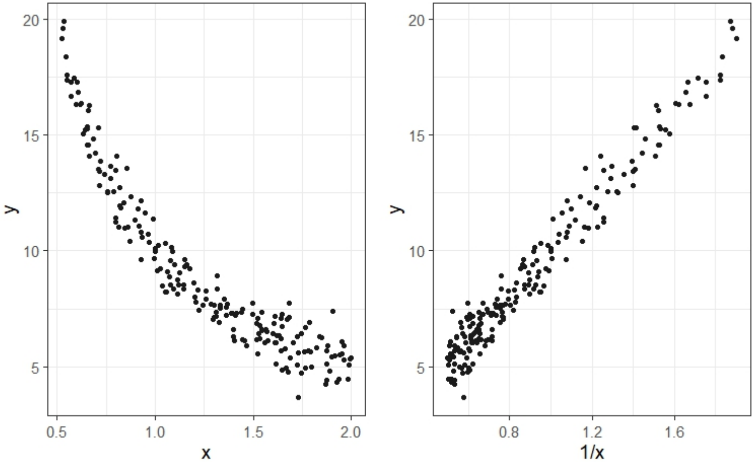

Figura 5: Ilustração de transformação induzindo linearidade (2)

Transformação de variáveis- caso de não linearidade

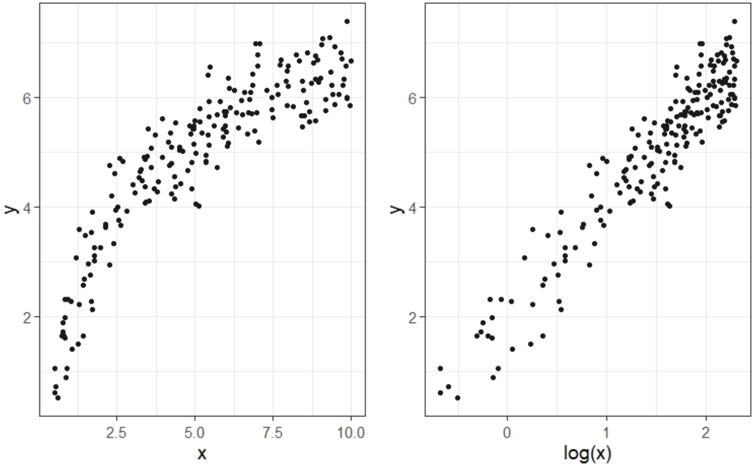

Figura 6: Ilustração de transformação induzindo linearidade (3)

Transformação de variáveis- caso de não linearidade

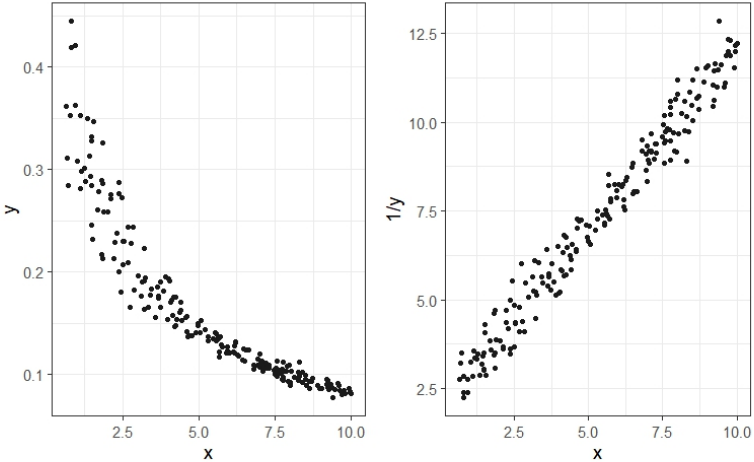

Figura 7: Ilustração de transformação induzindo linearidade (4)

Transformação de variáveis- caso de não linearidade

Tabela 3: Exemplos de funções linearizáveis

Função linearizável	Transformação	Forma linear
(a,b): $y = \beta_0 x^{\beta_1}$	$y' = log(y); x' = log(x)$	$y' = log(\beta_0) + \beta_1 x$
(c,d): $y = \beta_0 e^{\beta_1 x}$	$y' = ln(y)$	$y' = ln\beta_0 + \beta_1 x$
(e,f): $y = \beta_0 + \beta_1 log(x)$	$x' = log(x)$	$y' = \beta_0 + \beta_1 x'$
(g,h): $y = \frac{x}{\beta_0 x - \beta_1}$	$y' = \frac{1}{y}; x' = \frac{1}{x}$	$y' = \beta_0 - \beta_1 x'$

Transformação de variáveis- caso de não linearidade

- Ao usar qualquer uma dessas transformações assumimos que os erros, **na escala transformada**, sejam independentes, normalmente distribuídos com média zero e variância σ^2.

Transformação de variáveis- caso de não linearidade

- Ao usar qualquer uma dessas transformações assumimos que os erros, **na escala transformada**, sejam independentes, normalmente distribuídos com média zero e variância σ^2.

- Quando o método de mínimos quadrados é aplicado após transformação as propriedades dos estimadores, que estudamos anteriormente, valem para os dados transformados e não necessariamente para os dados originais.

Exemplo- Exame de Matemática

Exemplo- Exame de Matemática

- Nesta aplicação as seguintes variáveis são consideradas:

Exemplo- Exame de Matemática

- Nesta aplicação as seguintes variáveis são consideradas:
    - `math`: Desempenho médio do distrito na prova de Matemática (variável resposta);

Exemplo- Exame de Matemática

- Nesta aplicação as seguintes variáveis são consideradas:
  - `math`: Desempenho médio do distrito na prova de Matemática (variável resposta);
  - `income`: Renda média dos habitantes do distrito (variável explicativa).

Exemplo- Exame de Matemática

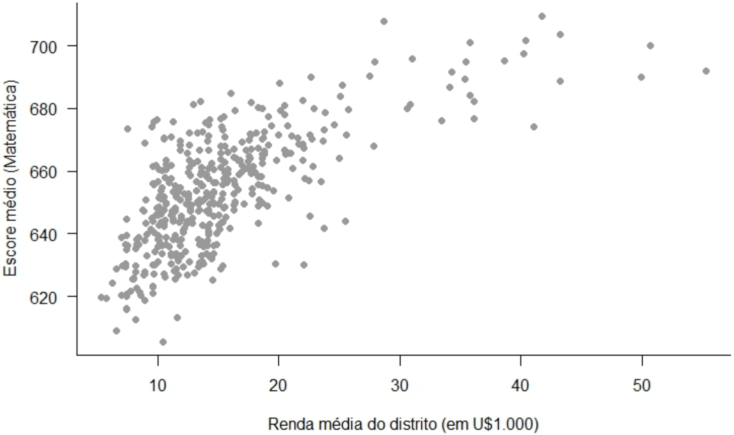

Figura 8: Resultados da prova de Matemática por distritos segundo a renda

Exemplo- Exame de Matemática

- Regressão linear simples

$$\widehat{\texttt{math}} = 625.54 + 1.82\texttt{income}$$

Exemplo- Exame de Matemática

- Regressão linear simples

$$\widehat{\texttt{math}} = 625.54 + 1.82\texttt{income}$$

- Regressão com transformação logarítmica para `income`:

$$\widehat{\texttt{math}} = 561.66 + 34.66\log(\texttt{income})$$

Exemplo- Exame de Matemática

- Regressão linear simples

$$\widehat{\texttt{math}} = 625.54 + 1.82\texttt{income}$$

- Regressão com transformação logarítmica para `income`:

$$\widehat{\texttt{math}} = 561.66 + 34.66\log(\texttt{income})$$

- O escore de Matemática ajustado para um distrito com `income=10` é dado por:

$$\widehat{\texttt{math}} = 561.66 + 34.66\log(10) = 641.47$$

Exemplo- Exame de Matemática

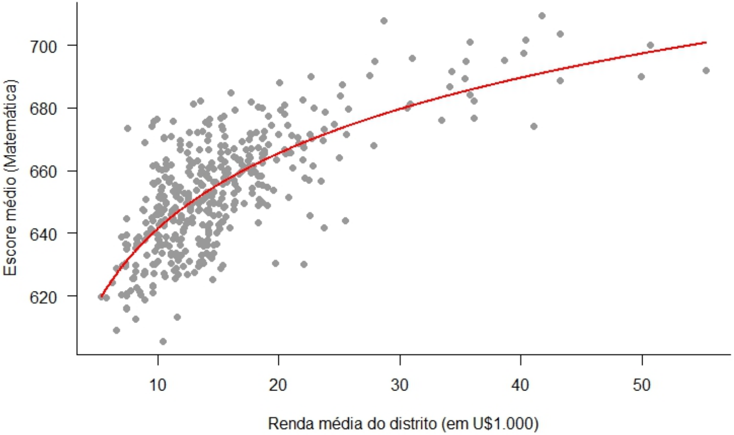

Figura 9: Resultados da prova de Matemática por distritos segundo a renda com regressão ajustada

Exemplo- Energia gerada e velocidade do vento

Exemplo- Energia gerada e velocidade do vento

- Nesta aplicação são consideradas as variáveis:

Exemplo- Energia gerada e velocidade do vento

- Nesta aplicação são consideradas as variáveis:
  - `energy`: energia gerada (variável resposta);

Exemplo- Energia gerada e velocidade do vento

- Nesta aplicação são consideradas as variáveis:
    - `energy`: energia gerada (variável resposta);
    - `wind`: velocidade dos moinhos de vento (variável explicativa).

Exemplo- Energia gerada e velocidade do vento

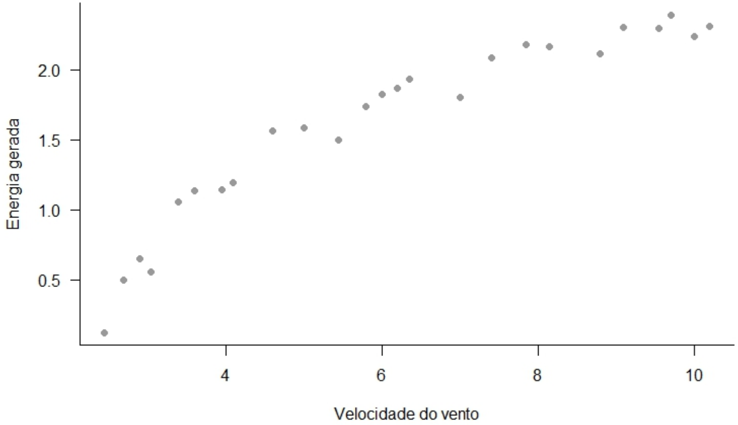

Figura 10: Dados de geração de energia e velocidade de moinhos de vento

Exemplo- Energia gerada e velocidade do vento

- Regressão ajustada com variável **`wind`** transformada (transformação inversa):

$$\widehat{\texttt{energy}} = 2.98 - 6.93 \times \frac{1}{\texttt{wind}}$$

Exemplo- Energia gerada e velocidade do vento

- Regressão ajustada com variável `wind` transformada (transformação inversa):

$$\widehat{\texttt{energy}} = 2.98 - 6.93 \times \frac{1}{\texttt{wind}}$$

- A energia ajustada pelo modelo para `wind=6.50` é dada por:

$$\widehat{\texttt{energy}} = 2.98 - 6.93 \times \frac{1}{6.50} = 1.91$$

Exemplo- Energia gerada e velocidade do vento

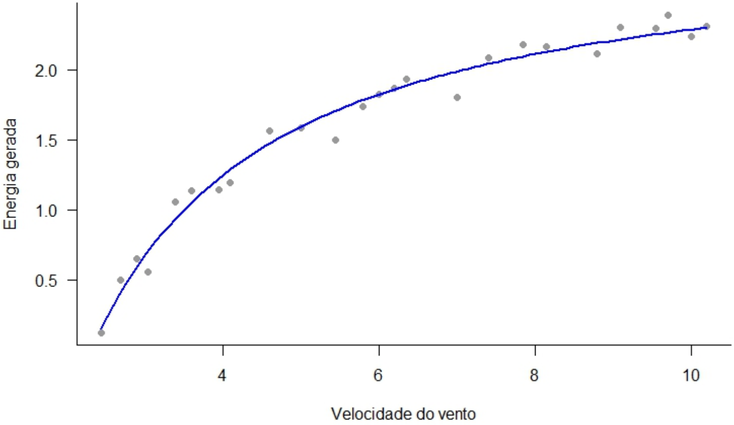

Figura 11: Dados de geração de energia e velocidade de moinhos de vento com regressão ajustada

Exemplo- Pressão de vapor e temperatura da água

Exemplo- Pressão de vapor e temperatura da água

- Nesta aplicação, são consideradas as seguintes variáveis:

Exemplo- Pressão de vapor e temperatura da água

- Nesta aplicação, são consideradas as seguintes variáveis:
  - `p`: Pressão do vapor (variável resposta);

Exemplo- Pressão de vapor e temperatura da água

- Nesta aplicação, são consideradas as seguintes variáveis:
  - `p`: Pressão do vapor (variável resposta);
  - `t`: Temperatura da água (variável explicativa).

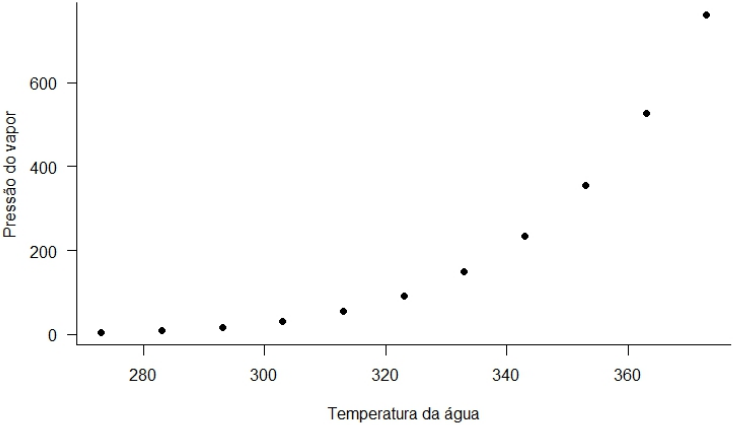

Figura 12: Dados de temperatura de água e pressão de vapor

Exemplo- Pressão de vapor e temperatura da água

- Regressão ajustada com transformação nas duas variáveis:

$$\widehat{\log(\mathtt{p})} = 20.61 - 5201 \times \frac{1}{\mathtt{t}}$$

Exemplo- Pressão de vapor e temperatura da água

- Regressão ajustada com transformação nas duas variáveis:

$$\widehat{\log(\mathtt{p})} = 20.61 - 5201 \times \frac{1}{\mathtt{t}}$$

- De forma equivalente:

$$\hat{p} = \exp\left\{20.61 - 5201 \times \frac{1}{\mathtt{t}}\right\}$$

Exemplo- Pressão de vapor e temperatura da água

- Regressão ajustada com transformação nas duas variáveis:

$$\widehat{\log(\texttt{p})} = 20.61 - 5201 \times \frac{1}{\texttt{t}}$$

- De forma equivalente:

$$\hat{p} = \exp\left\{20.61 - 5201 \times \frac{1}{\texttt{t}}\right\}$$

- Logo, a pressão ajustada para `t=320` é igual a:

$$\hat{p} = \exp\left\{20.61 - 5201 \times \frac{1}{\mathsf{320}}\right\} = 78.01$$

Exemplo- Pressão de vapor e temperatura da água

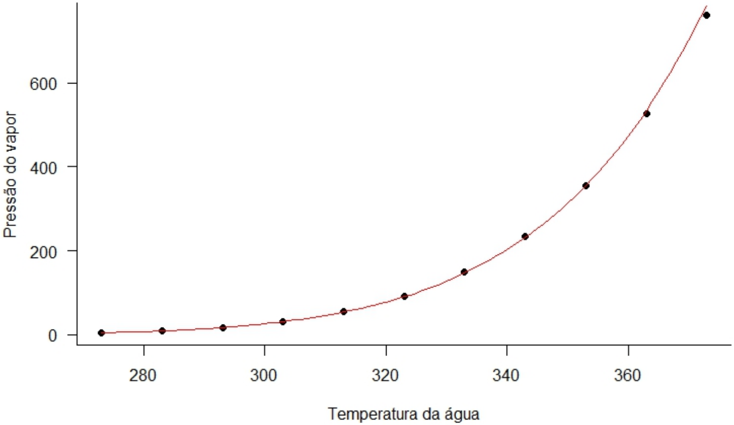

Figura 13: Dados de temperatura de água e pressão de vapor com regressão ajustada

Exercícios adicionais

Exercício- Capacidade pulmonar de jovens

- Nesta aplicação, vamos analisar dados de capacidade pulmonar de 654 jovens, disponíveis na base de dados `lungcap`, que pode ser acessada na biblioteca `GLMsData` do `R`.

Exercício- Capacidade pulmonar de jovens

- Nesta aplicação, vamos analisar dados de capacidade pulmonar de 654 jovens, disponíveis na base de dados `lungcap`, que pode ser acessada na biblioteca `GLMsData` do `R`.

- As variáveis a serem consideradas na análise são as seguintes:

Exercício- Capacidade pulmonar de jovens

- Nesta aplicação, vamos analisar dados de capacidade pulmonar de 654 jovens, disponíveis na base de dados `lungcap`, que pode ser acessada na biblioteca `GLMsData` do `R`.

- As variáveis a serem consideradas na análise são as seguintes:
  - `FEV`: volume expiratório forçado em litros, uma medida de capacidade pulmonar (resposta);

Exercício- Capacidade pulmonar de jovens

- Nesta aplicação, vamos analisar dados de capacidade pulmonar de 654 jovens, disponíveis na base de dados `lungcap`, que pode ser acessada na biblioteca `GLMsData` do `R`.

- As variáveis a serem consideradas na análise são as seguintes:
  - `FEV`: volume expiratório forçado em litros, uma medida de capacidade pulmonar (resposta);
  - `Ht`: altura em polegadas.

Exercício- Capacidade pulmonar de jovens

- Nesta aplicação, vamos analisar dados de capacidade pulmonar de 654 jovens, disponíveis na base de dados `lungcap`, que pode ser acessada na biblioteca `GLMsData` do `R`.

- As variáveis a serem consideradas na análise são as seguintes:
  - `FEV`: volume expiratório forçado em litros, uma medida de capacidade pulmonar (resposta);
  - `Ht`: altura em polegadas.

- Ajuste uma regressão linear simples e, na sequência, procure um melhor modelo transformando a variável resposta e/ou a explicativa.

Exercício- Capacidade pulmonar de jovens

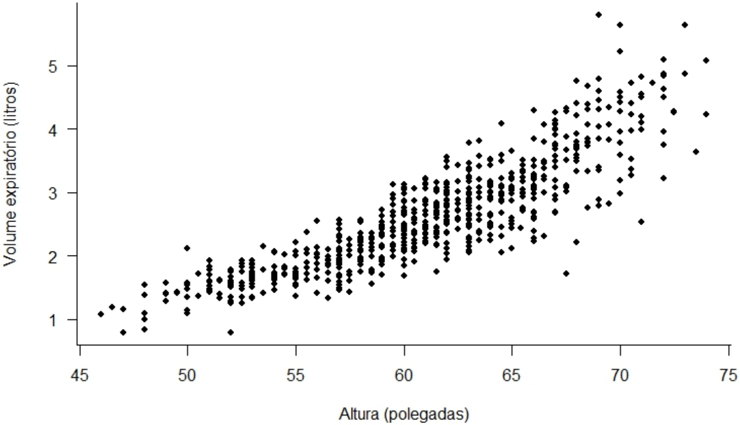

Figura 14: Dados sobre capacidade pulmonar e altura de jovens

Exercício- Saúde dental de crianças

- Nesta aplicação, vamos analisar dados de saúde dental crianças de 90 jovens, disponíveis na base de dados `dental`, que pode ser acessada na biblioteca `GLMsData` do `R`.

Exercício- Saúde dental de crianças

- Nesta aplicação, vamos analisar dados de saúde dental crianças de 90 jovens, disponíveis na base de dados `dental`, que pode ser acessada na biblioteca `GLMsData` do `R`.

- As variáveis a serem consideradas na análise são as seguintes:

Exercício- Saúde dental de crianças

- Nesta aplicação, vamos analisar dados de saúde dental crianças de 90 jovens, disponíveis na base de dados `dental`, que pode ser acessada na biblioteca `GLMsData` do `R`.

- As variáveis a serem consideradas na análise são as seguintes:
  - `DMFT`: estimativa do número médio de dentes cariados, perdidos e obturados (CPOD) na idade 12 anos (resposta);

Exercício- Saúde dental de crianças

- Nesta aplicação, vamos analisar dados de saúde dental crianças de 90 jovens, disponíveis na base de dados `dental`, que pode ser acessada na biblioteca `GLMsData` do `R`.

- As variáveis a serem consideradas na análise são as seguintes:
  - `DMFT`: estimativa do número médio de dentes cariados, perdidos e obturados (CPOD) na idade 12 anos (resposta);
  - `Sugar`: consumo médio de açúcar em quilogramas por pessoa por ano, computado nos últimos cinco anos.

Exercício- Saúde dental de crianças

- Nesta aplicação, vamos analisar dados de saúde dental crianças de 90 jovens, disponíveis na base de dados `dental`, que pode ser acessada na biblioteca `GLMsData` do `R`.

- As variáveis a serem consideradas na análise são as seguintes:
  - `DMFT`: estimativa do número médio de dentes cariados, perdidos e obturados (CPOD) na idade 12 anos (resposta);
  - `Sugar`: consumo médio de açúcar em quilogramas por pessoa por ano, computado nos últimos cinco anos.

- Ajuste uma regressão linear simples e, na sequência, procure um melhor modelo transformando a variável resposta e/ou a explicativa.

Exercício- Saúde dental de crianças

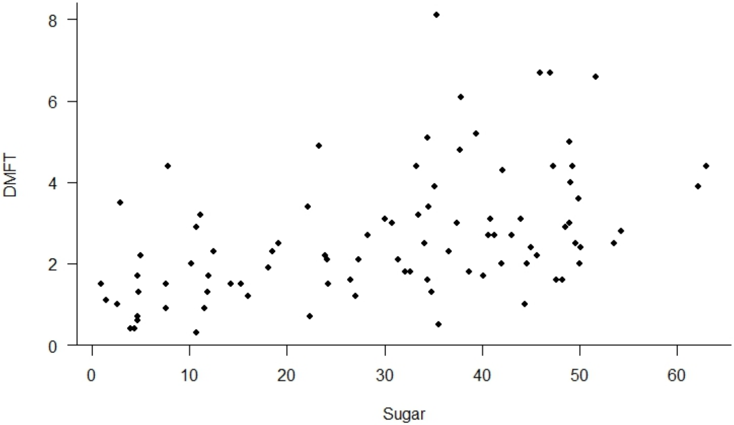

Figura 15: Dados de saúde dental e consumo de açúcar de crianças

Mínimos quadrados ponderados para o caso de variância não constante

Método de mínimos quadrados ponderados

- O método de mínimos quadrados ponderados se aplica caso os erros sejam não correlacionados mas com variâncias diferentes.

Método de mínimos quadrados ponderados

- O método de mínimos quadrados ponderados se aplica caso os erros sejam não correlacionados mas com variâncias diferentes.

- No cenário de erros com variâncias heterogêneas ou autocorrelacionados os estimadores de mínimos quadrados (ordinários) ainda são não viciados, mas não têm variância mínima.

Método de mínimos quadrados ponderados

- O método de mínimos quadrados ponderados se aplica caso os erros sejam não correlacionados mas com variâncias diferentes.

- No cenário de erros com variâncias heterogêneas ou autocorrelacionados os estimadores de mínimos quadrados (ordinários) ainda são não viciados, mas não têm variância mínima.

- Na obtenção dos estimadores por mínimos quadrados ponderados, os componentes da soma de quadrados dos erros são ponderados por pesos ω_i inversamente proporcionais às variâncias dos correspondentes $y_i's$.

Método de mínimos quadrados ponderados

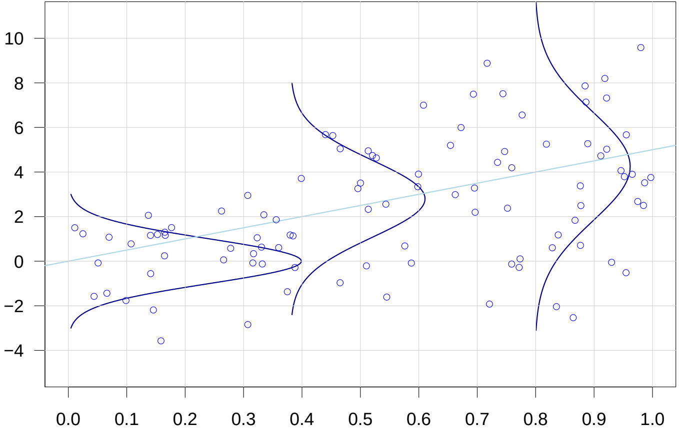

Figura 16: Erros com variância não constante

Método de mínimos quadrados ponderados

- Para o caso da regressão linear simples, por exemplo:

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

e novamente os estimadores de mínimos quadrados são obtidos pela solução do sistema:

$$\frac{\partial S(\beta_0, \beta_1)}{\partial \beta_0} = 0; \quad \frac{\partial S(\beta_0, \beta_1)}{\partial \beta_1} = 0.$$

Método de mínimos quadrados ponderados

- Vamos admitir que a matriz de covariâncias para os erros tenha a seguinte forma:

$$\mathrm{Var}(\epsilon) = \sigma^2 \mathrm{V} = \sigma^2 \begin{bmatrix} \frac{1}{\omega_1} & 0 & \dots & 0 \\ 0 & \frac{1}{\omega_2} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \frac{1}{\omega_n} \end{bmatrix}$$

de forma que $\mathrm{W} = \mathrm{V}^{-1}$ configura a matriz de pesos do método de mínimos quadrados ponderados.

Método de mínimos quadrados ponderados

- Vamos admitir que a matriz de covariâncias para os erros tenha a seguinte forma:

$$\mathrm{Var}(\epsilon) = \sigma^2 \mathrm{V} = \sigma^2 \begin{bmatrix} \frac{1}{\omega_1} & 0 & \dots & 0 \\ 0 & \frac{1}{\omega_2} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \frac{1}{\omega_n} \end{bmatrix}$$

de forma que $\mathrm{W} = \mathrm{V}^{-1}$ configura a matriz de pesos do método de mínimos quadrados ponderados.

- Como V é uma matriz diagonal, W também é uma matriz diagonal com elementos ω_i, $i = 1, 2, ..., n$.

Método de mínimos quadrados ponderados

- O estimador de mínimos quadrados de $\boldsymbol{\beta}$ é $\hat{\boldsymbol{\beta}}$ que é a solução de:

$$(\mathrm{X}'\mathrm{W}\mathrm{X})\hat{\beta} = \mathrm{X}'\mathrm{W}\mathrm{y}$$

Método de mínimos quadrados ponderados

- O estimador de mínimos quadrados de $\boldsymbol{\beta}$ é $\hat{\boldsymbol{\beta}}$ que é a solução de:

$$(\mathrm{X}'\mathrm{W}\mathrm{X})\hat{\beta} = \mathrm{X}'\mathrm{W}\mathrm{y}$$

- Multiplicando-se ambos os lados da igualdade por $(\boldsymbol{X}'\boldsymbol{W}\boldsymbol{X})^{-1}$:

$$\hat{\beta} = (\mathrm{X}'\mathrm{W}\mathrm{X})^{-1}\mathrm{X}'\mathrm{W}\mathrm{y}$$

Método de mínimos quadrados ponderados

- O estimador de mínimos quadrados de $\boldsymbol{\beta}$ é $\hat{\boldsymbol{\beta}}$ que é a solução de:

$$(\mathrm{X}'\mathrm{W}\mathrm{X})\hat{\beta} = \mathrm{X}'\mathrm{W}\mathrm{y}$$

- Multiplicando-se ambos os lados da igualdade por $(\boldsymbol{X}'\boldsymbol{W}\boldsymbol{X})^{-1}$:

$$\hat{\beta} = (\mathrm{X}'\mathrm{W}\mathrm{X})^{-1}\mathrm{X}'\mathrm{W}\mathrm{y}$$

- A matriz de covariâncias de $\hat{\boldsymbol{\beta}}$ fica dada por:

$$\mathrm{Var}(\hat{\beta}) = \sigma^2(\mathrm{X}'\mathrm{W}\mathrm{X})^{-1},$$

que pode ser estimada substituindo σ^2 por $\hat{\sigma}^2 = \frac{\sum_{i=1}^{n}(w_i r_i)^2}{n-p}$, que é a soma de quadrados de resíduos ponderados (r_i é o i-ésimo resíduo).

Mínimos quadrados ponderados

- Na sequência algumas situações práticas que sugerem o uso de ponderação.

Mínimos quadrados ponderados

- Na sequência algumas situações práticas que sugerem o uso de ponderação.

1. Suponha que as observações sejam, na verdade, médias de amostras de m_i observações, ou seja:

$$y_i = \bar{u}_i = \frac{1}{m_i}\sum_{k=1}^{m_i} u_{ik}, \quad i = 1, 2, ..., n.$$

Mínimos quadrados ponderados

- Adicionalmente, vamos considerar que as observações individuais (u_{ik}'s) satisfazem $\text{Var}(u_{ik}|\text{x}_i) = \sigma^2$, constante para todo u_{ik}.

- Adicionalmente, vamos considerar que as observações individuais (u_{ik}'s) satisfazem $\text{Var}(u_{ik}|\text{x}_i) = \sigma^2$, constante para todo u_{ik}.

- Neste caso:

$$\text{Var}(y_i|\text{x}_i) = \frac{\sigma^2}{m_i}, \quad i = 1, 2, ..., n,$$

de tal forma que deveríamos adotar $\omega_i = m_i$.

Exemplo- consumo de água per capita

- Vamos supor que o objetivo seja ajustar um modelo de regressão para o consumo de água por pessoa dos habitantes de uma população.

Exemplo- consumo de água per capita

- Vamos supor que o objetivo seja ajustar um modelo de regressão para o consumo de água por pessoa dos habitantes de uma população.

- Considere que os consumos individuais sejam representados por uma variável aleatória U com variância $\sigma^2 = 2$.

Exemplo- consumo de água per capita

- Vamos supor que o objetivo seja ajustar um modelo de regressão para o consumo de água por pessoa dos habitantes de uma população.

- Considere que os consumos individuais sejam representados por uma variável aleatória U com variância $\sigma^2 = 2$.

- No entanto, na prática se dispõe apenas dos consumos domiciliares, com base nos registros dos medidores de consumo de água.

Exemplo- consumo de água per capita

- Neste caso, vamos considerar como medida de consumo per capita, para cada domicílio, o consumo médio dos moradores, ou seja:

$$y_i = \bar{u}_i = \frac{1}{m_i}\sum_{k=1}^{m_i} u_{ik}, \quad i = 1, 2, ..., n, \quad k = 1, 2, ..., m_i,$$

onde n representa o número de domicílios e m_i o número de habitantes no domicílio i.

Exemplo- consumo de água per capita

- Neste caso, vamos considerar como medida de consumo per capita, para cada domicílio, o consumo médio dos moradores, ou seja:

$$y_i = \bar{u}_i = \frac{1}{m_i}\sum_{k=1}^{m_i} u_{ik}, \quad i = 1, 2, ..., n, \quad k = 1, 2, ..., m_i,$$

onde n representa o número de domicílios e m_i o número de habitantes no domicílio i.

- Como $\text{Var}(U_{ik}) = \sigma^2 = 2$, segue que:

$$\text{Var}(Y_i) = \text{Var}(\bar{U}_i) = \frac{\sigma^2}{m_i} = \frac{2}{m_i}, \quad i = 1, 2, ..., n.$$

Exemplo- consumo de água per capita

Tabela 4: Dados de consumo de água por domicílio e pesos associados

Domicilio	Consumo médio (Y_i)	Habitantes (m_i)	Variância ($\sigma_i^2 = \sigma^2/m_i$)	Peso ($1/\sigma_i^2$)
1	3.5	5	$2/5 = 0.40$	$1/0.40 = 2.50$
2	3.9	2	$2/2 = 1.00$	$1/1.00 = 1.00$
3	2.5	8	$2/8 = 0.25$	$1/0.25 = 4.00$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
999	4.2	1	$2/1 = 2.00$	$1/2.00 = 0.50$
1000	4.0	5	$2/5 = 0.40$	$1/0.40 = 2.50$

Mínimos quadrados ponderados

2. Suponha que o padrão não constante da variância possa ser descrito por alguma função de uma ou mais covariáveis. Como exemplo:

$$\text{Var}(y_i|\boldsymbol{x}_i) = x_{ij}\sigma^2,$$

ou seja, a variância está linearmente relacionada à variável x_j.

Mínimos quadrados ponderados

2. Suponha que o padrão não constante da variância possa ser descrito por alguma função de uma ou mais covariáveis. Como exemplo:

$$\mathrm{Var}(y_i|\boldsymbol{x}_i) = x_{ij}\sigma^2,$$

ou seja, a variância está linearmente relacionada à variável x_j.

- Neste caso, os pesos ficam definidos por $\omega_i = \frac{1}{x_{ij}}$.

Mínimos quadrados ponderados

2. Suponha que o padrão não constante da variância possa ser descrito por alguma função de uma ou mais covariáveis. Como exemplo:

$$\text{Var}(y_i|\boldsymbol{x}_i) = x_{ij}\sigma^2,$$

ou seja, a variância está linearmente relacionada à variável x_j.

- Neste caso, os pesos ficam definidos por $\omega_i = \frac{1}{x_{ij}}$.

- De maneira semelhante, se tivéssemos $\text{Var}(y_i|\text{x}_i) = x_{ij}^2\sigma^2$, poderíamos definir $\omega_i = \frac{1}{x_{ij}^2}$.

Exemplo- Tempos de prova e notas de candicatos de um exame

- Nesta aplicação vamos considerar os dados de 1000 candidatos em um particular exame.

Exemplo- Tempos de prova e notas de candicatos de um exame

- Nesta aplicação vamos considerar os dados de 1000 candidatos em um particular exame.

- Para isso, foram coletadas as notas e os tempos necessários para realização do exame para cada candidato.

Exemplo- Tempos de prova e notas de candicatos de um exame

- Nesta aplicação vamos considerar os dados de 1000 candidatos em um particular exame.
- Para isso, foram coletadas as notas e os tempos necessários para realização do exame para cada candidato.
- O objetivo é ajustar um modelo de regressão que permita explicar a nota em função do tempo de prova.

Exemplo- Tempos de prova e notas de candicatos de um exame

Tabela 5: Dados dos tempos de prova e notas dos candidatos

Candidato	Tempo (min)	Nota
1	44.1	35.0
2	53.8	41.9
3	78.3	47.5
⋮	⋮	⋮
999	83.0	45.5
1000	95.4	67.0

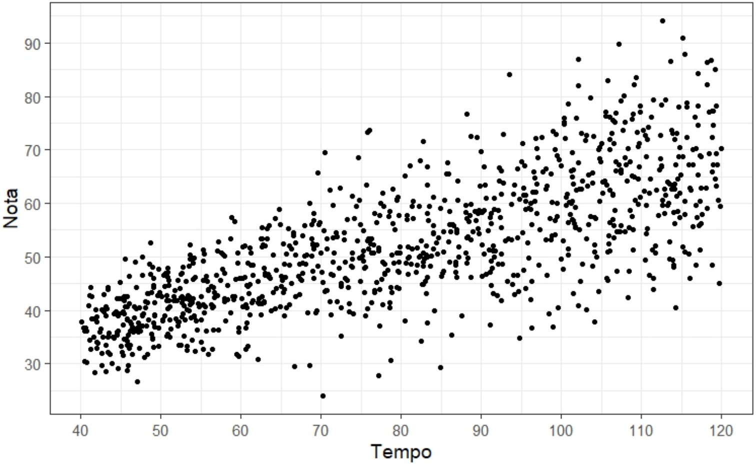

Figura 17: Nota vs tempo de prova para os 1000 candidatos

Exemplo- Tempos de prova e notas de candicatos de um exame

- Embora a relação entre as variáveis seja aparentemente linear, percebe-se que a variância das notas aumenta conforme o tempo de prova.

Exemplo- Tempos de prova e notas de candicatos de um exame

- Embora a relação entre as variáveis seja aparentemente linear, percebe-se que a variância das notas aumenta conforme o tempo de prova.
- Para investigar a relação entre a variância das notas e o tempo de prova, procedemos da seguinte forma:

Exemplo- Tempos de prova e notas de candicatos de um exame

- Embora a relação entre as variáveis seja aparentemente linear, percebe-se que a variância das notas aumenta conforme o tempo de prova.

- Para investigar a relação entre a variância das notas e o tempo de prova, procedemos da seguinte forma:

1. Ordenamos os tempos de prova e dividimos a amostra em faixas conforme os decis desta variável, ou seja:

Exemplo- Tempos de prova e notas de candicatos de um exame

- Embora a relação entre as variáveis seja aparentemente linear, percebe-se que a variância das notas aumenta conforme o tempo de prova.

- Para investigar a relação entre a variância das notas e o tempo de prova, procedemos da seguinte forma:

1. Ordenamos os tempos de prova e dividimos a amostra em faixas conforme os decis desta variável, ou seja:

    - A primeira faixa contém 10% dos candidatos com os (100) menores tempos de prova;

Exemplo- Tempos de prova e notas de candicatos de um exame

- Embora a relação entre as variáveis seja aparentemente linear, percebe-se que a variância das notas aumenta conforme o tempo de prova.

- Para investigar a relação entre a variância das notas e o tempo de prova, procedemos da seguinte forma:

1. Ordenamos os tempos de prova e dividimos a amostra em faixas conforme os decis desta variável, ou seja:
    - A primeira faixa contém 10% dos candidatos com os (100) menores tempos de prova;
    - A segunda faixa contém os 10% seguintes, com os tempos de prova nas posições 101 a 200...

Exemplo- Tempos de prova e notas de candicatos de um exame

- Embora a relação entre as variáveis seja aparentemente linear, percebe-se que a variância das notas aumenta conforme o tempo de prova.

- Para investigar a relação entre a variância das notas e o tempo de prova, procedemos da seguinte forma:

1. Ordenamos os tempos de prova e dividimos a amostra em faixas conforme os decis desta variável, ou seja:
    - A primeira faixa contém 10% dos candidatos com os (100) menores tempos de prova;
    - A segunda faixa contém os 10% seguintes, com os tempos de prova nas posições 101 a 200...
    - A décima faixa contém 10% dos candidatos com os (100) maiores tempos de prova.

Exemplo- Tempos de prova e notas de candicatos de um exame

- Embora a relação entre as variáveis seja aparentemente linear, percebe-se que a variância das notas aumenta conforme o tempo de prova.

- Para investigar a relação entre a variância das notas e o tempo de prova, procedemos da seguinte forma:

1. Ordenamos os tempos de prova e dividimos a amostra em faixas conforme os decis desta variável, ou seja:
    - A primeira faixa contém 10% dos candidatos com os (100) menores tempos de prova;
    - A segunda faixa contém os 10% seguintes, com os tempos de prova nas posições 101 a 200...
    - A décima faixa contém 10% dos candidatos com os (100) maiores tempos de prova.

2. Calculamos, para os dados em cada faixa, a variância das notas dos respectivos candidatos;

Exemplo- Tempos de prova e notas de candicatos de um exame

- Embora a relação entre as variáveis seja aparentemente linear, percebe-se que a variância das notas aumenta conforme o tempo de prova.

- Para investigar a relação entre a variância das notas e o tempo de prova, procedemos da seguinte forma:

1. Ordenamos os tempos de prova e dividimos a amostra em faixas conforme os decis desta variável, ou seja:
    - A primeira faixa contém 10% dos candidatos com os (100) menores tempos de prova;
    - A segunda faixa contém os 10% seguintes, com os tempos de prova nas posições 101 a 200...
    - A décima faixa contém 10% dos candidatos com os (100) maiores tempos de prova.

2. Calculamos, para os dados em cada faixa, a variância das notas dos respectivos candidatos;

3. Plotamos as variâncias calculadas no passo 2 vs os pontos médios das faixas definidas no passo 1.

Exemplo- Tempos de prova e notas de candicatos de um exame

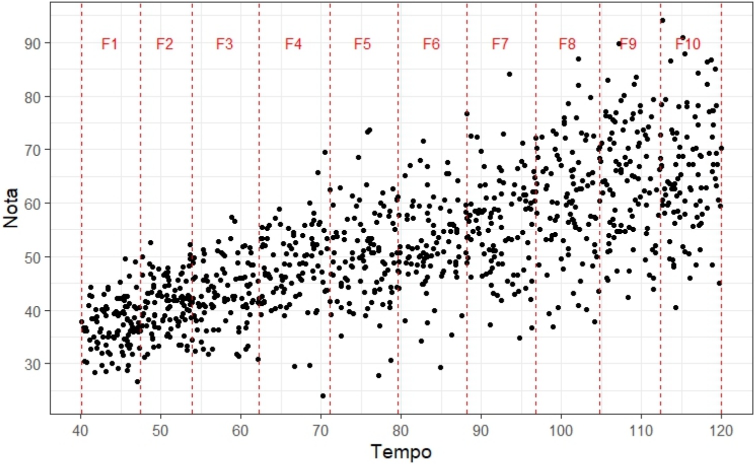

Figura 18: Nota vs tempo de prova para os 1000 candidatos com os dados agrupados em 10 faixas segundo os tempos de prova

Exemplo- Tempos de prova e notas de candicatos de um exame

Tabela 6: Faixas de tempos de prova e variâncias para as notas dos candidatos

Faixa	Tempo de prova		Nota
	Intervalo	Ponto médio	(Variância)
1	(40.1 ; 47.4]	43.75	21.30
2	(47,4 ; 53.9]	50.65	23.21
3	(53.9 ; 62.3]	58.10	25.16
4	(62.3 ; 71.1]	66.70	49.94
5	(71.1 ; 79.6]	75.35	66.47
6	(79.6 ; 88.2]	83.90	57.96
7	(88.2 ; 96.8]	92.50	81.57
8	(96.8 ; 105]	100.9	107.46
9	(105 ; 112]	108.5	108.65
10	(112 ; 120]	116.0	120.80

Exemplo- Tempos de prova e notas de candicatos de um exame

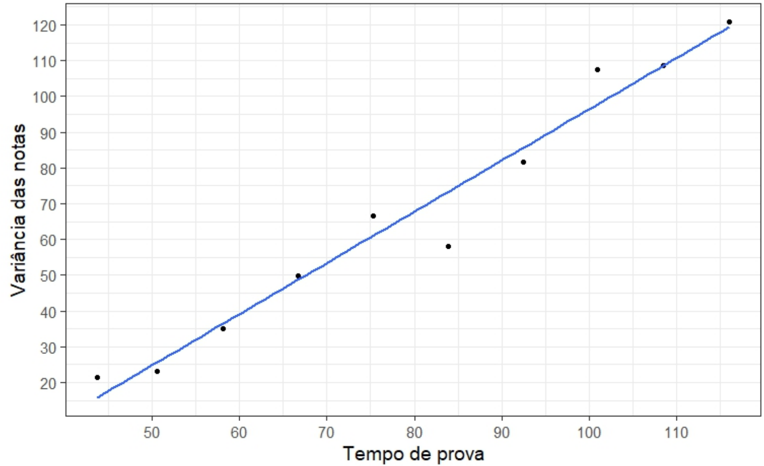

Figura 19: Variância das notas nas dez faixas de tempos de prova

Exemplo- Tempos de prova e notas de candicatos de um exame

- Podemos observar uma relação linear entre as variâncias das notas e os tempos de prova, com a variância aumentando conforme o tempo.

Exemplo- Tempos de prova e notas de candicatos de um exame

- Podemos observar uma relação linear entre as variâncias das notas e os tempos de prova, com a variância aumentando conforme o tempo.

- Com base nisso, vamos ajustar uma regressão linear que descreva a relação entre as variâncias das notas e os tempos de prova. O modelo resultante é o seguinte:

$$\widehat{\texttt{Var(Nota)}} = -46.944 + 1.434 \times \texttt{tempo}$$

Exemplo- Tempos de prova e notas de candicatos de um exame

- Podemos observar uma relação linear entre as variâncias das notas e os tempos de prova, com a variância aumentando conforme o tempo.

- Com base nisso, vamos ajustar uma regressão linear que descreva a relação entre as variâncias das notas e os tempos de prova. O modelo resultante é o seguinte:

$$\widehat{\texttt{Var(Nota)}} = -46.944 + 1.434 \times \texttt{tempo}$$

- Desta forma, podemos usar essa equação para estimação da variância e atribuição dos pesos para cada observação.

Exemplo- Tempos de prova e notas de candicatos de um exame

- Para a primeira observação da base, por exemplo, temos `Tempo=44.1`, tal que:

Exemplo- Tempos de prova e notas de candicatos de um exame

- Para a primeira observação da base, por exemplo, temos `Tempo=44.1`, tal que:

$$\widehat{\texttt{Var(Nota)}}_1 = -46.944 + 1.434 \times 44.1 = 16.2$$

Exemplo- Tempos de prova e notas de candicatos de um exame

- Para a primeira observação da base, por exemplo, temos `Tempo=44.1`, tal que:

$$\widehat{\texttt{Var(Nota)}}_1 = -46.944 + 1.434 \times 44.1 = 16.2$$

- Já para a segunda observação da base `Tempo=53.8`, de forma que:

Exemplo- Tempos de prova e notas de candicatos de um exame

- Para a primeira observação da base, por exemplo, temos `Tempo=44.1`, tal que:

$$\widehat{\texttt{Var(Nota)}}_1 = -46.944 + 1.434 \times 44.1 = 16.2$$

- Já para a segunda observação da base `Tempo=53.8`, de forma que:

$$\widehat{\texttt{Var(Nota)}}_2 = -46.944 + 1.434 \times 53.8 = 30.2,$$

e assim por diante para as demais observações.

Exemplo- Tempos de prova e notas de candicatos de um exame

Tabela 7: Dados dos tempos de prova e notas dos candidatos e pesos para estimação por mínimos quadrados ponderados

Candidato	Tempo (min)	Nota	$\widehat{\texttt{Var(Notas)}}$	Peso $= 1/\widehat{\texttt{Var(Notas)}}$
1	44.1	35.0	16.2	0.0617
2	53.8	41.9	30.2	0.0331
3	78.3	47.5	65.3	0.0153
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
999	83.0	45.5	72.2	0.0138
1000	95.4	67.0	89.8	0.0111

Mínimos quadrados ponderados

3. Em muitos estudos as observações estão sujeitas a erros de medida que podem assumir diferentes distribuições para subconjuntos de observações.

Mínimos quadrados ponderados

3. Em muitos estudos as observações estão sujeitas a erros de medida que podem assumir diferentes distribuições para subconjuntos de observações.

- Como exemplo, considere um experimento em que cada observação é medida por um de três equipamentos disponíveis (A, B e C).

Mínimos quadrados ponderados

3. Em muitos estudos as observações estão sujeitas a erros de medida que podem assumir diferentes distribuições para subconjuntos de observações.

- Como exemplo, considere um experimento em que cada observação é medida por um de três equipamentos disponíveis (A, B e C).

- Considere que os três equipamentos têm diferentes níveis de precisão, sendo as respectivas variâncias dadas por $\sigma_A^2 = 2$, $\sigma_B^2 = 4$ e $\sigma_C^2 = 8$.

Mínimos quadrados ponderados

3. Em muitos estudos as observações estão sujeitas a erros de medida que podem assumir diferentes distribuições para subconjuntos de observações.

- Como exemplo, considere um experimento em que cada observação é medida por um de três equipamentos disponíveis (A, B e C).

- Considere que os três equipamentos têm diferentes níveis de precisão, sendo as respectivas variâncias dadas por $\sigma_A^2 = 2$, $\sigma_B^2 = 4$ e $\sigma_C^2 = 8$.

- Neste caso, os pesos para cada observação seriam determinados pelo inverso das variâncias (eventualmente estimadas) do equipamento que a produziu.

Exemplo- Experimento químico

Tabela 8: Dados ilustrativos de experimento químico com pesos para estimação por mínimos quadrados ponderados

Ensaio	x	y	Equipamento	σ^2_{Equip}	Peso $= 1/\sigma^2_{\text{Equip}}$
1	0	0.8	A	2	1/2
2	0	1.2	B	4	1/4
3	0	0.5	C	8	1/8
4	1	2.0	A	2	1/2
5	1	2.4	B	4	1/4
6	1	2.6	C	8	1/8
7	2	4.8	A	2	1/2
8	2	4.4	B	4	1/4
9	2	5.1	C	8	1/8
10	4	9.5	A	2	1/2
11	4	9.2	B	4	1/4
12	4	8.5	C	8	1/8

Exemplo- Velocidade e distância de frenagem

Exemplo- Velocidade e distância de frenagem

- Neste exemplo de aplicação vamos considerar a base de dados `cars`, disponível no `R`.

Exemplo- Velocidade e distância de frenagem

- Neste exemplo de aplicação vamos considerar a base de dados `cars`, disponível no `R`.

- As variáveis analisadas são as seguintes:

Exemplo- Velocidade e distância de frenagem

- Neste exemplo de aplicação vamos considerar a base de dados `cars`, disponível no `R`.

- As variáveis analisadas são as seguintes:
  - `Dist`: distância percorrida da frenagem até a parada total (resposta);

Exemplo- Velocidade e distância de frenagem

- Neste exemplo de aplicação vamos considerar a base de dados `cars`, disponível no `R`.

- As variáveis analisadas são as seguintes:
  - `Dist`: distância percorrida da frenagem até a parada total (resposta);
  - `Speed`: velocidade do veículo no momento da frenagem.

Exemplo- Velocidade e distância de frenagem

- Neste exemplo de aplicação vamos considerar a base de dados `cars`, disponível no `R`.

- As variáveis analisadas são as seguintes:
  - `Dist`: distância percorrida da frenagem até a parada total (resposta);
  - `Speed`: velocidade do veículo no momento da frenagem.

- Códigos e resultados disponíveis no script `R`.

Exercício- Dados de mandíbulas e idade gestacional de recém-nascidos

- Neste exercício vamos considerar a base de dados `mandible`, disponível na biblioteca `GLMsData` do `R`.

Exercício- Dados de mandíbulas e idade gestacional de recém-nascidos

- Neste exercício vamos considerar a base de dados `mandible`, disponível na biblioteca `GLMsData` do `R`.

- As variáveis a serem analisadas são as seguintes:

Exercício- Dados de mandíbulas e idade gestacional de recém-nascidos

- Neste exercício vamos considerar a base de dados `mandible`, disponível na biblioteca `GLMsData` do `R`.

- As variáveis a serem analisadas são as seguintes:
  - `Length:` tamanho da mandíbula em mm (variável resposta);

Exercício- Dados de mandíbulas e idade gestacional de recém-nascidos

- Neste exercício vamos considerar a base de dados `mandible`, disponível na biblioteca `GLMsData` do `R`.

- As variáveis a serem analisadas são as seguintes:
  - `Length:` tamanho da mandíbula em mm (variável resposta);
  - `Age:` idade gestacional (em semanas).

Exercício- Dados de mandíbulas e idade gestacional de recém-nascidos

- Neste exercício vamos considerar a base de dados `mandible`, disponível na biblioteca `GLMsData` do `R`.

- As variáveis a serem analisadas são as seguintes:
  - `Length:` tamanho da mandíbula em mm (variável resposta);
  - `Age:` idade gestacional (em semanas).

- O objetivo é ajustar um modelo de regressão linear levando em conta a variância não constante dos resíduos, usando:

Exercício- Dados de mandíbulas e idade gestacional de recém-nascidos

- Neste exercício vamos considerar a base de dados `mandible`, disponível na biblioteca `GLMsData` do `R`.

- As variáveis a serem analisadas são as seguintes:
  - `Length:` tamanho da mandíbula em mm (variável resposta);
  - `Age:` idade gestacional (em semanas).

- O objetivo é ajustar um modelo de regressão linear levando em conta a variância não constante dos resíduos, usando:
  - Transformação na resposta (Box-Cox);

Exercício- Dados de mandíbulas e idade gestacional de recém-nascidos

- Neste exercício vamos considerar a base de dados `mandible`, disponível na biblioteca `GLMsData` do `R`.

- As variáveis a serem analisadas são as seguintes:
  - `Length:` tamanho da mandíbula em mm (variável resposta);
  - `Age:` idade gestacional (em semanas).

- O objetivo é ajustar um modelo de regressão linear levando em conta a variância não constante dos resíduos, usando:
  - Transformação na resposta (Box-Cox);
  - Mínimos quadrados ponderados.

Exercício- Dados de peso ao nascer e idade gestacional de recém-nascidos

- Neste exercício vamos considerar a base de dados `gestation`, disponível na biblioteca `GLMsData` do `R`.

Exercício- Dados de peso ao nascer e idade gestacional de recém-nascidos

- Neste exercício vamos considerar a base de dados `gestation`, disponível na biblioteca `GLMsData` do `R`.

- As variáveis a serem analisadas são as seguintes:

Exercício- Dados de peso ao nascer e idade gestacional de recém-nascidos

- Neste exercício vamos considerar a base de dados `gestation`, disponível na biblioteca `GLMsData` do `R`.

- As variáveis a serem analisadas são as seguintes:
    - `Weight:` peso ao nascer em kg (variável resposta);

Exercício- Dados de peso ao nascer e idade gestacional de recém-nascidos

- Neste exercício vamos considerar a base de dados `gestation`, disponível na biblioteca `GLMsData` do `R`.

- As variáveis a serem analisadas são as seguintes:
  - `Weight`: peso ao nascer em kg (variável resposta);
  - `Age`: idade gestacional (em semanas);

Exercício- Dados de peso ao nascer e idade gestacional de recém-nascidos

- Neste exercício vamos considerar a base de dados `gestation`, disponível na biblioteca `GLMsData` do `R`.

- As variáveis a serem analisadas são as seguintes:
  - `Weight`: peso ao nascer em kg (variável resposta);
  - `Age`: idade gestacional (em semanas);
  - `Births`: total de nascimentos.

Exercício- Dados de peso ao nascer e idade gestacional de recém-nascidos

- Neste exercício vamos considerar a base de dados `gestation`, disponível na biblioteca `GLMsData` do `R`.

- As variáveis a serem analisadas são as seguintes:
  - `Weight`: peso ao nascer em kg (variável resposta);
  - `Age`: idade gestacional (em semanas);
  - `Births`: total de nascimentos.

- O objetivo é ajustar um modelo de regressão linear por mínimos quadrados ponderados incorporando como pesos os totais de nascimentos em cada idade gestacional.

Material complementar

Regressão robusta- Estimadores M

Regressão robusta- Estimadores M

- Os estimadores de mínimos quadrados podem ser seriamente afetados se a distribuição dos erros apresentar caudas pesadas.

Regressão robusta- Estimadores M

- Os estimadores de mínimos quadrados podem ser seriamente afetados se a distribuição dos erros apresentar caudas pesadas.

- Em particular, os estimadores de mínimos quadrados são vulneráveis a outliers e a pontos de alavanca.

Regressão robusta- Estimadores M

- Os estimadores de mínimos quadrados podem ser seriamente afetados se a distribuição dos erros apresentar caudas pesadas.

- Em particular, os estimadores de mínimos quadrados são vulneráveis a outliers e a pontos de alavanca.

- Se as observações atípicas forem decorrentes de erros no processo de coleta, registro ou tabulação dos dados, deverão ser corrigidas ou excluídas da análise.

Regressão robusta- Estimadores M

- Os estimadores de mínimos quadrados podem ser seriamente afetados se a distribuição dos erros apresentar caudas pesadas.

- Em particular, os estimadores de mínimos quadrados são vulneráveis a outliers e a pontos de alavanca.

- Se as observações atípicas forem decorrentes de erros no processo de coleta, registro ou tabulação dos dados, deverão ser corrigidas ou excluídas da análise.

- Caso contrário, a utilização de métodos robustos de regressão é indicada.

Regressão robusta- Estimadores M

- Os estimadores M configuram uma classe de estimadores, obtidos mediante minimização de uma família de funções objetivas dos erros, sendo o método de mínimos quadrados um caso particular.

Regressão robusta- Estimadores M

- Os estimadores M configuram uma classe de estimadores, obtidos mediante minimização de uma família de funções objetivas dos erros, sendo o método de mínimos quadrados um caso particular.

- Estimadores M generalizam a ideia de mínimos quadrados ao identificar $\hat{\beta}$ que minimiza:

$$\sum_{i=1}^{n} \rho(\epsilon_i) = \sum_{i=1}^{n} \rho(y_i - \boldsymbol{x}_i' \boldsymbol{\beta}). \tag{1}$$

Regressão robusta- Estimadores M

- Diferenciando a função objetiva (1) com relação a β e igualando a 0, obtemos:

$$\sum_{i=1}^{n} \rho'(y_i - \mathrm{x}'_i\beta)\mathrm{x}'_i = 0,$$

que é um sistema de p equações nos p componentes de $\boldsymbol{\beta}$.

Regressão robusta- Estimadores M

- Diferenciando a função objetiva (1) com relação a β e igualando a 0, obtemos:

$$\sum_{i=1}^{n} \rho'(y_i - \mathrm{x'}_i\beta)\mathrm{x'}_i = 0,$$

que é um sistema de p equações nos p componentes de $\boldsymbol{\beta}$.

- Tomando $\epsilon_i = y_i - \mathrm{x}'_i\beta$, o mesmo sistema pode ser escrito de forma equivalente:

$$\sum_{i=1}^{n} \frac{\rho'(\epsilon_i)}{\epsilon_i}(y_i - \mathrm{x}'_i\boldsymbol{\beta})\mathrm{x}'_i = \sum_{i=1}^{n} \omega_i(y_i - \mathrm{x}'_i\beta)\mathrm{x}'_i = 0,$$

em que os termos $\omega_i = \rho'(\epsilon_i)/\epsilon_i$ atuam como pesos na obtenção dos estimadores de $\boldsymbol{\beta}$.

Regressão robusta- Estimadores M

- Assim, os estimadores M correspondem aos estimadores de mínimos quadrados ponderados de $\boldsymbol{\beta}$ com pesos definidos por $\omega_i = \rho'(\epsilon_i)/\epsilon_i = \psi(\epsilon_i)/\epsilon_i$.

Regressão robusta- Estimadores M

- Assim, os estimadores M correspondem aos estimadores de mínimos quadrados ponderados de $\boldsymbol{\beta}$ com pesos definidos por $\omega_i = \rho'(\epsilon_i)/\epsilon_i = \psi(\epsilon_i)/\epsilon_i$.

- A Tabela 9 apresenta algumas escolhas usuais para $\rho(\epsilon)$ e as correspondentes funções peso.

Tabela 9: Função objetiva e função peso para estimadores M

Estimador	$\rho(\epsilon)$	$\omega(\epsilon)$	
Least squares	ϵ^2	1	
Least absolute values	$\lvert\epsilon\rvert$	$1/\lvert\epsilon\rvert$	para $\epsilon \neq 0$
Huber	$\frac{\epsilon^2}{2}$	1	para $\lvert\epsilon\rvert \leq k$
	$k\lvert\epsilon\rvert - \frac{k^2}{2}$	$k/\lvert\epsilon\rvert$	para $\lvert\epsilon\rvert > k$
Biweight	$\frac{k^2}{6}\left\{1 - \left[1 - \left(\frac{\epsilon}{k}\right)^2\right]^3\right\}$	$\left[1 - \left(\frac{\epsilon}{k}\right)^2\right]^2$	para $\lvert\epsilon\rvert \leq k$
	$\frac{k^2}{6}$	0	para $\lvert\epsilon\rvert > k$

Regressão robusta- Estimadores M

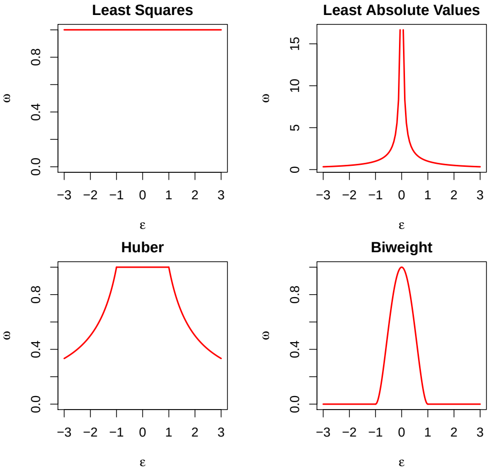

Figura 20: Função peso para diferentes tipos de estimadores M. Para os estimadores Huber e Biweight foi fixado $k = 1$

Regressão robusta- Estimadores M

- No processo de estimação os pesos dependem dos resíduos, os $\beta's$ estimados dependem dos pesos e os resíduos dependem dos $\beta's$ estimados.

Regressão robusta- Estimadores M

- No processo de estimação os pesos dependem dos resíduos, os $\beta's$ estimados dependem dos pesos e os resíduos dependem dos $\beta's$ estimados.

- Dessa forma, o processo de estimação baseia-se num algoritmo de mínimos quadrados ponderados iterativamente, definido pelos seguintes passos:

Regressão robusta- Estimadores M

1. Escolha estimativas iniciais para β ($\beta^{(0)}$) e calcule os resíduos, $\epsilon_i^{(0)} = y_i - \mathrm{x_i'}\beta^{(0)}$, e os pesos, $\omega_i^{(0)} = \omega(\epsilon_i^{(0)})$;

Regressão robusta- Estimadores M

1. Escolha estimativas iniciais para β ($\beta^{(0)}$) e calcule os resíduos, $\epsilon_i^{(0)} = y_i - \mathrm{x_i'}\beta^{(0)}$, e os pesos, $\omega_i^{(0)} = \omega(\epsilon_i^{(0)})$;

2. Na iteração l do algoritmo, obtenha $\hat{\beta}^{(l)}$ minimizando a soma de quadrados ponderada $\sum_{i=1}^{n} \omega_i^{(l-1)} \epsilon_i^{2^{(l-1)}}$:

$$\hat{\beta}^{(l)} = (\mathrm{X'W^{(l-1)}X})^{-1}\mathrm{X'W^{(l-1)}y},$$

onde $\mathrm{X}_{n\times p}$ é a matriz do modelo e $\mathrm{W^{(l-1)}}_{n\times n}$ é a matriz diagonal com elementos $\omega_i^{(l-1)}$.

Regressão robusta- Estimadores M

1. Escolha estimativas iniciais para β ($\beta^{(0)}$) e calcule os resíduos, $\epsilon_i^{(0)} = y_i - \mathrm{x_i'}\beta^{(0)}$, e os pesos, $\omega_i^{(0)} = \omega(\epsilon_i^{(0)})$;

2. Na iteração l do algoritmo, obtenha $\hat{\beta}^{(l)}$ minimizando a soma de quadrados ponderada $\sum_{i=1}^{n} \omega_i^{(l-1)} \epsilon_i^{2^{(l-1)}}$:

$$\hat{\beta}^{(l)} = (\mathrm{X'W^{(l-1)}X})^{-1}\mathrm{X'W^{(l-1)}y},$$

onde $\mathrm{X}_{n\times p}$ é a matriz do modelo e $\mathrm{W^{(l-1)}}_{n\times n}$ é a matriz diagonal com elementos $\omega_i^{(l-1)}$.

3. Os passos 2 e 3 são repetidos até que $\hat{\beta}^{(l)} - \hat{\beta}^{(l-1)}$ seja suficientemente próxima de zero.

Regressão robusta- Estimadores M

- A matriz de covariância assintótica de $\hat{\boldsymbol{\beta}}$ fica dada por:

$$\text{Var}(\hat{\boldsymbol{\beta}}) = \frac{E(\rho'^2)}{[E(\rho')]^2}(\mathrm{X}'\mathrm{X})^{-1}.$$

- A matriz de covariância assintótica de $\hat{\boldsymbol{\beta}}$ fica dada por:

$$\mathrm{Var}(\hat{\boldsymbol{\beta}}) = \frac{E(\rho'^2)}{[E(\rho')]^2}(\mathrm{X}'\mathrm{X})^{-1}.$$

- A matriz de covariância assintótica estimada é obtida substituindo $E(\rho'^2)$ por $\sum_{i=1}^{n}[\rho'(r_i)]^2/n$ e $[E(\rho')]^2$ por $[\sum_{i=1}^{n}\rho'(r_i)/n]^2$.

Exemplo- Gastos com apostas

- Nesta aplicação vamos retomar a base de dados `teengamb` da biblioteca `faraway` do `R`, com o comportamento de 47 apostadores jovens. As variáveis são as seguintes:

Exemplo- Gastos com apostas

- Nesta aplicação vamos retomar a base de dados `teengamb` da biblioteca `faraway` do `R`, com o comportamento de 47 apostadores jovens. As variáveis são as seguintes:
- `sex`: 0=masculino, 1=feminino;

Exemplo- Gastos com apostas

- Nesta aplicação vamos retomar a base de dados `teengamb` da biblioteca `faraway` do `R`, com o comportamento 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

- Nesta aplicação vamos retomar a base de dados `teengamb` da biblioteca `faraway` do `R`, com o comportamento 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

- Nesta aplicação vamos retomar a base de dados `teengamb` da biblioteca `faraway` do `R`, com o comportamento 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

- Nesta aplicação vamos retomar a base de dados `teengamb` da biblioteca `faraway` do `R`, com o comportamento 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

- Nesta aplicação vamos retomar a base de dados `teengamb` da biblioteca `faraway` do `R`, com o comportamento 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).
- Códigos R e discussão das análises disponíveis nos scripts disponibilizados na página da disciplina.

Regressão robusta- Least trimmed squares

Regressão robusta- Least trimmed squares

- Em algumas situações, a relação entre a variável resposta e as explicativas pode ser *contaminada* por uma pequena parcela de observações.

Regressão robusta- Least trimmed squares

- Em algumas situações, a relação entre a variável resposta e as explicativas pode ser *contaminada* por uma pequena parcela de observações.

- O método least trimmed squares consiste na obtenção das estimativas dos $\beta's$ com base num subconjunto ótimo de $n' < n$ observações, que produzem os menores resíduos (as demais $n - n'$ não são utilizadas no ajuste).

Regressão robusta- Least trimmed squares

- Em algumas situações, a relação entre a variável resposta e as explicativas pode ser *contaminada* por uma pequena parcela de observações.

- O método least trimmed squares consiste na obtenção das estimativas dos $\beta's$ com base num subconjunto ótimo de $n' < n$ observações, que produzem os menores resíduos (as demais $n - n'$ não são utilizadas no ajuste).

- Para motivar o uso de least trimmed squares, na sequência são apresentados dados sobre o número de ligações telefônicas realizadas na Bélgica (em milhões) no período de 1950 a 1973.

Regressão robusta- Least trimmed squares

- Em algumas situações, a relação entre a variável resposta e as explicativas pode ser *contaminada* por uma pequena parcela de observações.

- O método least trimmed squares consiste na obtenção das estimativas dos $\beta's$ com base num subconjunto ótimo de $n' < n$ observações, que produzem os menores resíduos (as demais $n - n'$ não são utilizadas no ajuste).

- Para motivar o uso de least trimmed squares, na sequência são apresentados dados sobre o número de ligações telefônicas realizadas na Bélgica (em milhões) no período de 1950 a 1973.

- Observe os resultados atípicos (demasiadamente altos) registrados no período de 1965 a 1970.

Regressão robusta- Least trimmed squares

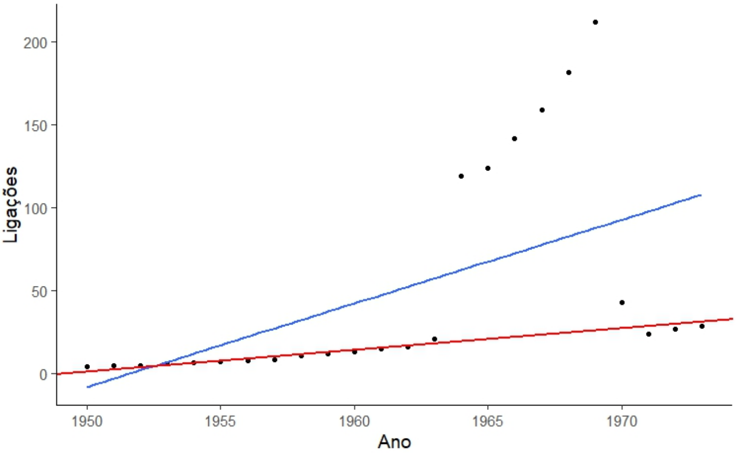

Figura 21: Ligações telefônicas anuais na Bélgica no período de 1950 a 1973 com ajuste de regressão linear por mínimos quadrados (azul) e least trimmed squares (vermelho)

Regressão robusta- Least trimmed squares

- Fica evidente a diferença entre os modelos ajustados com a base completa e com a remoção dos dados atípicos.

Regressão robusta- Least trimmed squares

- Fica evidente a diferença entre os modelos ajustados com a base completa e com a remoção dos dados atípicos.

- O modelo ajustado por mínimos quadrados claramente sofre de falta de ajuste, não permitindo explicar a relação entre as variáveis em nenhum dos períodos.

Regressão robusta- Least trimmed squares

- Fica evidente a diferença entre os modelos ajustados com a base completa e com a remoção dos dados atípicos.

- O modelo ajustado por mínimos quadrados claramente sofre de falta de ajuste, não permitindo explicar a relação entre as variáveis em nenhum dos períodos.

- Já o modelo ajustado por least trimmed squares (LTS) é robusto com relação aos pontos atípicos, por não utilizá-los na estimação dos parâmetros de regressão.

Regressão robusta- Least trimmed squares

- No contexto de regressão linear, os estimadores dos β's por LTS são aqueles tais que:

$$S = \sum_{i=1}^{n'} r_{(i)}^2(\beta)$$

é mínima, em que $r_{(i)}^2(\beta)$ representa o i-ésimo menor resíduo quadrático.

Regressão robusta- Least trimmed squares

- A obtenção dos estimadores via LTS pode se dar:

Regressão robusta- Least trimmed squares

- A obtenção dos estimadores via LTS pode se dar:
  - Avaliando as soluções para todas as $\binom{n}{n'}$ sub amostras (computacionalmente inviável para grandes amostras);

Regressão robusta- Least trimmed squares

- A obtenção dos estimadores via LTS pode se dar:
  - Avaliando as soluções para todas as $\binom{n}{n'}$ sub amostras (computacionalmente inviável para grandes amostras);
  - Usando métodos de otimização que permitem encontrar uma solução (sub) ótima mediante menor número de avaliações.

Regressão robusta- Least trimmed squares

- A escolha de n' é um ponto crítico do método, tal que:

Regressão robusta- Least trimmed squares

- A escolha de n' é um ponto crítico do método, tal que:
  - Essa escolha deve satisfazer $\frac{n}{2} < n' \leq n$;

Regressão robusta- Least trimmed squares

- A escolha de n' é um ponto crítico do método, tal que:
    - Essa escolha deve satisfazer $\frac{n}{2} < n' \leq n$;
    - Uma escolha usual para n' é $\lfloor n/2 \rfloor + \lfloor (p+1)/2 \rfloor$, onde $\lfloor x \rfloor$ representa o maior inteiro menor ou igual a x.

Regressão robusta- Least trimmed squares

- A escolha de n' é um ponto crítico do método, tal que:
    - Essa escolha deve satisfazer $\frac{n}{2} < n' \leq n$;
    - Uma escolha usual para n' é $\lfloor n/2 \rfloor + \lfloor (p+1)/2 \rfloor$, onde $\lfloor x \rfloor$ representa o maior inteiro menor ou igual a x.
    - Se tomarmos $n = n'$, os estimadores de LTS serão equivalentes aos de mínimos quadrados ordinários.

Regressão robusta- Least trimmed squares

- A escolha de n' é um ponto crítico do método, tal que:
  - Essa escolha deve satisfazer $\frac{n}{2} < n' \leq n$;
  - Uma escolha usual para n' é $\lfloor n/2 \rfloor + \lfloor (p+1)/2 \rfloor$, onde $\lfloor x \rfloor$ representa o maior inteiro menor ou igual a x.
  - Se tomarmos $n = n'$, os estimadores de LTS serão equivalentes aos de mínimos quadrados ordinários.

- Na prática, diferentes valores de n' podem ser testados e as soluções obtidas comparadas.

Exemplo- Temperatura e luminosidade de estrelas

Exemplo- Temperatura e luminosidade de estrelas

- Nesta aplicação vamos considerar a base de dados `star` da biblioteca `faraway`.

Exemplo- Temperatura e luminosidade de estrelas

- Nesta aplicação vamos considerar a base de dados `star` da biblioteca `faraway`.

- Os dados referem-se a 47 estrelas no aglomerado estelar CYG OB1, que está na direção de Cygnus. As variáveis são as seguintes:

Exemplo- Temperatura e luminosidade de estrelas

- Nesta aplicação vamos considerar a base de dados `star` da biblioteca `faraway`.

- Os dados referem-se a 47 estrelas no aglomerado estelar CYG OB1, que está na direção de Cygnus. As variáveis são as seguintes:
  - `light`: log intensidade da luz (resposta);

Exemplo- Temperatura e luminosidade de estrelas

- Nesta aplicação vamos considerar a base de dados `star` da biblioteca `faraway`.

- Os dados referem-se a 47 estrelas no aglomerado estelar CYG OB1, que está na direção de Cygnus. As variáveis são as seguintes:
  - `light`: log intensidade da luz (resposta);
  - `temp`: log temperatura da superfície.

Exemplo- Temperatura e luminosidade de estrelas

- Nesta aplicação vamos considerar a base de dados `star` da biblioteca `faraway`.

- Os dados referem-se a 47 estrelas no aglomerado estelar CYG OB1, que está na direção de Cygnus. As variáveis são as seguintes:
  - `light`: log intensidade da luz (resposta);
  - `temp`: log temperatura da superfície.

- Para efeito de comparação, vamos ajustar três modelos, usando:

Exemplo- Temperatura e luminosidade de estrelas

- Nesta aplicação vamos considerar a base de dados `star` da biblioteca `faraway`.

- Os dados referem-se a 47 estrelas no aglomerado estelar CYG OB1, que está na direção de Cygnus. As variáveis são as seguintes:
  - `light`: log intensidade da luz (resposta);
  - `temp`: log temperatura da superfície.

- Para efeito de comparação, vamos ajustar três modelos, usando:
  - Mínimos quadrados ordinários;

Exemplo- Temperatura e luminosidade de estrelas

- Nesta aplicação vamos considerar a base de dados `star` da biblioteca `faraway`.

- Os dados referem-se a 47 estrelas no aglomerado estelar CYG OB1, que está na direção de Cygnus. As variáveis são as seguintes:
  - `light`: log intensidade da luz (resposta);
  - `temp`: log temperatura da superfície.

- Para efeito de comparação, vamos ajustar três modelos, usando:
  - Mínimos quadrados ordinários;
  - Estimadores M;

Exemplo- Temperatura e luminosidade de estrelas

- Nesta aplicação vamos considerar a base de dados `star` da biblioteca `faraway`.

- Os dados referem-se a 47 estrelas no aglomerado estelar CYG OB1, que está na direção de Cygnus. As variáveis são as seguintes:
  - `light`: log intensidade da luz (resposta);
  - `temp`: log temperatura da superfície.

- Para efeito de comparação, vamos ajustar três modelos, usando:
  - Mínimos quadrados ordinários;
  - Estimadores M;
  - Least trimmed squares.

Exemplo- Temperatura e luminosidade de estrelas

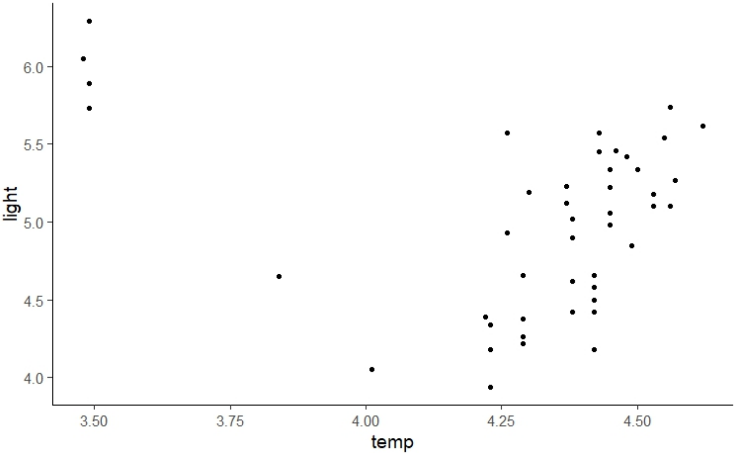

Figura 22: Dados sobre intensidade de luz e temperatura de estrelas

Exemplo- Temperatura e luminosidade de estrelas

- Modelo ajustado por mínimos quadrados ordinários:

$$\widehat{\texttt{light}} = 6.793 - 0.413 \times \texttt{temp}$$

Exemplo- Temperatura e luminosidade de estrelas

- Modelo ajustado por mínimos quadrados ordinários:

$$\widehat{\texttt{light}} = 6.793 - 0.413 \times \texttt{temp}$$

- Modelo ajustado usando estimadores M:

$$\widehat{\texttt{light}} = 6.866 - 0.429 \times \texttt{temp}$$

Exemplo- Temperatura e luminosidade de estrelas

- Modelo ajustado por mínimos quadrados ordinários:

$$\widehat{\texttt{light}} = 6.793 - 0.413 \times \texttt{temp}$$

- Modelo ajustado usando estimadores M:

$$\widehat{\texttt{light}} = 6.866 - 0.429 \times \texttt{temp}$$

- Modelo ajustado por least trimmed squares:

$$\widehat{\texttt{light}} = -8.500 + 3.046 \times \texttt{temp}$$

Exemplo- Temperatura e luminosidade de estrelas

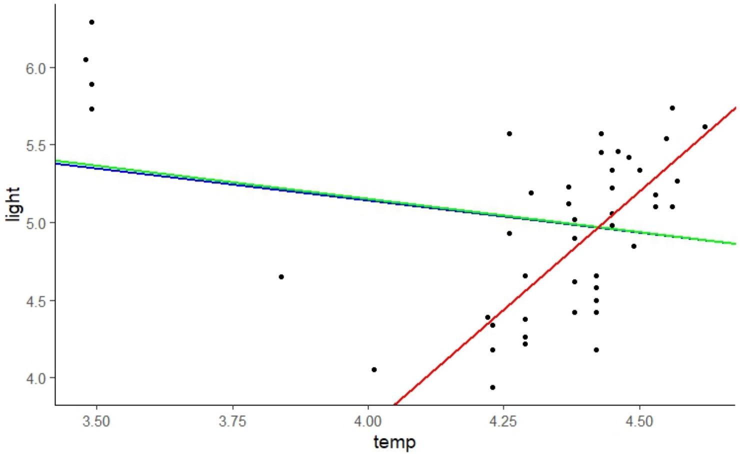

Figura 23: Dados sobre intensidade de luz e temperatura de estrelas com ajuste via OLS (azul), estimadores M (verde) e LTS (vermelho)

Exercícios

Exercícios

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