Transformações e Ponderação para corrigir violações do modelo

Diagnóstico na análise de regressão

Relembrando suposições

- Os erros do modelo tem média zero e variância constante.
- Os erros do modelo tem distribuição normal.
- A forma paramétrica estabelecida para o modelo está correta.

Algumas medidas para contornar problemas do modelo de regressão

Modelo de regressão linear simples não é adequado $\longrightarrow$
- Usar um modelo apropriado
- Usar transformações

Não linearidade do modelo de regressão

- Mudar o modelo

$$E(Y) = \beta_0 + \beta_1 X + \beta_2 X^2$$

$$E(Y) = \beta_0 \beta_1^X \quad (Exponencial)$$

$$E(Y) = \frac{\beta_0}{1 + \beta_1 \exp(\beta_2 X_i)} \quad (logístico)$$

- Usar transformação

Variâncias heterogêneas

⇒ Usar o método de mínimos quadrados ponderados para estimar os parâmetros

⇒ Usar transformação

Algumas medidas para contornar problemas do modelo de regressão

Falta de normalidade

A falta de normalidade geralmente vem junto com falta de homogeneidade de variâncias. Frequentemente, a mesma transformação estabiliza a variância e aproxima para normalidade, portanto, primeiro usar uma transformação para estabilizar a variância (será visto na próxima seção).

Omissão de variável preditora importante

→ Modificar o modelo de regressão múltipla

Outliers

→ Usar procedimentos de estimação robustos (método dos mínimos quadrados reponderados iterativamente), pois os métodos de mínimos quadrados e máxima verossimilhança produzem estimativas distorcidas.

Transformações para linearizar o modelo

Ocasionalmente dectecta-se a suposição de linearidade é violada.
Como:

- diagramas de dispersão
- gráfico dos resíduos (gráfico de regressão parcial)
- experiência a priori (não usual).

No caso de experência a priori, a função não linear pode ser linearizada. Esses modelos são chamados de intrisicamente linear.

Transformações para linearizar

Transformação da variável *Y* ou da variável preditora *X,* ou de *ambas,* frequentemente é suficiente para tornar o modelo de regressão linear simples apropriado para os dados transformados.

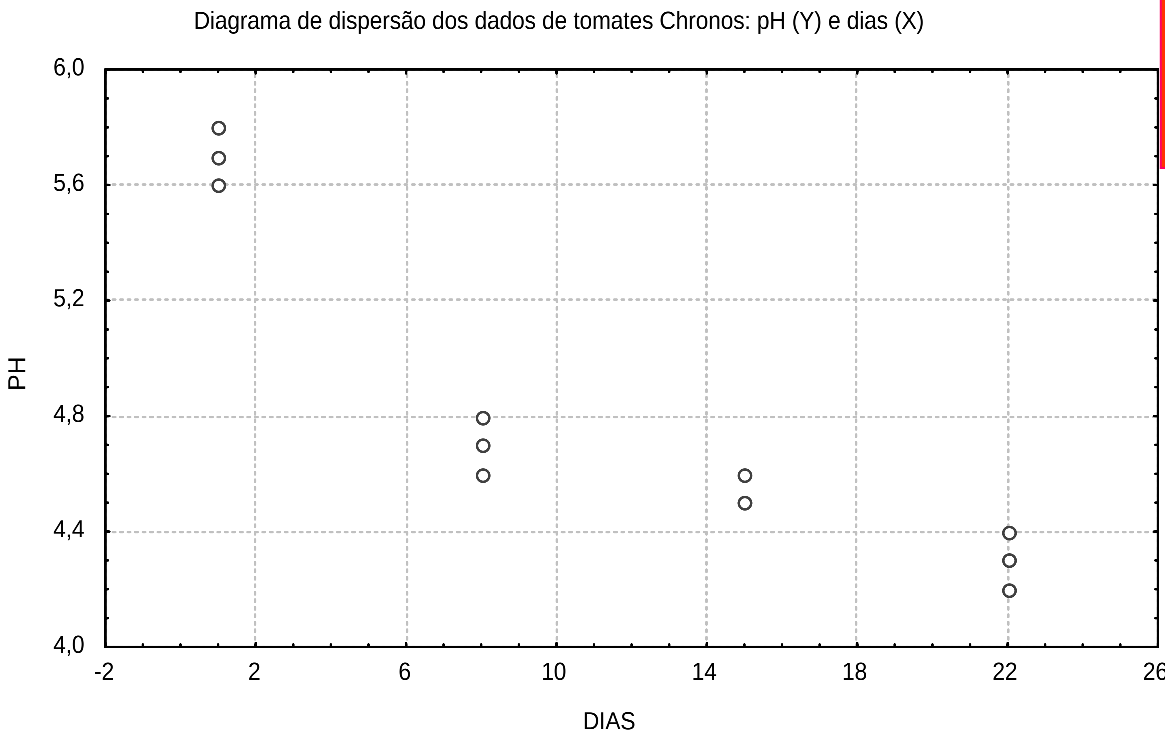
Padrões de relação entre *X* e *Y*

Modelo	$y = \beta_0 + \beta_1 \log x$	$y = \beta_0 e^{\beta_1 x}$
Transformação	$x' = \log x$	$y' = \ln y$
Forma linear	$y' = \beta_0 +_0 \beta_1 x'$	$y' = \ln \beta +_0 \beta_1 x$

Exemplo: Transformação nos regressores

Uma pesquisadora estava interessada em estudar o comportamento do pH de tomates Chronos *(Y)*, inteiros minimamente processados, submetidos ao tratamento vácuo, durante 22 dias de estocagem *(X)*, a uma temperatura média de 8ºC e umidade relativa de 62,78%.

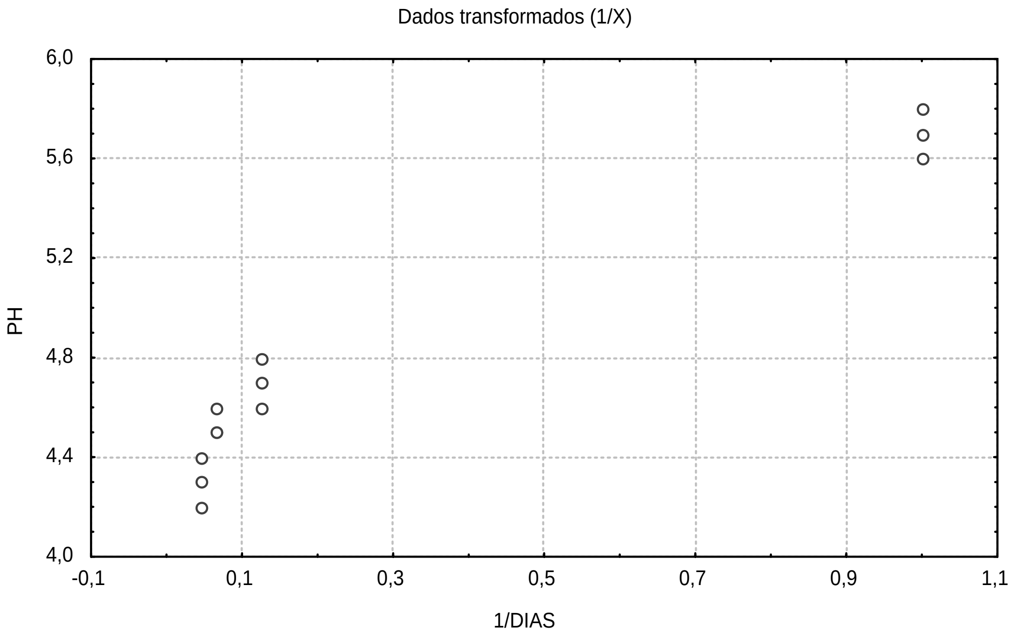
Condições aproximadamente satisfeitas:
Normalidade
Independência
e variância constante



O diagrama de dispersão indica uma relação curvilínea. A variabilidade nos diferentes níveis de *X* parece constante.

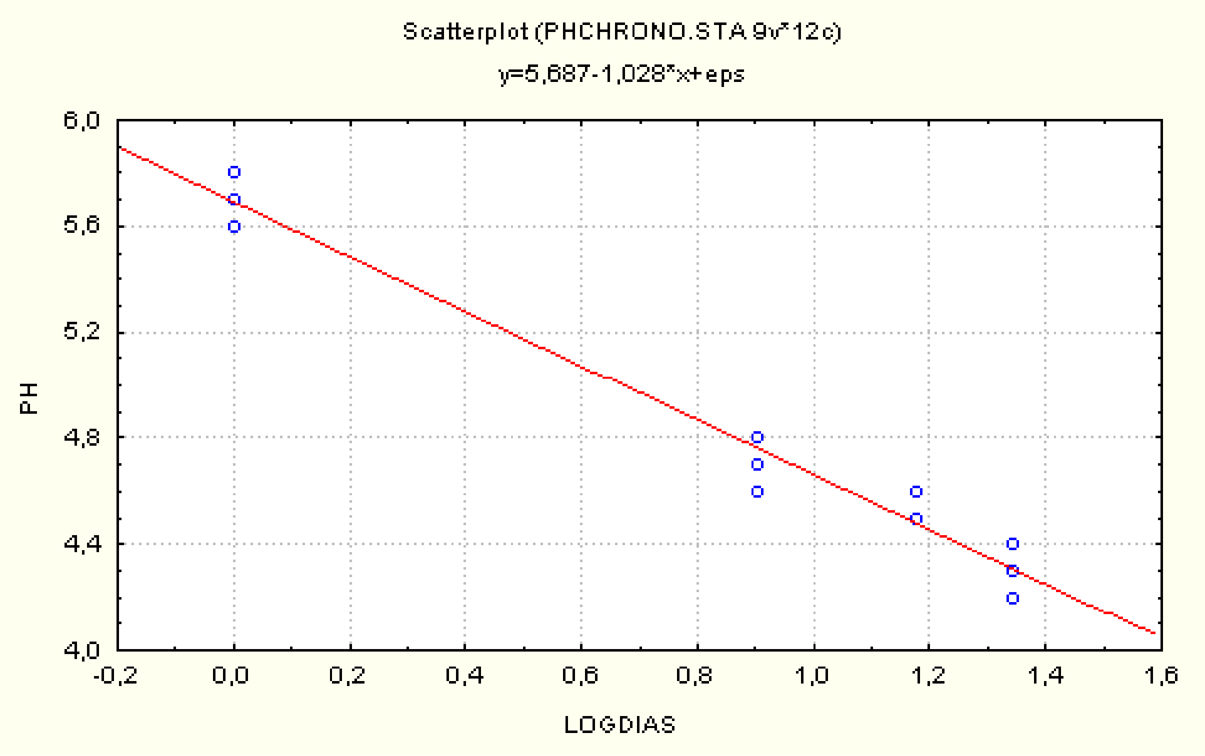
Vamos considerar a transformação $X' = 1/X$.

Exemplo



Os dados continuam mostrando um comportamento curvilíneo. A variabilidade nos diferentes níveis de X continua constante.

Exemplo



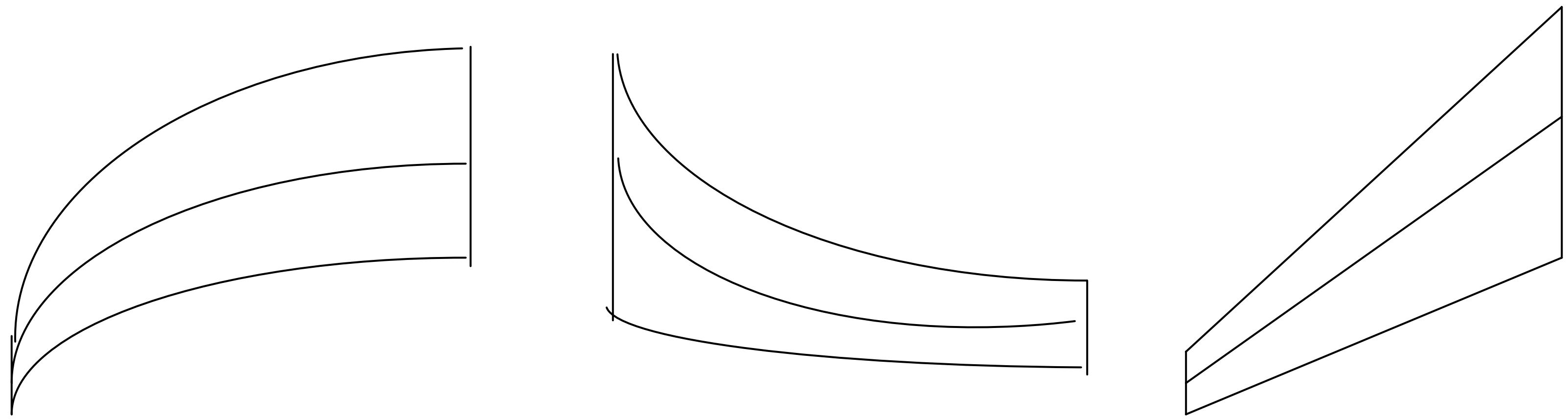
A transformação $\log_{10}$ (dias) linearizou a função de regressão. A variabilidade permanece constante.

Outra transformação muito usada é potência de x.

Transformações para estabilizar variância

Variâncias heterogêneas e não normalidade dos erros frequentemente aparecem juntas. Necessita-se fazer uma transformação em *Y*, pois a forma e a dispersão em *Y* precisam ser modificadas. A transformação em *Y* pode também eliminar o problema de não linearidade do modelo. Outras vezes uma transformação também em *X* é necessária para manter ou obter uma relação linear.

A figura ilustra algumas formas de relacionamento onde a assimetria e as variâncias aumentam com a reposta média *E(Y).*



Transformações sobre *Y*:

Seleciona-se empiricamente

$$Y' = \sqrt{Y}$$

$$Y' = \log_{10} Y$$

$$Y' = 1/Y$$

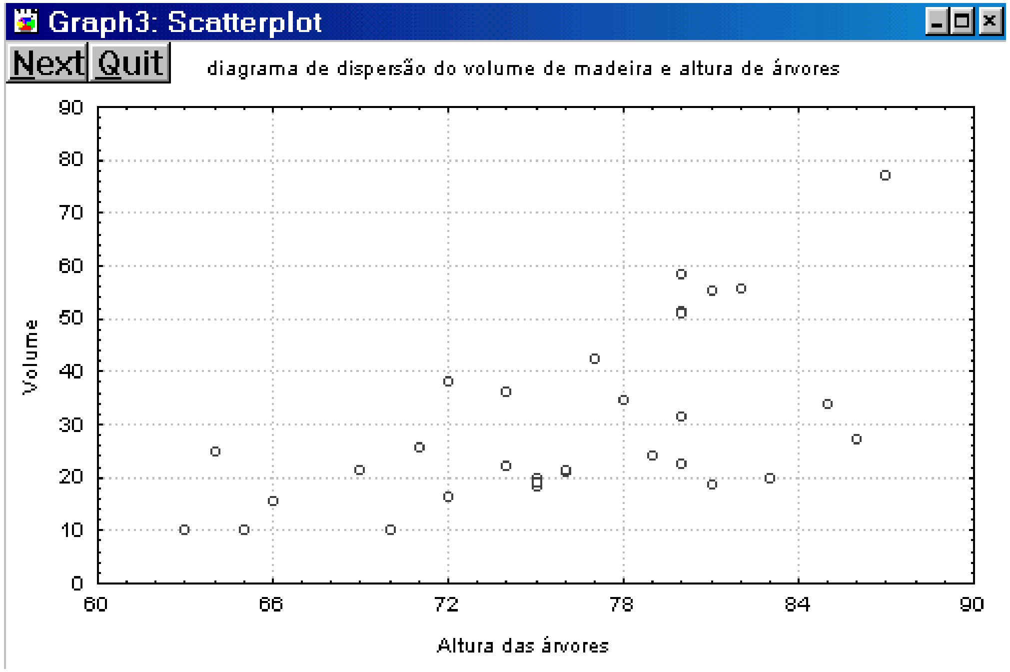
Nota: uma transformação em *X* pode ser útil ou necessário.

Fazer análise de resíduos

Exemplo

objetivo: estimar o volume da árvore em pé a partir de medidas mais facilmente obtidas.

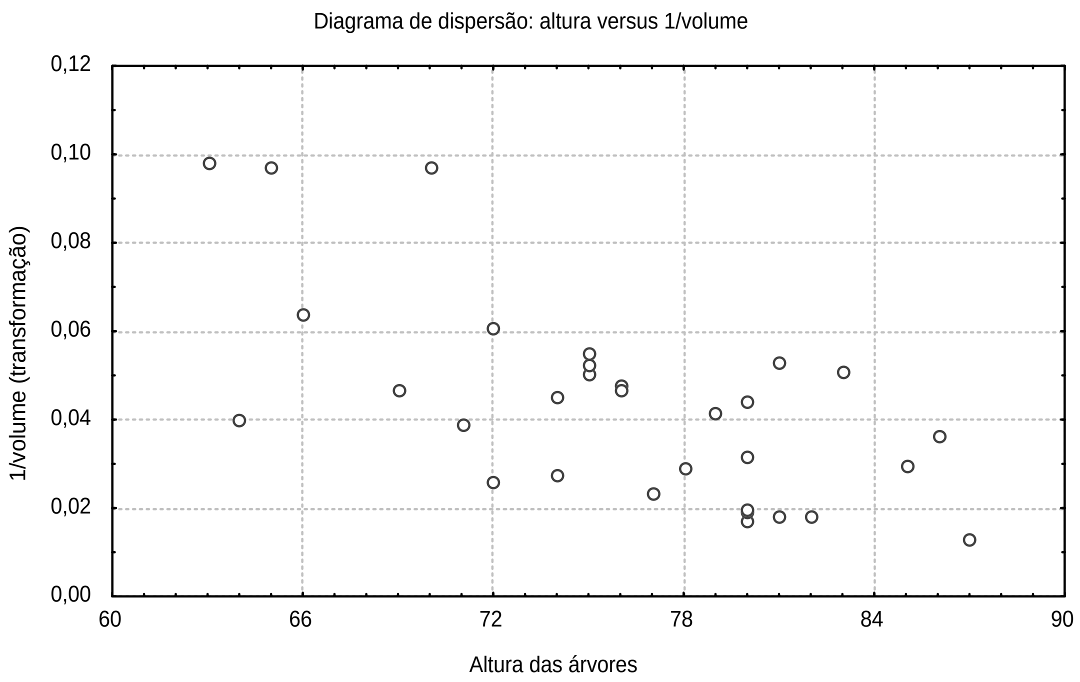
Y=volume da árvore em pés cúbicos; X1=diâmetro da árvore em polegadas a 4 pés e 6 polegadas acima do solo; X2=altura da árvore em pés.



Observamos maior variabilidade para valores maiores de altura. A relação entre volume e altura é linear.

Exemplo

Transformação: valores inverso de Y (1/Y).

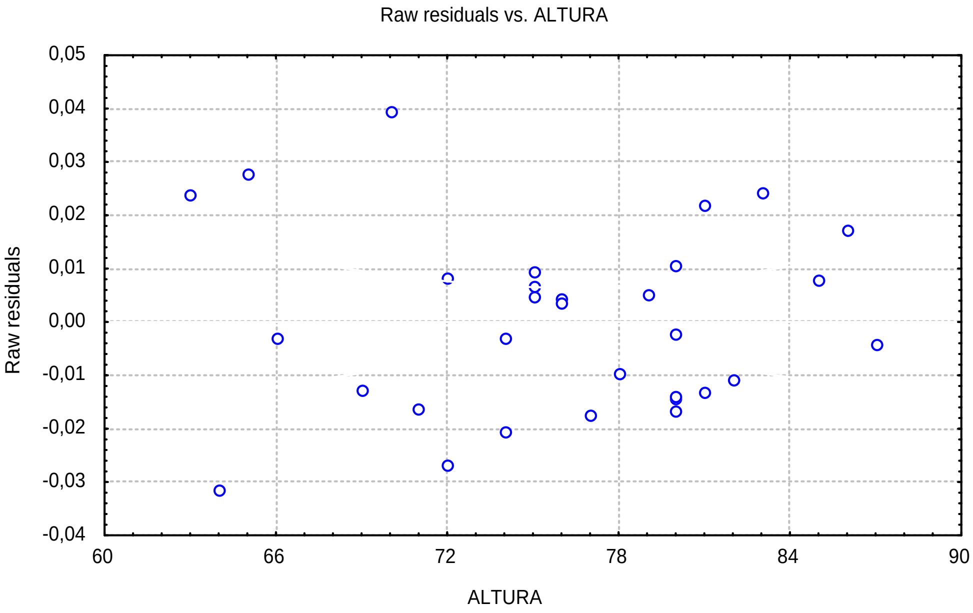


Note que a transformação tornou a variância razoavelmente constante para os diferentes níveis de X.

O modelo de regressão linear simples ajustado aos dados com a transformação $Y' = 1/Y$ é dado por:

$$\hat{Y}' = 0{,}22386 - 0{,}002377X$$

Exemplo



Indica que o modelo é apropriado para os dados transformados

Se desejamos estimar os valores de Y, na unidade original, fazemos:

$$\hat{Y} = \frac{1}{0{,}22386 - 0{,}002377X}$$

Método analítico de seleção da melhor Transformação

Transformação Box-Cox
(Box and Cox (1964))

A transformação Box-Cox automaticamente identifica uma transformação a partir de uma família de transformações potência de *Y*. A família de transformações potência é dada por:

$$\boxed{Y' = Y^{\lambda}}$$

Onde λ é um parâmetro a ser determinado a partir dos dados da amostra. Esta família inclui, por exemplo,

$$\lambda = 2 \rightarrow \boldsymbol{Y'} = Y^2$$

$$\lambda = \boldsymbol{0{,}5} \rightarrow \boldsymbol{Y'} = \sqrt{Y}$$

$$\lambda = -0{,}5 \rightarrow \boldsymbol{Y'} = \tfrac{1}{\sqrt{Y}}$$

$$\lambda = 0 \rightarrow \boldsymbol{Y'} = \log_e Y \ \textbf{\textit{(por definição)}}$$

$$\lambda = \boldsymbol{-1{,}0} \rightarrow \boldsymbol{Y'} = \tfrac{1}{Y}$$

Método analítico de seleção da melhor Transformação

O modelo de regressão com erros normais com a variável resposta pertencente a família de transformação potência fica:

$$Y_i^{(\lambda)} = \beta_0 + \beta_1 X_i + \varepsilon_i$$

onde

$$Y_i^{(\lambda)} = \begin{cases} \dfrac{y^{\lambda} - 1}{\lambda \dot{y}^{\lambda - 1}}, \lambda \neq 0 \\ \dot{y} \ln y, \lambda = 0 \end{cases} \qquad \dot{y} = \ln^{-1}\left[\frac{1}{n} \sum_{i=1}^{n} \ln y_i \right]$$

Seleciona-se diferentes valores para λ. Ajusta-se os modelos usando $y^{(\lambda)}$ e seleciona a transformação (λ) tal que a soma de quadrados de resíduos SS_R é mínima.

Usualmente 10 a 20 valores de λ são suficientes. Usar diferenças pequenas.

Exemplo

Continuamos com o exemplo das árvores (*X*=altura e *Y*=volume). Vamos tomar os seguintes valores para lambda

$$\lambda = -1 \quad \lambda = \text{-}0{,}3 \quad \lambda = -0{,}2 \quad \lambda = \text{-}0{,}1 \quad \lambda = 0 \quad \lambda = 0{,}1 \quad \lambda = 0{,}2 \quad \lambda = 0{,}3 \quad \lambda = 1$$

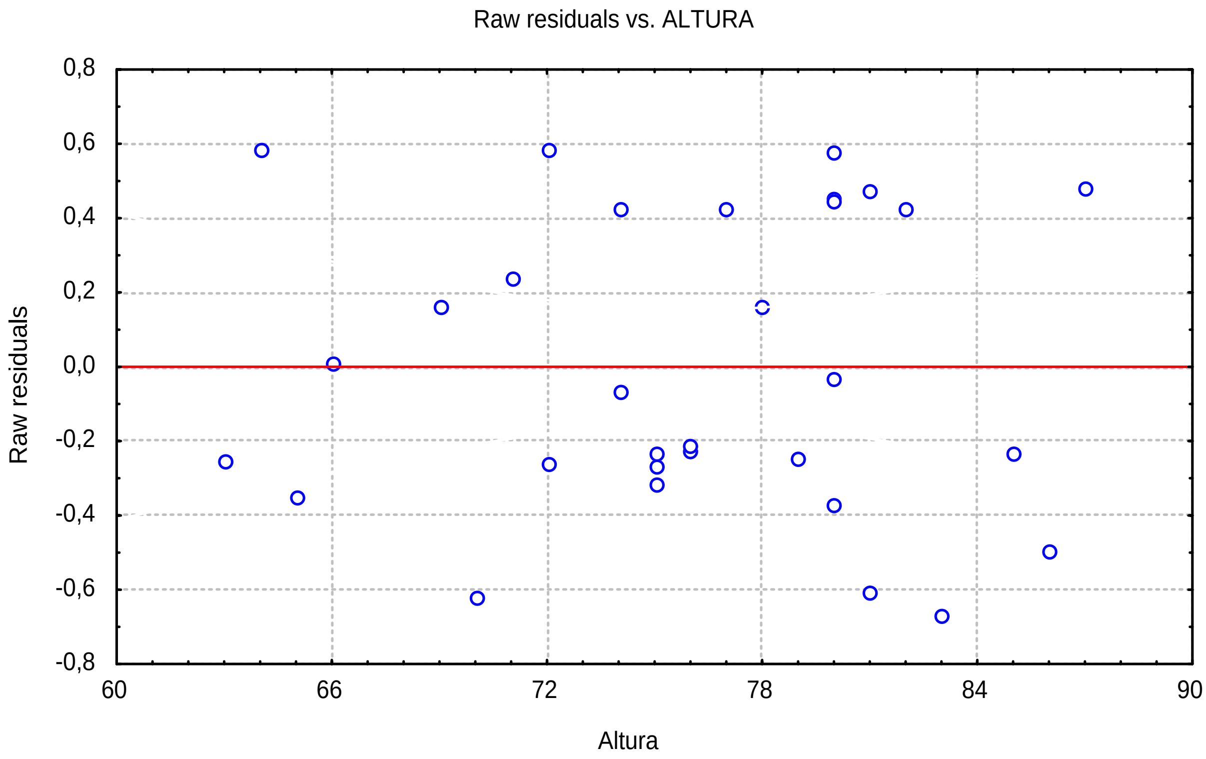
λ	$-1{,}00$	-0,30	$-0{,}2$	$-0{,}10$	0,00	0,10	0,20	0,30	1,00
SQR	4201,9	**3324,5**	**3310,3**	**3319,8**	**3352,9**	3409,7	3490,5	3596,3	5204,9

Observe na tabela acima que a transformação Box-Cox indica λ próximo de -0,20. Entretanto, a SQR é aproximadamente estável na faixa de -0,30 a 0,00, portanto, vamos usar a transformação logarítmica por ser a preferida dos pesquisadores (é uma transformação que os pesquisadores entendem melhor). A transformação Box-Cox dá um direção no sentido da escolha da melhor transformação.

Observe que a transformação usada anteriormente, *1/Y,* não foi razoável de acordo com transformação de Box-Cox. (compare os dois gráficos de resíduos).

Quando a transformação Box-Cox produz um λ próximo de 1, não é necessário transformar os dados.

Exemplo



Indica a adequação do modelo de regressão para os dados transformados (transformação logarítmica)

Estabilizar variância: Mínimos quadrados generalizados.

Uso: Observações y são correlacionadas e variâncias desiguais.

Modelo

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon} \qquad E(\boldsymbol{\varepsilon}) = \boldsymbol{0}, \quad Var(\boldsymbol{\varepsilon}) = \sigma^2 \boldsymbol{V}$$

V deve ser não singular e definida positiva. Então existe uma matriz simétrica **K** tal que

$$\boldsymbol{K}^T \boldsymbol{K} = \boldsymbol{V}$$

Tipicamente σ^2 é desconhecido. Nesse caso, ***V*** é assumida como a matriz de variâncias covariâncias entre os erros a menos de uma constante. ***K*** é a matriz raiz quadrada de V.

Estabilizar variância: Mínimos quadrados generalizados.

Novas variáveis.

$$z = K^{-1}y$$

$$B = K^{-1}X$$

$$g = K^{-1}\boldsymbol{\varepsilon}$$

Modelo original $$y = X\beta + \boldsymbol{\varepsilon}$$

Novo Modelo

$$K^{-1}y = K^{-1}X + K^{-1}\boldsymbol{\varepsilon}$$

$$z = \boldsymbol{B\beta} + g$$

Após uma algebra: $$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{y}$$

Estimador de mínimos quadrados generalizados

Estabilizar variância: Mínimos quadrados ponderado

Uso: Observações y independentes e variâncias desiguais.

$$Var(\boldsymbol{\varepsilon}) = \sigma^2 \boldsymbol{V} = \sigma^2 \begin{bmatrix} 1/w_1 & & 0 \\ & & \\ 0 & & 1/w_n \end{bmatrix}$$

As observações com grandes variâncias tem peso menores.

Seja $\boldsymbol{W}=\boldsymbol{V}^{-1}$, após uma algebra

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T \boldsymbol{W} \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{W} \boldsymbol{y}$$

Estimador de mínimos quadrados ponderados.

Estabilizar variância: Mínimos quadrados ponderado

Estimativas de Mínimos quadrados ponderados podem ser facilmente obtidos pelos Mínimos quadrados ordinários.

Pode-se reescrever o estimador dos coeficientes da seguinte forma:

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{B}^T \boldsymbol{B})^{-1} \boldsymbol{B}^T \boldsymbol{z}$$

onde:

$$\boldsymbol{B} = \begin{bmatrix} 1\sqrt{w_1} & x_1\sqrt{w_1} & x_{p-1}\sqrt{w_1} \\ 1\sqrt{w_2} & x_1\sqrt{w_2} & x_{p-1}\sqrt{w_2} \\ \cdots & \cdots\cdot & \cdots \\ 1\sqrt{w_n} & x_1\sqrt{w_n} & x_{p-1}\sqrt{w_n} \end{bmatrix} \qquad \boldsymbol{z} = \begin{bmatrix} y_1\sqrt{w_1} \\ y_2\sqrt{w_2} \\ \cdots \\ y_n\sqrt{w_n} \end{bmatrix}$$

Como estimar *V*

1. Ordene as obeservações em y.
2. Encontre clusters de valores de x na ordem obtida.
3. Dentro de cada cluster calcule a média de x e para os correspondentes valores de y calcule a variância.
4. Os valores da variância de y devem aumentar ou diminuir quando os valores médios de x aumentam.
5. Construa uma regressão para a variância de y usando a média de x como regressor.
6. Para cada valor de x do conjunto encontre o valor da estimativa da média da variancia de y.
7. Calule os pesos como o inverso dos valores obtidos no passo 6.
8. Obtenha **B** e **z**.
9. Ajuste o modelo:

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{B}^T\boldsymbol{B})^{-1}\boldsymbol{B}^T\boldsymbol{z}$$