Modelo de Regressão Linear Simples

O modelo de regressão linear simples considera apenas uma variável explicativa e a função de regressão é linear. O modelo é definido por:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, \tag{1}$$

em que:

- Y_i é o valor da variável resposta para a i-ésima observação,
- β_0 é o intercepto e β_1 é o coeficiente angular, ambos são parâmetros desconhecidos,
- X_i é uma constante conhecida, o valor da variável explicativa para a i-ésima observação. X_i é uma variável fixa (ou sem erro ou determinística),
- ε_i são independentes e $N(0, \sigma^2)$.

Análise de Variância

- A análise de variância é baseada na partição da soma de quadrados e no número de graus de liberdade associados à variável resposta, Y.
- Hipóteses:

$$H_0 : \beta_1 = 0$$
$$H_1 : \beta_1 \neq 0$$

ou,

$$H_0 : \text{ausência de regressão}$$
$$H_1 : \text{existe regressão}$$

Análise de Variância

Tabela: Tabela de Análise de Variância

Fonte de Variação	SQ	g.l.	QM	F
Regressão	SQReg	1	MSReg	MSReg/MSRes
Resíduo	SQRes	n-2	MSRes	
Total	SQT	n-1		

Análise de Variância

- Estatística do teste

$$F = \frac{MSReg}{MSRes},$$

tem distribuição F com 1 grau de liberdade no numerador e (n-2) graus de liberdade no denominador.

- Se $F \leq F_{(1-\alpha,1,n-2)} \implies$ Há evidência para aceitar H_0
- Se $F > F_{(1-\alpha,1,n-2)} \implies$ Há evidência para rejeitar H_0

Coeficiente de Determinação

- SQT é a medida da variação das observações Y_i sem considerar o efeito da variável regressora X.
- SQRes é a medida da variação das observações Y_i quando o modelo de regressão utilizando a variável regressora X é considerado.
- Um medida do efeito de X em reduzir a variação de Y é

$$SQT - SQRes = SQReg$$

- Essa medida pode ser expressa em termos da proporção da variação total:

$$R^2 = \frac{SQReg}{SQT} = 1 - \frac{SQRes}{SQT}$$

Coeficiente de Determinação

- A medida R^2 é chamada de **Coeficiente de Determinação** e representa a proporção da variação total explicada pela relação X e Y (regressão).
- Desde de que $0 \leq SQRes \leq SQT$, segue que

$$0 \leq R^2 \leq 1$$

- Valores grandes de R^2 indicam que a variação total de Y é mais reduzida pela introdução da variável preditora X.
- Quando todas as observações estão na reta de regressão ajustada, SQRes=0 e $R^2 = 1$.
- Quando a reta de regressão ajustada é horizontal, então, $\hat{\beta}_1 = 0$ e SQRes=SQT. Logo, $R^2 = 0$.

Coeficiente de Correlação

- Uma medida de associação entre Y e X é o coeficiente de correlação. E pode ser encontrado por:

$$r = \pm\sqrt{R^2},$$

  o sinal de mais ou menos está associado ao sinal positivo ou negativo de $\hat{\beta}_1$.

- Assim, r é definido: $-1 \leq r \leq 1$.

- Encontre as medidas R^2 e r para os dados do Exercício.

Diagnóstico

- É fundamental verificar a adequabilidade do modelo antes de tirar conclusões sobre o processo inferencial.
- Verificar as características do modelo.
- Para avaliar a adequabilidade do modelo serão usados:
  - Métodos gráficos
  - Testes estatísticos
- Também será considerado algumas técnicas para remediar quando os dados não seguem as suposições do modelo.

Diagnóstico para variável explicativa

- É importante fazer análise de diagnóstico para a variável explicativa para verificar se existe alguns valores que estão afastados (valores atípicos) que podem influenciar a adequabilidade do modelo de regressão ajustado.

- Diagnóstico sobre a amplitude e concentração dos níveis da variável explicativa X no estudo é usado também para determinar a amplitude da validade da análise de regressão.

Diagnóstico para variável explicativa

- Métodos gráficos
  - Diagrama de pontos (dot plot) - é usado quando o número de observações do conjunto de dados não é grande.
  - Gráfico de sequência - deve ser utilizado quando os dados são obtidos em uma sequência, tal como de tempo ou para áreas geográficas adjacentes. Construir o gráfico da sequência de tempo versus X.
  - Ramo e folhas - fornece informação similar ao histograma.
  - Diagrama em Caixas (Box Plot) - é utilizado quando o número deobservações do conjunto de dados é grande. Verifica-se o comportamento da distribuição dos dados.

Resíduo

- Gráficos de diagnóstico para variável resposta Y não são tão utilizados na análise de regressão porque os valores das observações sobre a variável resposta são uma função do nível da variável explicativa.

- O diagnóstico para variável resposta é feito indiretamente através da análise dos resíduos.

Resíduo

- Como definido anteriormente, o resíduo é a diferença entre o valore observado Y_i e o valor ajustado $\hat{Y}_i$:

$$e_i = Y_i - \hat{Y}_i. \tag{2}$$

- O resíduo (2) pode ser considerado como o erro observado, distinto do erro verdadeiro do modelo ε_i que é desconhecido:

$$\varepsilon_i = Y_i - E(Y_i). \tag{3}$$

Resíduo

- Para o modelo de regressão (1), assume-se que o erros, ε_i são variáveis aleatórias independentes, identicamente distribuídos com distribuição Normal com média 0 e variância σ^2.

- Se o modelo de regressão é apropriado para os dados em estudo, os resíduos observados, e_i, devem refletir as propriedades assumidas para o erro do modelo, ε_i.

Propriedades dos Resíduos

- **Média**: a média de n resíduos, e_i, para o modelo (1) é definida por:

$$\bar{e} = \frac{\sum_{i=1}^{n} e_i}{n} = 0.$$

- Desde que $\bar{e}$ é sempre ZERO, não fornece qualquer informação se os erros verdadeiros, ε_i, tem valor esperado $E(\varepsilon_i) = 0$.

Propriedades dos Resíduos

- **Variância**: a variância de n resíduos, e_i, para o modelo (1) é definida por:

$$\frac{\sum_{i=1}^{n}(e_i - \bar{e})^2}{n-2} = \frac{\sum_{i=1}^{n} e_i^2}{n-2} = \frac{SQRes}{n-2} = MSRes = \hat{\sigma}^2.$$

- Se o modelo (1) é apropriado, $\hat{\sigma}^2$ é um estimador não viesado da variância do erro, $\mathrm{Var}(\varepsilon_i) = \sigma^2$.

Propriedades dos Resíduos

- **Não independência**: os resíduos e_i não são variáveis aleatórias independentes, pois eles dependem dos valores ajustados que são baseados na mesma função de regressão ajustada.
- Os resíduos do modelo de regressão estão sujeitos a duas condições:

$$\sum_{i=1}^{n} e_i = 0 \qquad \text{e} \qquad \sum_{i=1}^{n} X_i e_i = 0$$

- Quando o tamanho da amostra é grande em relação ao número de parâmetros no modelo de regressão, o efeito da dependência entre os resísuos é relativamente pouco importante e pode ser ignorado para maioria dos propósitos.

Resíduos

- As vezes é útil padronizar os resíduos para a análise de resíduos.
- Como o desvio padrão dos erros, ε_i, é σ, que pode ser estimado por $\hat{\sigma} = \sqrt{\text{MSRes}}$. É natural considerar a seguinte padronização:

$$e_i^* = \frac{e_i - \bar{e}}{\sqrt{\text{MSRes}}} \tag{4}$$

- Se $\hat{\sigma} = \sqrt{\text{MSRes}}$ é uma estimativa do desvio padrão do resíduo e_i, e_i^* é chamado de resíduo studentizado.
- Se o desvio padrão de e_i é complexo e varia para os diferentes resíduos e_i e $\hat{\sigma} = \sqrt{\text{MSRes}}$ é apenas uma aproximação do desvio padrão de e_i, e_i^* é chamado de resíduo semistudentizado.

Afastamentos do Modelo a serem estudados pelos Resíduos

- A função de regressão não é linear.
- Os erros do modelo não tem variância constante.
- Os erros não são independentes.
- O modelo apresenta um ou poucos valores atípicos.
- Os erros não são normalmente distribuidos.
- Uma ou mais variáveis explicativas importantes foram omitidos do modelo.

Análise de Resíduos

- Para verificar qualquer dos seis tipos de afastamento do modelo de regressão linear (1) serão utililizados os seguintes gráficos dos resíduos ou gráficos dos resíduos studentizado/semistudentizado.
  - Gráfico dos resíduos versus a variável explicativa.
  - Gráfico dos resíduos ao quadrado ou em valor absoluto versus a variável explicativa.
  - Gráfico dos resíduos versus valores ajustados.
  - Gráfico dos resíduos versus tempo ou outra sequência.
  - Gráfico dos resíduos versus variáveis explicativas omitidas.
  - Diagrama de caixas (box plot) dos resíduos.
  - Gráfico de probabilidade normal dos resíduos.

Não linearidade da função de regressão

- Gráfico dos resíduos versus a variável explicativa ou gráfico dos resíduos versus valores ajustados, mostram se a função de regressão é apropriada para os dados analisados.

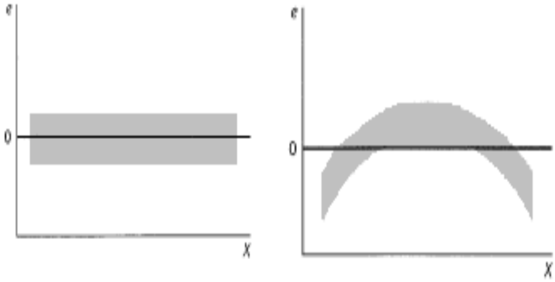
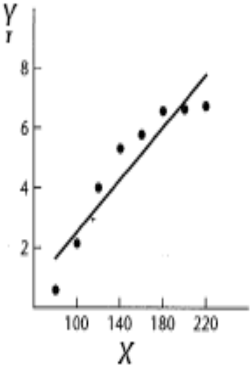


FIGURA: (a) Modelo de regressão linear é apropriado. (b) Modelo de regressão linear não é apropriado

Não linearidade da função de regressão

- O diagrama de dispersão também pode ser usado para verificar se a função de regressão é apropriada para os dados analisados.
- Porém, nem sempre é tão eficaz como os gráficos dos resíduos.
- Existem situações em que a escala do diagrama de dispersão colocam as observações Y_i próximas dos valores ajustados $\hat{Y}_i$. Por exemplo, quando a declividade é muito acentuada.

Não linearidade da função de regressão



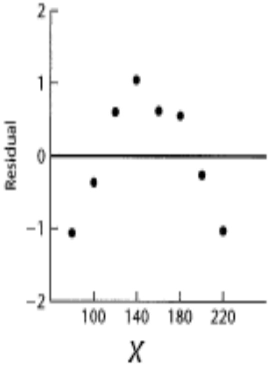
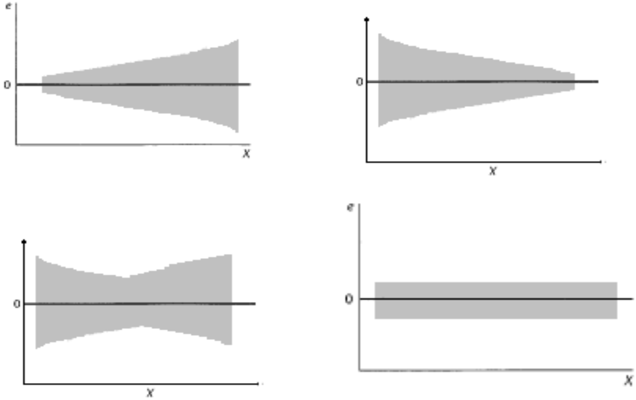


Figura: Exemplos

Heterocedasticidade da Variância do Erro

- Os gráficos dos resíduos versus a variável explicativa ou gráfico dos resíduos versus valores ajustados, também serão utilizados para verificar se a variância dos erros é constante.



Heterocedasticidade da Variância do Erro

- Gráficos:
  - do valor absoluto do resíduo versus a variável explicativa
  - do valor absoluto do resíduo versus valores ajustados
  - do quadrado do resíduo versus a variável explicativa
  - do quadrado do resíduo versus valores ajustados

  também podem ser usados para verificar se a variância dos erros é constante. Pois, os sinais dos resíduos não são importantes para examinar a constância da variância do erro.

- Eles são especialmente úteis quando não há um grande número de observações na amostra, pois toda a informação sobre a magnitude dos resíduos é colocado acima da linha zero horizontal e pode-se ver prontamente se as magnitudes dos resíduos está mudando com os níveis de X_i ou de $\hat{Y}_i$.

Heterocedasticidade da Variância do Erro

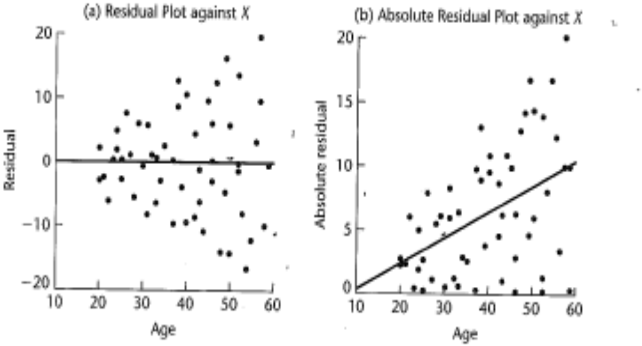


Figura: Exemplos

Presença de Valores Atípicos

- Valores atípicos são observações extremas.
- Gráficos que podem ser utilizados para identificar valores atípicos:
  - Gráfico dos resíduos versus a variável explicativa.
  - Gráfico dos resíduos versus valores ajustados.
  - Diagram de caixas (box plot).
  - Ramo e folhas
  - Diagrama de pontos dos resíduos.
- Gráficos de resíduos studentizados são particularmente úteis para distinguir valores atípicos , visto que eles tornam mais fácil identificar resíduos que se afastam vários desvios padrões de zero.
- Regra: quando o número de casos é grande, se valor absoluto do resíduo studentizado for ≥ 4, considerar valor atípico.

Presença de Valores Atípicos

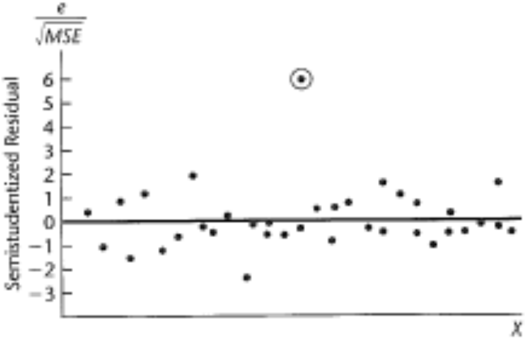


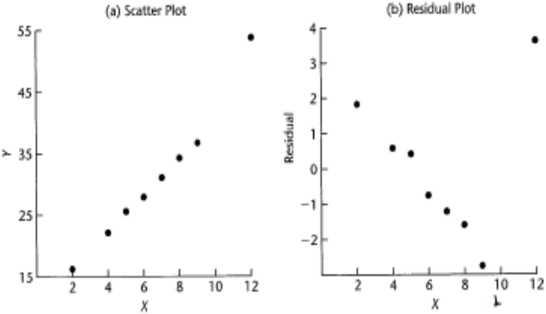
Figura: Exemplo de valor atípico

Presença de Valores Atípicos

- A presença de valor atípico nos dados pode ocorrer por diferentes motivos, como:
  - Observação resultou de um erro ou outro efeito estranho, deve ser descartado. Pois, a reta ajustada pode ser puxada desproporcionalmente em sua direção.
  - Entretanto, valores atípicos podem ser resultados de uma interação com outra variável explicativa omitida no modelo e portanto traz informação importante.
- **Regra**: Só descartar um valor atípico se existe uma evidência direta que ele representa um erro, um erro de cálculo, um mal funcionamento do equipamento, ou um tipo similar de circunstância.

Presença de Valores Atípicos

- Quando um modelo de regressão linear é ajustado a um conjunto de dados com poucas observações e existe um valor atípico, a reta ajustada pode estar tão distorcida por esse valor que o gráfico de resíduos pode sugerir, não apropriadamente, falta de ajustamento do modelo linear, além da presença de valor atípico.



Os erros não são independentes

- Quando os dados são obtidos em uma sequência de tempo ou algum outro tipo de sequência ou para áeras geográficas adjacentes, é bom construir um gráfico de resíduos em sequência: resíduos versus tempo ou outro tipo de sequência.
- O objetivo é verificar se existe qualquer correlação entre os erros que são próximos um dos outros na sequência.
- Quando os erros são independentes espera-se que os resíduos em um gráfico de sequência flutuem em um padrão mais ou menos aleatório em torno da linha de referência Zero.

Os erros não são independentes

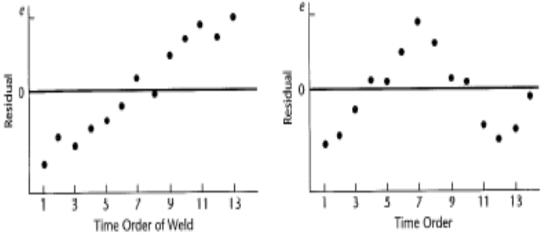


Figura: Exemplo

Os erros não são normalmente distribuidos

- A normalidade dos erros pode ser investigada por meio dos seguintes recursos gráficos:
  - Box plot, histograma, ramo e folhas e dot plot dos resíduos. Pode-se observar simetria dos resíduos e possíveis valores atípicos.
  - No entanto, o número de observações deve ser razoavelmente grande para se obter informações confiáveis sobre a forma da distribuição dos erro.

Os erros não são normalmente distribuidos

- Outro recurso gráfico importante é o **Gráfico de Probabilidade Normal dos resíduos** - cada resíduo é colocado em um gráfico versus seu valor esperado sob normalidade.
- Se o resultado gráfico for aproximadamente linear, há evidências da suposição de normalidade.
- Se o resultado gráfico afasta substancialmente da linearidade, não há evidências de que a distribuição do erro é normal.

Os erros são normalmente distribuidos

- Teste Shapiro Wilks - um estudo concluiu que esse teste tem o melhor poder quando comparados com os testes abaixo. Porém, deve ser usado com cuidado para amostras com muitos valores identicos.
- Teste Kolmogorov-Smirnov
- Teste Lillieforts - é baseado no teste Kolmogorov-Smirnov

Homogeneidade da Variância do Erro

Será apresentado dois testes para verificar se a variância do erro é constante. Os testes são:

- Teste Brown-Forsythe
- Teste Breusch-Pagan

Teste Brown-Forsythe

- É uma modificação do Teste de Levene - não depende da normalidade dos erros.
- É um teste robusto para afastamentos sérios da normalidade - o nível de significância nominal permanece aproximadamente correto quando os erros têm a mesma variância mesmo que a distribuição dos erros não seja normal.
- É aplicado para modelos de regressão linear simples quando a variância dos erros ou cresce ou decresce com X.
- O tamanho da amostra deve ser suficientemente grande de modo que a dependência entre os resíduos possa ser ignorada.

Teste Brown-Forsythe

- O primeiro passo é dividir os dados em dois grupos de acordo com os níveis da variável explicativa X:
  - O grupo 1 é formado pelas observações de X com nível baixo: e_{i1} é o i-ésimo resíduo do grupo 1, $i = 1, \ldots, n_1$
  - O grupo 2 é formado pelas observações de X com nível alto: e_{i2} é o i-ésimo resíduo do grupo 2, $i = 1, \ldots, n_2$
- Sejam $\tilde{e}_1$ e $\tilde{e}_2$ as medianas dos resíduos dos grupos 1 e 2, respectivamente, e os desvios em valor absoluto:

$$d_{i1} = |e_{i1} - \tilde{e}_1| \quad d_{i2} = |e_{i2} - \tilde{e}_2|$$

Teste Brown-Forsythe

- Estatística do teste

$$t^*_{BF} = \frac{\bar{d}_1 - \bar{d}_2}{s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

em que $\bar{d}_1$ e $\bar{d}_2$ são as médias de d_{i1} e d_{i2}, respectivamente. E a variância agrupada é definida por:

$$s^2 = \frac{\sum(d_{i1} - \bar{d}_1)^2 + \sum(d_{i2} - \bar{d}_2)^2}{n-2}$$

- Se variância dos erros é constante e as amostras n_1 e n_2 não são muito pequenas t^*_{BF} tem aproximadamentre distribuição t-Student com n-2 graus de liberdade.

Teste Breusch-Pagan

- Teste para grandes amostras

- Assume que os erros são independentes com distribuição normal e que a variância dos erros ε_i, designada por σ_i^2 , está relacionada com o nível de X da seguinte forma:

$$\log_e \sigma_i^2 = \gamma_0 + \gamma_1 X_i$$

Teste Breusch-Pagan

- Hipóteses

$$H_0 : \gamma_1 = 0$$
$$H_1 : \gamma_1 \neq 0$$

- Estatística do teste

$$\chi^2_{\text{BP}} = \frac{SQReg^*}{2} \div \left(\frac{SQRes}{n}\right)^2,$$

em que:

  - $SQReg^*$ é a soma de quadrados da regressão para a regressão de σ_i^2 e X_i,
  - SQRes é a soma de quadrados do resíduo para a regressão de Y e X,
  - Se H_0 é verdadeira e n é razoavelmente grande, χ^2_{BP} tem aproximadamente distribuição Qui-quadrado com 1 grau de liberdade.

Transformação Box Cox

Para auxiliar a descobrir qual a melhor transformação que deve ser adotada, surge a Família de Transformações Box e Cox, dada por:

$$Y' = Y^{\lambda} = \begin{cases} \frac{Y^{\lambda}-1}{\lambda}, & \lambda \neq 0 \\ \log(Y), & \lambda = 0, \end{cases}$$

sendo λ o parâmetro da transformação e Y a variável resposta. Na ausência de uma transformação, $\lambda = 1$.
O modelo de regressão com a variável resposta um membro da família de transformação é definido por:

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

$\Longrightarrow$ Para estimar os parâmetros do modelo, incluindo o parâmetro λ, usa-se o método de máxima verossimilhança perfilada.