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 $N(0, \sigma^2)$.

Modelo de regressão linear geral

Em geral, a variável resposta Y pode estar relacionada com p variáveis explicativas. O modelo de regressão é definido por:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_p X_{ip} + \varepsilon_i, \tag{2}$$

é conhecido como modelo de primeira ordem, pois é linear nos parâmetros e linear nas variáveis explicativas. Esse modelo é conhecido como um modelo de regressão linear múltipla ou também como um modelo de regressão linear geral.

- Y_i é o valor da variável resposta para a i-ésima observação,
- β_0, $\beta_1, \ldots, \beta_p$ são parâmetros desconhecidos,
- $X_{i1}, X_{i2}, \ldots, X_{ip}$ são constantes conhecidas,
- ε_i são independentes e $N(0, \sigma^2)$, $i = 1, \ldots, n$.

Diagnóstico

- A análise de diagnóstico é um importante procedimento para avaliar a adequabilidade do modelo de regressão múltipla.

- Para verificar a adequabilidade do modelo serão usados:
  - Métodos Gráficos
  - Testes de Hipóteses.

Diagnóstico

- É interessante iniciar a análise de dados com os seguintes gráficos: Box plot, ramo e folhas, diagrama de dispersão univariado de cada uma das variáveis explicativas e também da variável resposta.
- O próximo passo é realizar um estudo bidimensional entre cada variável explicativa e a variável resposta e também entre duas variáveis explicativas. Para esse estudo pode-se usar o diagrama de dispersão, bem como, o coeficiente de correlação linear de Pearson.
- Esse estudo dará informações preliminares sobre os dados.

Diagnóstico

- Gráficos do resíduo versus os valores ajustados é utilizado para:
  - Avaliar a adequação da função de regressão múltipla
  - Os erros do modelo tem Variância constante
  - Presença de valor atípico
- Gráfico do resíduo absoluto versus os valores ajustados também é usado para verificar se os erros do modelo tem Variância constante.
- Gráfico dos resíduos versus tempo ou uma outra sequência é usado para verificar se os Erros são Independentes.
- Box plot e o gráfico normal de probabilidade são usados para verificar se os Erros são Normais.

Diagnóstico

- Gráficos dos resíduos versus cada variável explicativa dará informação sobre a adequação da função de regressão com respeito a variável explicativa considerada.
- Gráficos dos resíduos versus cada variável explicativa também informará possível variação da variância que pode estar relacionada com a variável explicativa considerada.
- Gráficos dos resíduos versus variável explicativa não introduzida no modelo ou ainda, versus interação, $X_1X_2, X_1X_3, X_2X_3, ...,$ mostrará se a variável ou a interação devem ser introduzidas no modelo.

Diagnóstico

- Forma gráfica ideal e não ideal para os pressupostos do modelo serem válidos:

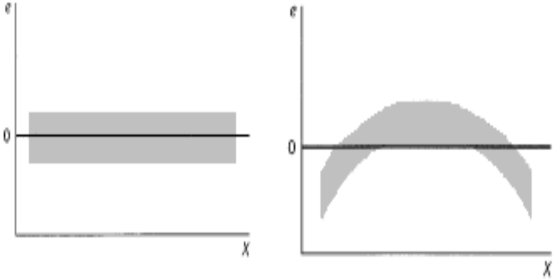

Figura : Modelo de regressão linear é apropriado. (b) Modelo de regressão linear não é apropriado

Diagnóstico

- Teste Shapiro Wilks - para normalidade dos erros
- Teste Brown-Forsythe e Breusch-Pagan para variância constante dos erros. São utlizados quando há evidências de que a variância do erro ou cresce ou decresce com uma determinada variável explicativa.
- Teste F para falta de ajustamento do modelo de regressão

Identificando observações discrepantes em relação à Y

Resíduos

- Resíduos - como já definido, o resíduo é dado por:

$$e_i = Y_i - \hat{Y}_i$$

- Resíduo semistudentizado ou padronizado

$$e_i^* = \frac{e_i}{\sqrt{MSRes}}$$

Matriz H - "chapéu"

- A matriz **H** é definida por:

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

- Os valores de $\hat{Y}_i$ podem ser expressos como uma combinação linear dos Y_i através da matriz **H**:

$$\hat{\mathbf{Y}} = \mathbf{H}\mathbf{Y}$$

- E os resíduos e_i também podem ser expressos como uma combinação linear dos Y_i através da matriz **H**:

$$\mathbf{e} = (\mathbf{I} - \mathbf{H})\mathbf{Y}$$

Matriz H - "chapéu"

- A matriz de variância e covariância do resíduo é definida como:

$$Cov(\mathbf{e}) = \sigma^2(\mathbf{I} - \mathbf{H})$$

- Dessa forma, a variância do resíduo, e_i é dada por:

$$Var(e_i) = \sigma^2(1 - h_{ii})$$

- E a covariância entre e_i e e_j, para $i \neq j$, é:

$$Cov(e_i, e_j) = \sigma^2(0 - h_{ij}) = -h_{ij}\sigma^2$$

em que h_{ii} são os elementos da diagonal principal da matriz **H** e h_{ij} são os elementos da i-ésima linha e j-ésima coluna da matriz **H**.

Matriz H - "chapéu"

- Ao usar MSRes, o estimador da variância do erro - σ^2, as estimativas das variâncias e covariâncias são definidas por:

$$\hat{Var}(e_i) = MSRes(1 - h_{ii})$$

e

$$\hat{Cov}(e_i, e_j) = -h_{ij} MSRes$$

Resíduo Studentizado

- Os resíduos e_i podem ter variâncias substancialmente diferentes.

- Por isso, é importante considerar a magnitude de cada resíduo relativa ao desvio padrão estimado de cada resíduo. Dando origem ao Resíduo Studentizado:

$$r_i = \frac{e_i}{\sqrt{MSRes(1 - h_{ii})}}$$

- O resíduo studentizado tem variância constante, $Var(r_i) = 1$.

Resíduo excluído

- Se a i-ésima observação, Y_i, é realmente incomum, discrepante, o modelo de regressão ajustado usando todas as observações pode ser influenciado por essa observação.
- Uma outra medida para verificar se a i-ésima observação, Y_i, é discrepante, é o resíduo excluído definido por:

$$e_{(i)} = Y_i - \hat{Y}_{i(i)},$$

  também conhecido como erro de predição PRESS para o i-ésimo caso.
- Uma expressão equivalente para $e_{(i)}$ é:

$$e_{(i)} = \frac{e_i}{(1 - h_{ii})}$$

- Dessa forma, o critério PRESS pode ser obtido por:

$$PRESS_{p+1} = \sum_{i=1}^{n} \left(\frac{e_i}{(1 - h_{ii})} \right)^2$$

Resíduo excluído

- A variância do resíduo excluído é dada por:

$$
\begin{aligned}
Var(e_{(i)}) &= \frac{1}{(1-h_{ii})^2} Var(e_i) = \frac{1}{(1-h_{ii})^2}[\sigma^2_{(i)}(1-h_{ii})] \\
&= \frac{\sigma^2_{(i)}}{(1-h_{ii})}
\end{aligned}
$$

- Consequentemente, a variância estimada é definida por:

$$
Var(\hat{e}_{(i)}) = \frac{MSRes_{(i)}}{(1-h_{ii})},
$$

em que $MSRes_{(i)}$ é o MSRes do modelo quando a i-ésima observação é excluída.

Resíduo excluído studentizado

- Seguindo a definição do resíduo studentizado, o resíduo excluído studentizado é definido por:

$$t_i = \frac{e_{(i)}}{\sqrt{\hat{Var}(e_{(i)})}} \sim \text{t-Student}_{n-1-(p+1)}$$

Identificando observações discrepantes em relação à X

- Pontos potencialmente distantes tem impacto nas estimativas dos parâmetros, erro padrão, valores preditos e estatísticas do modelo. A matriz **H**

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

  é importante para detectar observações influentes.
- Os elementos h_{ii} da matriz **H** são definidos por:

$$h_{ii} = \mathbf{X}_i^{'}(\mathbf{X}^{'}\mathbf{X})^{-1}\mathbf{X}_i,$$

  em que $\mathbf{X}_i$ é a i-ésima linha da matriz **X**. A diagonal da matriz **H** é uma medida padronizada da distância da i-ésima observação do centro do espaço de X.

- Valor grande de h_{ii} indica que a i-ésima observação está distante do centro das observações e que essa observação pode ser considerada um ponto de alavanca (leverage point).

- Outra forma de verificar ponto de alavanca é quando: $h_{ii} > 2\bar{h}$, em que $\bar{h} = \sum_{i=1}^{n} h_{ii}/n$.

- Após identificar valores discrepantes com respeito aos valores de Y ou X, o próximo passo é verificar se essas observações são ou não observações influentes.
- Medidas para identificar observações influentes são:
  - DFFITS
  - Distância de Cook
  - DFBETAS

DFFITS

- Medida da influência que a i-ésima observação tem sobre o valor ajustado $\hat{Y}_i$ é dada por:

$$(\text{DFFITS})_i = \frac{\hat{Y}_i - \hat{Y}_{i(i)}}{\sqrt{MSRes_{(i)} h_{ii}}}$$

em que:

- $\hat{Y}_i$ valor ajustado para o i-ésimo caso quando todas as n observações são usadas no ajuste do modelo
- $\hat{Y}_{i(i)}$ valor ajustado para o i-ésimo caso quando o i-ésimo caso é omitido no ajuste do modelo
- $MSRes_{(i)}$ quando o i-ésimo caso é omitido no ajuste do modelo.

DFFITS

- O valor de DFFITS pode ser obtido usando apenas o resultado do ajuste do modelo com todos os dados por meio de:

$$(\text{DFFITS})_i = t_i \left(\frac{h_{ii}}{1 - h_{ii}} \right)^{1/2}$$

- Para identificar se uma observação é influente:
- Se $|(\text{DFFITS})_i| > 1$, para conjunto de dados pequenos ou médios,
- Se $|(\text{DFFITS})_i| > 2\sqrt{p/n}$, para conjunto de dados grandes.

Distância de Cook

- Cook(1977,1979) sugeriu uma medida usando a distância ao quadrado entre todas as estimativas $\hat{\boldsymbol{\beta}}$, e a estimativa obtida ao exluir a i-ésima observação, $\hat{\boldsymbol{\beta}}_{(i)}$. Essa medida da distância pode ser expressa por:

$$D_i = \frac{(\hat{\boldsymbol{\beta}}_{(i)} - \hat{\boldsymbol{\beta}})' \mathbf{M} (\hat{\boldsymbol{\beta}}_{(i)} - \hat{\boldsymbol{\beta}})}{c},$$

em que $\mathbf{M}$ e c são usualmente $\mathbf{M} = (\mathbf{X}'\mathbf{X})$ e $c = (p+1)MSRes$. Logo:

$$D_i = \frac{(\hat{\boldsymbol{\beta}}_{(i)} - \hat{\boldsymbol{\beta}})' (\mathbf{X}'\mathbf{X}) (\hat{\boldsymbol{\beta}}_{(i)} - \hat{\boldsymbol{\beta}})}{(p+1)MSRes}.$$

Distância de Cook

- A distância de Cook, medida da influência que a i-ésima observação tem sobre todos os n valores ajustados, também pode ser definida por:

$$D_i = \sum_{i=1}^{n} \frac{(\hat{Y}_i - \hat{Y}_{i(i)})^2}{(p+1)MSRes}$$

- A distância de Cook pode ser obtida usando apenas o resultado do ajuste do modelo com todos os dados por meio de:

$$D_i = \frac{e_i^2}{pMSRes}\left[\frac{h_{ii}}{(1-h_{ii})^2}\right]$$

- Valores de $D_i > 1$ são consideradas observações influentes.

DFBETAS

- Medida da influência que a i-ésima observação tem sobre cada um dos coeficientes de regressão estimados é dada por:

$$(\text{DFBETAS})_{k(i)} = \frac{\hat{\beta}_k - \hat{\beta}_{k(i)}}{\sqrt{MSRes_i C_{kk}}}$$

em que:

- $\hat{\beta}_k$ coeficiente de regressão estimado considerando todos os n casos no ajuste do modelo
- $\hat{\beta}_{k(i)}$ coeficiente de regressão estimado considerando que o i-ésimo caso é omitido no ajuste do modelo
- $MSRes_i$ é o MSRes quando o i-ésimo caso é omitido no ajuste do modelo
- C_{kk} o k-ésimo elemento da diagonal da matriz $(\mathbf{X}'\mathbf{X})^{-1}$.

DFBETAS

Interpretação da medida DFBETAS:

- Sinal - indica se a inclusão do i-ésimo caso leva ao aumento ou a diminuição da estimativa do coeficiente de regressão
- Valor grande de $(\text{DFBETAS})_{k(i)}$ indica grande influência do i-ésimo caso na estimativa do k-ésimo coeficiente de regressão.
- Se $|(\text{DFBETAS})_{k(i)}| > 1$, para conjunto de dados pequenos ou médios, $\Rightarrow$ observação é influente
- Se $|(\text{DFBETAS})_{k(i)}| > 2\sqrt{p/n}$, para conjunto de dados grandes $\Rightarrow$ observação é influente.

Diagnóstico de Multicolinearidade

Diagnóstico de Multicolinearidade

Diagnóstico informal

- Coeficientes de correlação alto entre pares de variáveis explicativas.
- Mudanças grandes nas estimativas dos coeficientes de regressão quando:
  - uma variável explicativa é adicionada ou retirada do modelo;
  - uma observação é alterada ou apagada.
- Nos testes individuais sobre os coeficientes de regressão para variáveis explicativas importantes do modelo aceita-se $H_0 : \beta_k = 0$
- Coeficientes de regressão estimados com sinal algébrico oposto do que se espera através de considerações teóricas ou experiência anterior
- Intervalos de confiança para os coeficientes de regressão de variáveis explicativas importantes apresentam grande amplitude.

Diagnóstico de Multicolinearidade

Limitações:

- Estes métodos informais não fornecem medidas quantitativas do impacto da multicolinearidade e não podem identificar a natureza da multicolinearidade.

- Algumas vezes o comportamento observado pode ocorrer sem que exista de fato multicolinearidade.

Fator de inflação da Variância (VIF)

- Esse fator mede quanto a variância dos estimadores de mínimos quadrados são influenciadas quando comparada com variáveis explicativas que não são correlacionadas.
- Para entender esse fator, começa-se com a matriz de variâncias e covariâncias dos estimadores de mínimos quadrados:

$$Var(\boldsymbol{\beta}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$$

- Para medir o impacto da multicolinearidade, é útil utilizar o modelo de regressão padronizado definido em aulas anteriores para reduzir os erros de arredondamento no cálculo da matriz $(\mathbf{X}'\mathbf{X})^{-1}$.

Diagnóstico de Multicolinearidade

- No modelo transformado os coeficientes de regressão estimados $\hat{\beta}_k^*$ são padronizados e sua matriz de covariâncias é dada por:

$$Var(\boldsymbol{\beta}^*) = (\sigma^*)^2(\mathbf{r_{XX}})^{-1},$$

em que $\mathbf{r_{XX}}$ é a matriz de correlação simples entre os pares de variáveis X e $(\sigma^*)^2$ é a variância do erro do modelo transformado.

Diagnóstico de Multicolinearidade

- Porém, $(\mathbf{r_{XX}})^{-1}$ é o fator de inflação da variância (VIF) para β_k^*. Logo:

$$Var(\beta_k^*) = (\sigma^*)^2(\text{VIF})_k,$$

- Os elementos da diagonal de $(\text{VIF})_k$ é o fator de inflação da variância (VIF) para β_k^* e é igual a:

$$(\text{VIF})_k = (1 - R_k^2)^{-1},$$

em que R_k^2 é o coeficiente de determinação múltipla quando X_k é retirado do modelo que contém $(p = (p+1) - 1)$ variáveis .

Diagnóstico de Multicolinearidade

- Se $R_k^2 = 0 \Rightarrow (\text{VIF})_k = 1$ e X_k não está correlacionada com as demais variáveis.

- Se $R_k^2 \neq 0 \Rightarrow (\text{VIF})_k > 1$ e X_k está correlacionada com as demais variáveis.

- Quando X_k está perfeitamente relacionada com as outras variáveis do modelo $\Rightarrow R_k^2 = 1$ e $(\text{VIF})_k$ será ilimitado (valor grande).

Indicador de Multicolinearidade

- Se o máximo dos $(\text{VIF})_k > 10 \Rightarrow$ Multicolinearidade está influenciando as estimativas dos parâmetros.
- Se $R_k^2 = 0 \Rightarrow (\text{VIF})_k = 1 \Rightarrow$ Não existe Multicolinearidade
- Outra forma de verificar é obter:

$$(\bar{\text{VIF}})_k = \frac{\sum_{k=1}^{p} (\text{VIF})_k}{p}$$

- Se $(\bar{\text{VIF}})_k$ for um valor consideravelmente maior que $1 \Rightarrow$ Multicolinearidade está influenciando as estimativas dos parâmetros.