MODELO LINEAR CLÁSSICO

Frederico Machado Almeida
frederico.almeida@unb.br

Departamento de Estatística
Instituto de Exatas
Universidade de Brasília

Regressão Linear Simples

- Métodos estatísticos que objetivam analisar a relação entre duas variáveis são frequentes na literatura. Com esses métodos procura-se por uma função de X que explique as variações de Y, i.e., $Y \approx f(X)$.
- De forma geral, um modelo estatístico pode ser escrito da seguinte forma:

$$Y = f(X, \beta_0, \beta_1) + \epsilon \tag{1}$$

- Como iremos ver no decorrer do curso, existem diversas maneiras de especificar essas componentes. No modelo em (1), assumimos que X é uma quantidade `fixa e medida sem erro` (ou controlada), e Y é uma quantidade `aleatória`.

Regressão Linear Simples

Uma regressão linear simples tem como objetivo aproximar uma variável de resposta Y através de uma função linear de uma variável de interesse, ou seja,

$$Y = f(X, \beta_0, \beta_1) + \epsilon = \beta_0 + \beta_1 X + \epsilon,$$

sendo $f(\cdot)$ `uma função linear nos parâmetros`, e ϵ o termo de erro no qual assume que:

1. $\mathbb{E}(\epsilon) = 0$
2. $\text{Var}(\epsilon) = \sigma^2$ (constante ou, homoscedasticidade)
3. $\text{Cov}(\epsilon_i, \epsilon_j) = 0$ para $\forall i \neq j$
4. A relação entre X e Y deve ser linear
5. Os valores de X são fixos (controlados).

Em outras palavras, os erros tem média zero, variância constante e são não correlacionados.

Regressão Linear Simples

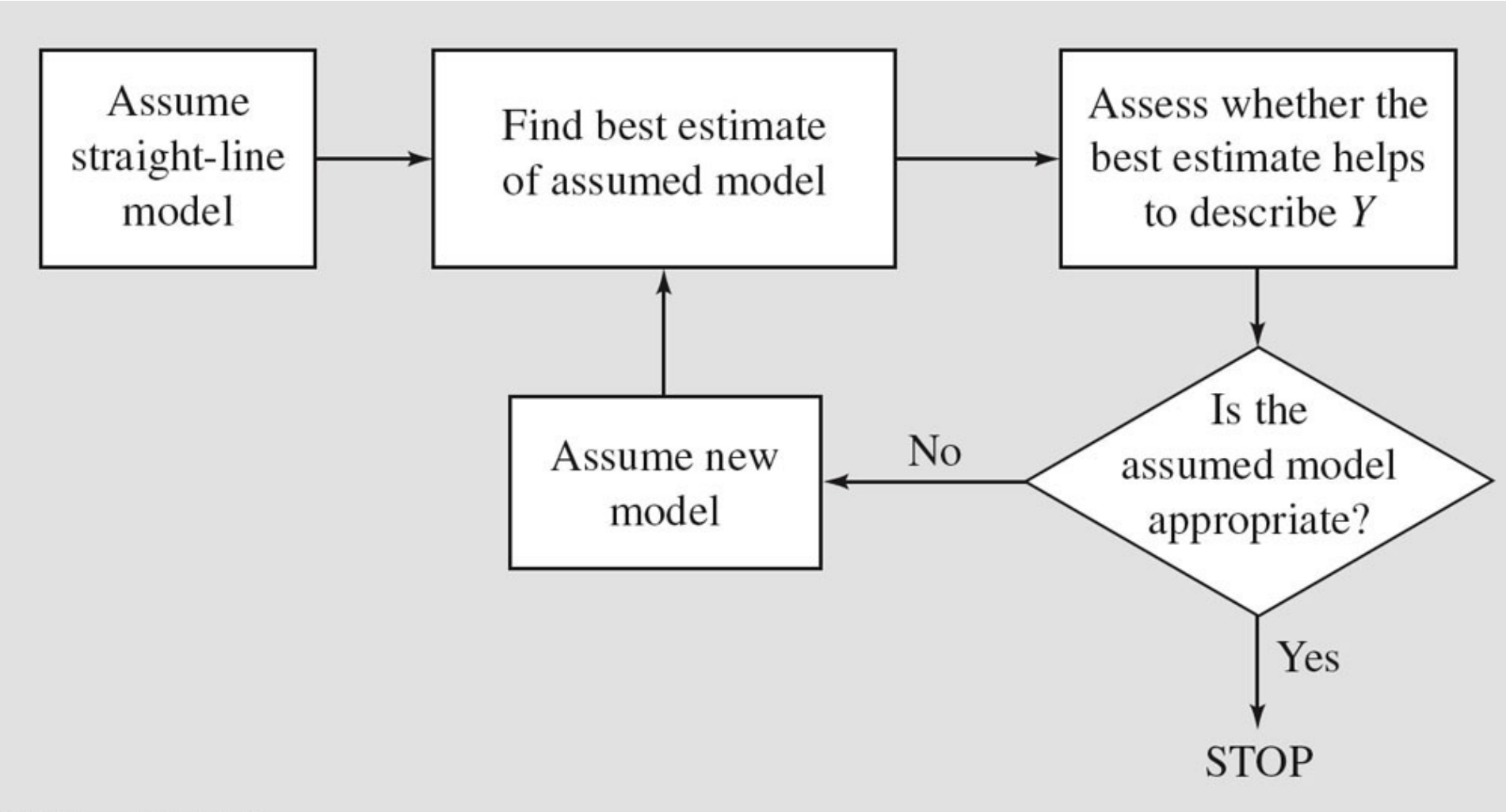
A variável preditora X pode vir de diversas fontes, como por exemplo:

- Inputs quantitativos (valores reais, medidas)
- Transformação de variáveis quantitativas (log ,$\sqrt{}$, etc)
- Inputs qualitativos ("dummy" e.x. gênero, escolaridade, classe social).

Dessa forma, um modelos de regressão consiste em 4 passos:

- Escolher o componente determinístico do modelo
- Utilizar os dados para estimar os parâmetros do modelo
- Especificar a distribuição do erro
- Avaliar o modelo estatístico.

Regressão Linear Simples



Regressão Linear Simples

- A função $f(\cdot)$ que caracteriza o modelo de regressão é dita ser linear, caso seja linear nos parâmetros, i.e.,

$$\frac{\partial}{\partial \beta_j} f(X, \beta_0, \beta_1) = h(X), \text{ com } j = 0, 1.$$

sendo $h(X)$ uma função que depende apenas dos dados observados.

Em suma, o modelo de regressão apresentado em (1) objetiva:

Estimação dos parâmetros: Consiste em utilizar a informação nos dados para fazer a estimação dos parâmetros do modelo de regressão.

Inferência: Ao ajustar o modelo de regressão linear, estamos interessados não só em estimar os parâmetros de regressão, mas também, realizar inferências sobre os mesmos. Como por exemplo, construção de intervalos de confiança, e os testes de hipóteses.

Regressão Linear Simples

Predição: Consiste em obter valores da variável dependente, Y usando valores de X que não foram considerados no ajuste do modelo.

Seleção de Variáveis: Consiste em manter no modelo apenas as variáveis relevantes (eliminar variáveis irrelevantes no modelo). Esta técnica é válida no caso da regressão múltipla.

O modelo estatístico da análise de regressão é tal que,

$$\mathbb{E}(Y|X = x) = \beta_0 + \beta_1 x$$

- Para $i \in \{1, \cdots, n\}$, assuma que $(x_1, y_1), \cdots, (x_n, y_n)$ são n combinações dos dados observados. Sendo x_i e y_i valores observados para X e Y, respectivamente.

Regressão Linear Simples

- O método mais utilizado é conhecido como método de mínimos quadrados, e é dados por:

$$\begin{aligned} SQE(\beta) &= \sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} (Y_i - f(X_i, \beta))^2 \\ &= \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2 \end{aligned}$$

- Assim, os estimadores para os parâmetros β_0 e β_1, digamos $\hat{\beta}_0$ e $\hat{\beta}_1$, são aqueles que minimizam a $SQE(\beta)$. Note que, $\beta = (\beta_0, \beta_1)^\top$.
- Diferenciando a $SQE\,(\beta)$ em relação a β_0 e β_1 chegamos as equações normais:

Regressão Linear Simples

$$\begin{aligned}
\frac{\partial}{\partial \beta_0} SQE(\beta) &= -2 \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i) \\
\frac{\partial}{\partial \beta_1} SQE(\beta) &= -2 \sum_{i=1}^{n} X_i (Y_i - \beta_0 - \beta_1 X_i) .
\end{aligned}$$

- Igualando as derivadas parciais a 0, e resolvendo o sistema de equações resultantes, obtemos facilmente os estimadores $\hat{\beta}_0$ e $\hat{\beta}_1$ que minimizam $SQE(\beta)$. Ou seja,

$$\begin{aligned}
\sum_{i=1}^{n} \left(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i \right) &= 0 \\
\sum_{i=1}^{n} X_i \left(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i \right) &= 0.
\end{aligned}$$

Regressão Linear Simples

- Por fim, obtemos:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}\left(X_i-\bar{X}\right)\left(Y_i-\bar{Y}\right)}{\sum_{i=1}^{n}\left(X_i-\bar{X}\right)^2} = \frac{\sum_{i=1}^{n}X_iY_i - n\bar{X}\bar{Y}}{\sum_{i=1}^{n}X_i^2 - n\bar{X}^2} = \frac{S_{XY}}{S_{XX}}$$

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1\bar{X} = \bar{Y} - \bar{X}\frac{S_{XY}}{S_{XX}}.$$

- Calculado as derivadas da segunda ordem obtemos:

$$\frac{\partial^2 SQE}{\partial\beta_0^2} = 2n$$

$$\frac{\partial^2 SQE}{\partial\beta_1^2} = 2\sum_{i=1}^{n}X_i^2$$

$$\frac{\partial^2 SQE}{\partial\beta_0\beta_1} = 2\sum_{i=1}^{n}X_i$$

Regressão Linear Simples

Calculando o determinante das derivadas segundas obtemos:

$$\left|\frac{\partial^2 SQE(\beta)}{\partial\beta\beta^\top}\right| = \begin{vmatrix} \frac{\partial^2 SQE}{\partial\beta_0^2} & \frac{\partial^2 SQE}{\partial\beta_0\beta_1} \\ \frac{\partial^2 SQE}{\partial\beta_0\beta_1} & \frac{\partial^2 SQE}{\partial\beta_1^2} \end{vmatrix} = \begin{vmatrix} 2n & 2\sum_{i=1}^{n} X_i \\ 2\sum_{i=1}^{n} X_i & 2\sum_{i=1}^{n} X_i^2 \end{vmatrix}.$$

Observe que,

$$\left|\frac{\partial^2 SQE(\beta)}{\partial\beta\beta^\top}\right| = 4n\sum_{i=1}^{n} X_i^2 - 4\left(\sum_{i=1}^{n} X_i\right)^2 = 4n\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2 \geq 0.$$

Assim, $\hat{\beta}_0$ e $\hat{\beta}_1$ são os mínimos globais.

Regressão Linear Simples

- Feito isso, a equação da reta que minimiza a distância entre Y_i e $\hat{Y}_i$, ou seja, a reta de regressão estimada pelo método dos mínimos quadrados é:

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i \tag{2}$$

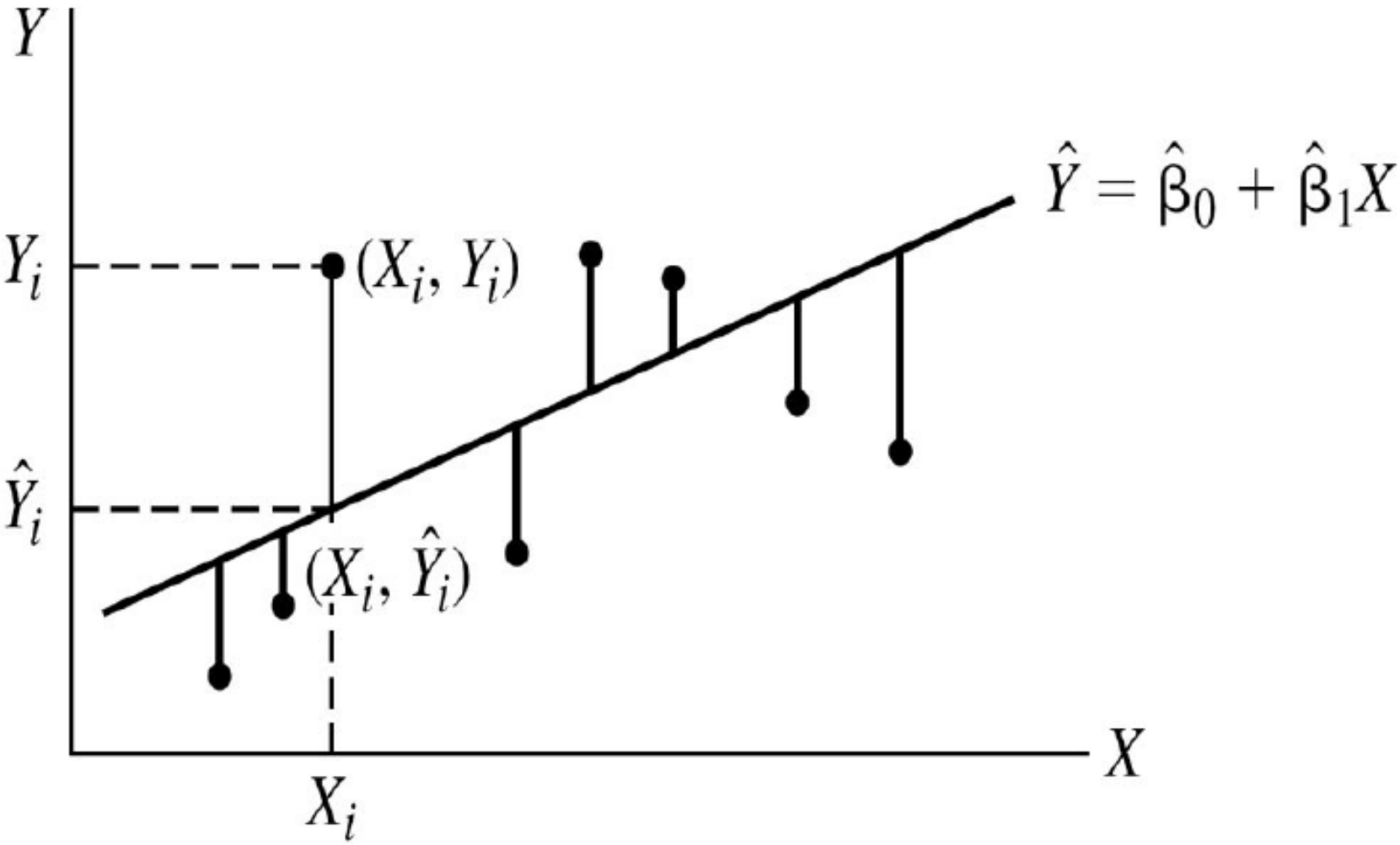
- Com base nos estimadores dos mínimos quadrados, é possível observar que, o $\left(\bar{X}, \bar{Y}\right)$ sempre cai no modelo ajustado.

- A distância entre Y_i e $\hat{Y}_i$ é comummente conhecida como resíduo, dada por:

$$e_i = Y_i - \hat{Y}_i = Y_i - \left(\hat{\beta}_0 + \hat{\beta}_1 X_i\right),$$

sendo e_i um estimador para ϵ_i.

Regressão Linear Simples



Regressão Linear Simples

- Desta forma, segue então que

$$SQE(\beta) = \sum_{i=1}^{n} \left(Y_i - f(X_i, \hat{\beta}) \right)^2 = \sum_{i=1}^{n} \left(Y_i - \hat{Y}_i \right)^2 = \sum_{i=1}^{n} e_i^2.$$

Propriedades dos resíduos:

- Conforme mostrado anteriormente, sendo $\hat{\beta}_0$ e $\hat{\beta}_1$ estimadores dos mínimos quadrados para β_0 e β_1 então, temos que:

$$\left. \frac{\partial^2 SQE(\beta)}{\partial \beta \beta^\top} \right|_{\beta = \hat{\beta}} = \mathbf{0}, \text{ portanto,}$$

- $\sum_{i=1}^{n} e_i = \sum_{i=1}^{n} \left(Y_i - (\hat{\beta}_0 + \hat{\beta}_1 X_i) \right) = 0$
- $\sum_{i=1}^{n} X_i e_i = \sum_{i=1}^{n} X_i \left(Y_i - (\hat{\beta}_0 + \hat{\beta}_1 X_i) \right) = \sum_{i=1}^{n} X_i \left(Y_i - \hat{Y}_i \right) = 0$

Regressão Linear Simples

- Até esse momento ainda não usamos a suposição de nenhuma distribuição para o erro.
- Conforme as suposição do erro dos itens (1)-(3) (`Slide 3`), comummente adota-se que $\epsilon_i \sim N\left(0, \sigma^2\right)$.
- A suposição de normalidade dos erros garante que todas as demais suposições do modelo sejam atendidas, e possibilita inferência sobre a modelagem.
- E portanto, com a suposição de normalidade temos que σ^2 pode ser estimado da seguinte forma:

$$\hat{\sigma}^2 = QME = \frac{1}{n-2}\sum_{i=1}^{n} e_i^2.$$

Regressão Linear Simples

Os resultados anteriores nos permitem afirmar que (verifique!):

- $\hat{\beta}_0$ e $\hat{\beta}_1$ são estimadores não-viesados para β_0 e β_1, respectivamente.
- A variância de $\hat{\beta}_0$ e $\hat{\beta}_1$ são dadas por: (i) $\text{Var}(\hat{\beta}_1) = \frac{\sigma^2}{S_{xx}}$

$$(ii)\ \text{Var}(\hat{\beta}_0) = \sigma^2\left(\frac{1}{n} + \frac{\bar{X}^2}{S_{xx}}\right), (iii)\ \text{Var}(\hat{Y}_i) = \sigma^2\left(\frac{1}{n} + \frac{\left(X_i - \bar{X}\right)^2}{S_{xx}}\right)$$

- Assumindo que o erro possui distribuição normal independente, podemos executar testes de hipóteses para os coeficientes β_0 e β_1. Ou seja, para todo $j = 0, 1$ e $i = 1, \cdots, n$ segue que:

$$\hat{\beta}_j \sim \text{N}\left(\beta_j, \text{Var}(\hat{\beta}_j)\right) \text{ e } \hat{Y}_i \sim \text{N}\left(f(X_i, \boldsymbol{\beta}), \text{Var}\left(\hat{Y}_i\right)\right).$$

Regressão Linear Simples

Testes de hipóteses para os parâmetros de regressão: Para $j = 0, 1$ as hipóteses de teste são:

$$H_0 : \beta_0 = 0 \text{ vs } H_0 : \beta_0 \neq 0,$$

seguimos de forma equivalente a testes de hipóteses normais. Assim, segue que, para $j = 0$ então,

$$t_{obs} = \frac{\hat{\beta}_0 - 0}{\sqrt{\text{Var}\left(\hat{\beta}_0\right)}} = \frac{\hat{\beta}_0}{\sqrt{QME\left(\frac{1}{n} + \frac{\bar{X}^2}{S_{XX}}\right)}} \sim t_{n-2}.$$

Portanto, a H_0 será rejeitada sempre que $|t_{obs}| > t_{(\alpha/2, n-2)}$.

Regressão Linear Simples

De forma similar, para $j = 1$ as hipóteses de teste são:

$$H_0 : \beta_1 = 0 \text{ vs } H_0 : \beta_1 \neq 0, \text{assim, segue que:}$$

$$t_{obs} = \frac{\hat{\beta}_1 - 0}{\sqrt{\text{Var}\left(\hat{\beta}_1\right)}} = \frac{\hat{\beta}_1}{\sqrt{\frac{QME}{S_{xx}}}} \sim t_{n-2}.$$

De igual forma, a H_0 será rejeitada sempre que $|t_{obs}| > t_{(\alpha/2, n-2)}$.

- Caso a amostra observada forneça bases que suportam a H_0, i.e., caso a hipótese nula não seja rejeitada, concluímos que o fator X_i não é importante para explicar as variações do Y_i.
- No caso em que temos mais de uma variável explicativa no modelo, podemos analisar a significância dos múltiplos coeficientes de regressão ao mesmo tempo.

Regressão Linear Simples

Exemplo 1: Um motor de foguete é fabricado unindo um propelente de ignição e um propelente sustentador dentro de um invólucro de metal. A resistência ao cisalhamento da ligação entre os dois tipos de propelente é uma importante característica de qualidade.
Suspeita-se que a resistência ao cisalhamento esteja relacionada com a idade em semanas do lote de propelente sustentador. Para tal, 20 observações sobre a resistência ao cisalhamento e a idade do lote correspondente de propelente foram coletadas e são mostradas no ficheiro `SimpleLM-ex1.txt`

ANÁLISE DE VARIÂNCIA

- Com base no modelo, é possível decompor a variação total $\left(Y_i - \bar{Y}\right)$ em dois componentes: variação devida ao erro $\left(Y_i - \hat{Y}_i\right)$ e, variação devida à reta de regressão $\left(\hat{Y}_i - \bar{Y}\right)$.
- A decomposição da variação total para todas as observações é comummente apresentada em termos quadráticos. Isto é,

$$\sum_{i=1}^{n}\left(Y_i - \bar{Y}\right)^2 = \sum_{i=1}^{n}\left(\hat{Y}_i - \bar{Y}\right)^2 + \sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2, \text{ ou seja,}$$

$$\text{SQT} = \text{SQReg} + \text{SQE}.$$

ANÁLISE DE VARIÂNCIA

- Portanto, pode-se mostrar que (Verifique!)
  1. $\mathbb{E}(SQT) = \sigma^2(n-1) + \beta_1^2 S_{xx}$
  2. $\mathbb{E}(SQReg) = \sigma^2 + \beta_1^2 S_{xx}$
  3. $\mathbb{E}(SQE) = \sigma^2(n-2)$.
- Os detalhes da análise de variância (ANOVA), serão apresentados posteriormente.

ANÁLISE DOS RESÍDUOS

Depois que a equação de regressão dos mínimos quadrados é obtida, uma série de questões devem ser feitas:

1. Até que ponto a equação de regressão ajusta aos dados disponíveis?
2. O modelo obtido pode ser usado para predizer valores da variável resposta?
3. As suposições básicas como: homoscedasticidade, normalidade, independência dos erros são cumpridas? Se não, qual é a magnitude da sua violação?

Todos estes problemas devem ser investigados antes que o modelo seja adotado para fazer previsões.

ANÁLISE DOS RESÍDUOS

Análise de Resíduos

- Existem diferentes métodos que podem ser usados para avaliar a qualidade de ajuste do modelo apresentado em (2). Quase todos os métodos são baseados na análise dos resíduos.

- Ao fazer uma análise dos resíduos temos por objetivo avaliar se:

  1. A reta de regressão é linear
  2. Os termos de erro tem variância constante
  3. Os termos de erros são independentes
  4. Há influência (no ajuste) de algumas observações discrepantes
  5. Os erros são normais
  6. Possível omissão de uma variável preditora no modelo.

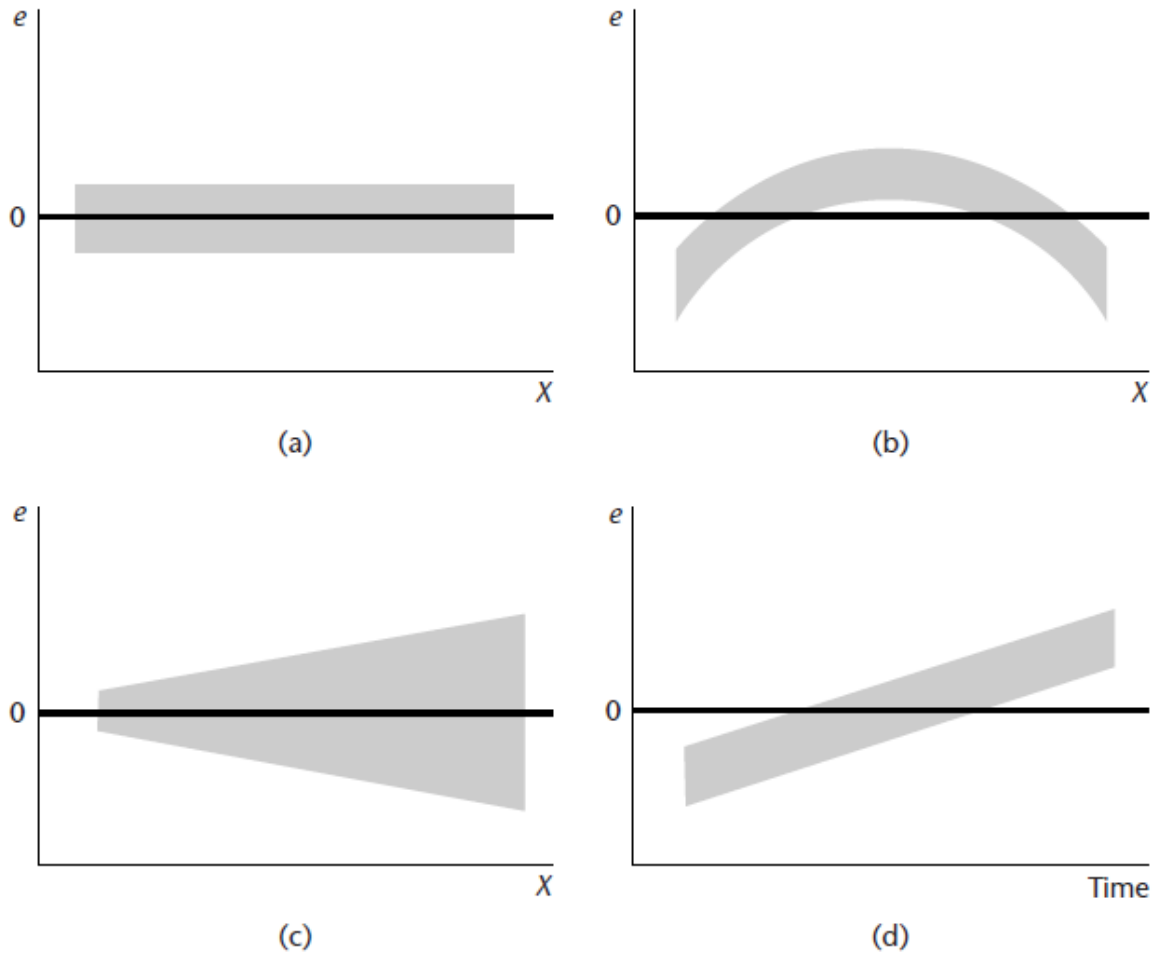
ANÁLISE DOS RESÍDUOS

- A análise dos resíduos pode ser feita usando as `técnicas informais` (métodos gráficos) e as `técnicas formais` (testes estatísticos).
- As técnicas informais são baseadas na análise dos seguintes gráficos:
  1. Gráfico dos resíduos vs $X_i's$
  2. Gráfico dos resíduos absolutos vs $X_i's$
  3. Gráfico dos resíduos vs $\hat{Y}_i's$
  4. Gráfico dos resíduos vs tempo/índice
  5. Boxplot dos resíduos
  6. Gráfico dos quantis normais (Q-Q plots), etc.

ANÁLISE DOS RESÍDUOS

Source: Applied linear regression models (4th Ed.), Kutner, Nachtsheim & Neter, 2004 Chap 3

FIGURE 3.4
Prototype Residual Plots.



ANÁLISE DOS RESÍDUOS

Pontos de Alavancagem (Leverage points)

- São observações que exercem uma considerável influência no modelo estimado. Podendo ser classificados em "`bons`" ou "`ruins`" dependendo da magnitude da sua influência/impacto no modelo.
- Um ponto de alavancagem é uma observação no qual o respectivo valor no `eixo das abcissas` (X) está distante das demais.
- É importante estudar os pontos de alavancagem porque, a sua presença pode afetar as estimativas dos mínimos quadrados e consequentemente a qualidade de ajuste do modelo.

ANÁLISE DOS RESÍDUOS

- O ponto de alavancagem será classificado como `ruim`, se além de estar distante das demais observações no eixo das abcissas, estiver também, distante das demais observações, porém, com relação ao eixo Y.
- O ponto de alavancagem será `bom` se a observação estiver distante das demais apenas no eixo das abcissas.
- Para tal, é importante estabelecer uma regra numérica que possa nos permitir identificar se uma observação com valor X_i afastado das demais é ou não um ponto de influência.

ANÁLISE DOS RESÍDUOS

- Com base na equação do modelo ajustado em (2), temos que $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$, onde

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}, \quad \text{e} \quad \hat{\beta}_1 = \frac{S_{XY}}{S_{XX}} = \sum_{j=1}^{n} K_j Y_j \text{ com } K_j = \frac{\left(X_j - \bar{X}\right)}{\sum_{j=1}^{n} \left(X_j - \bar{X}\right)^2}.$$

- Substituindo tais estimativas no modelo em (2) obtemos:

$$\begin{aligned} \hat{Y}_i &= \bar{Y} + \hat{\beta}_1 \left(X_i - \bar{X}\right) = \frac{1}{n} \sum_{j=1}^{n} Y_j + \left(X_i - \bar{X}\right) \sum_{j=1}^{n} K_j Y_j \\ &= \sum_{j=1}^{n} \left[\frac{1}{n} + \frac{\left(X_i - \bar{X}\right)\left(X_j - \bar{X}\right)}{\sum_{j=1}^{n} \left(X_j - \bar{X}\right)^2} \right] Y_j = \sum_{j=1}^{n} h_{ij} Y_j \end{aligned}$$

ANÁLISE DOS RESÍDUOS

- Sendo $h_{ij} = \frac{1}{n} + \frac{(x_i - \bar{x})(x_j - \bar{x})}{\sum_{j=1}^{n}(x_j - \bar{x})^2}$ valores de alavancagem.Ou seja, as entradas genéricas (i, j) da matriz projeção (hat matrix, $\mathbf{H}$).
- Observe que, com base na expressão apresentada anteriormente, h_{ij} depende apenas dos valores da variável independente.
- Podemos observar também, que os valores estimador $\hat{Y}_i$ podem ser reescritos como uma combinação linear dos valores de alavancagem, e os valores observados para a variável dependente. Isto é,

$$\hat{Y}_i = \sum_{j=1}^{n} h_{ij} Y_j = h_{ii} Y_i + \sum_{i \neq j} h_{ij} Y_j.$$

ANÁLISE DOS RESÍDUOS

Quando $i = j$, temos que h_{ii} são os elementos na diagonal de $\mathbf{H}$. Assim,

$$h_{ii} = \frac{1}{n} + \frac{\left(X_i - \bar{X}\right)^2}{\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2}.$$

Propriedades dos $h'_{ij}s$ (`verifique!`)

1. $0 \leq h_{ij} \leq 1$
2. $\sum_{j=1}^{n} h_{ij} = 1$
3. $\sum_{i=1}^{n} h_{ii} = 2$ (de forma geral, segue que $\sum_{i=1}^{n} h_{ii} = p$)
4. $\sum_{j=1}^{n} X_i h_{ij} = X_i$.

ANÁLISE DOS RESÍDUOS

- Uma observação (X_i, Y_i) na massa de dados pode ser considerado um ponto de alavancagem, se para todo $i \in \{1, \cdots, n\}$ obtivermos:

$$h_{ii} > 2 \times \texttt{Média}\ (h_{ii}) = 2 \times \frac{2}{n}$$

- No caso geral, um ponto será considerado de alavancagem quando

$$h_{ii} > 2 \times \frac{p}{n},$$

sendo p o número de coeficiente no modelo, incluindo o intercepto.

Como estratégias para tratar os pontos de alavancagem podemos:

- Remover o ponto de alavanca na massa de dados e ajustar novamente o modelo.
- Ajustar outros modelos capazes de acomodar o ponto classificado como de alavancagem (i.e., mudar a distribuição dos erros).

ANÁLISE DOS RESÍDUOS

- Fazer transformações$^{(*)}$
- Aumentar o modelo adicionando covariáveis, termos de interação e termos quadráticos.
- Adicionar observações para aumentar o poder do ajuste.

$^{(*)}$ `Pontos influentes comumente são causados por outliers nas covariáveis, e portanto transformações nas covariáveis e/ou resposta comumente corrigem esse problema.`

ANÁLISE DOS RESÍDUOS

Resíduos padronizados

- Conforme apresentado anteriormente, os valores estimados $\hat{Y}_i$ podem ser rescritos da seguinte forma $\hat{Y}_i = \sum_{j=1}^{n} h_{ij} Y_j$. Assim, os resíduos serão dados por,

$$e_i = Y_i - \hat{Y}_i = Y_i - \sum_{j=1}^{n} h_{ij} Y_j = (1 - h_{ii}) Y_i - \sum_{i \neq j} h_{ij} Y_j$$

- Segue então que (`verifique!`)
  1. $\mathbb{E}(e_i) = 0$
  2. $\text{Var}(e_i) = (1 - h_{ii})\sigma^2$
  3. $\text{Var}(\hat{Y}_i) = h_{ii}\sigma^2$.

ANÁLISE DOS RESÍDUOS

- O item (2) mostra claramente que, na presença de pontos de alavancagem, a `suposição de variância constante não será verificada`.
- O problema dos resíduos apresentarem variâncias diferentes para cada valor i, pode ser contornado por meio da padronização dos resíduos. Isto é,

$$r_i = \frac{e_i - 0}{\sqrt{(1 - h_{ii})\,\sigma^2}} = \frac{e_i}{s\sqrt{(1 - h_{ii})}}, \tag{3}$$

  onde $s = \sqrt{QME}$.
- Quando existem pontos de alto poder de alavancagem, o gráfico dos resíduos padronizados é `mais informativo` que o dos resíduos usuais.

ANÁLISE DOS RESÍDUOS

- Quando não há pontos de alavancagem, em geral, o padrão dos gráficos obtidos nos casos em que são usados os resíduos usuais ou padronizados são similares.
- Permitem analisar o quão o quão distante em termos de desvios-padrão, um determinado ponto está com relação ao modelo ajustado.
- Permitem identificar potenciais observações atípicas. Assim, observações com $-2 \leq r_i \leq +2$ são classificadas como usuais, e se $r_i \notin (-2, +2)$ são classificadas como outliers.
- Quando o ponto de alavancagem é bom, seu resíduo padronizado estará entre $(-2, +2)$. Caso contrário, $r_i \notin (-2, +2)$.

ANÁLISE DOS RESÍDUOS: Pontos de Influência

Métodos adicionais como por exemplo,

- Resíduos estudentizado externamente (RStudent or Jackknifed residuals)
- Distância de Cook
- DFBetaS
- DFFITS, entre outros, podem ser igualmente considerados.

Aplicação no R