# Análise de Regressão

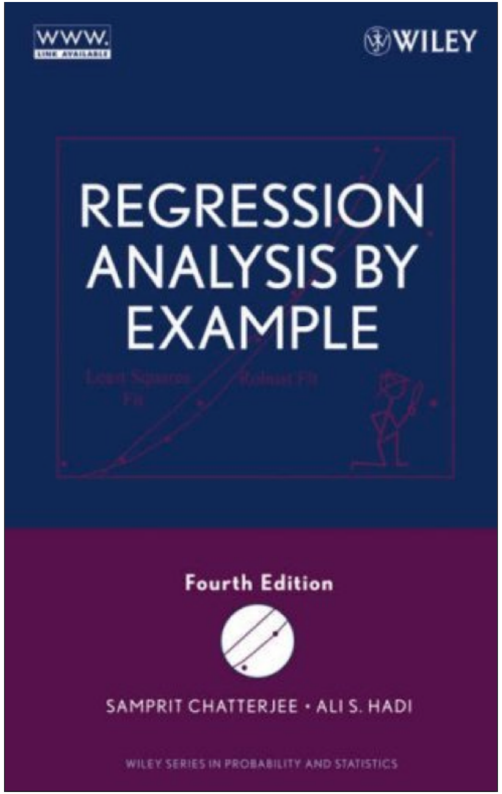
Tópicos Avançados em Avaliação de Desempenho

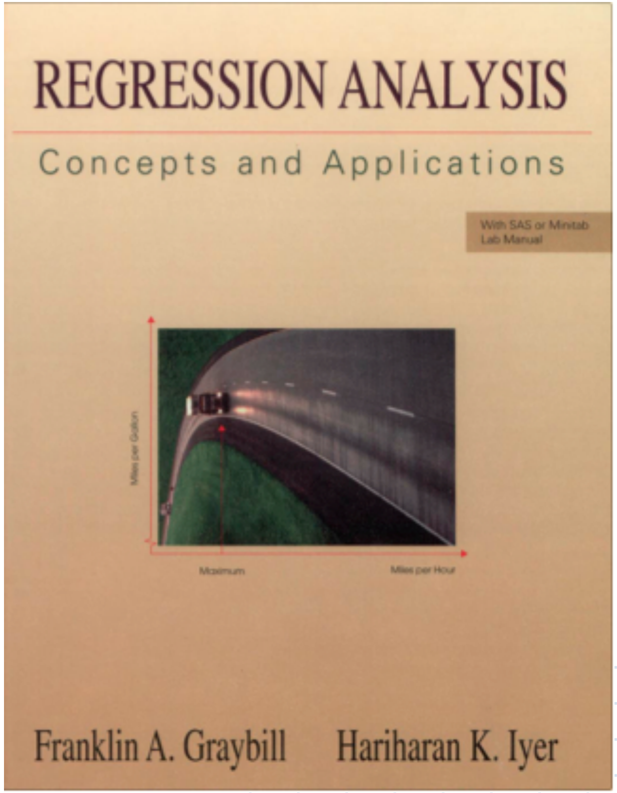
Cleber Moura
Edson Samuel Jr

# Agenda

- Introdução
- Passos para Realização da Análise
- Modelos para Análise de Regressão
- Regressão Linear Simples
- Regressão Linear Múltipla

# Referências



# O que é Análise de Regressão?

(GRAYBILL & IYER, 1994, p. 1)

- Área da Estatística que lida com métodos para investigação da existência de associações entre várias quantidades observáveis e, se presente, a natureza das associações.

(CHATTERJEE & HADI, 2006, p. 1)

- É um método simples para investigação de relacionamentos funcionais entre variáveis.

- Exemplo: Podemos examinar como o consumo do cigarro está relacionado a diversas variáveis socioeconomicas e demográficas, como: idade, educação, salário e preço do cigarro.

# Relação entre as variáveis

- O relacionamento entre as variáveis é expressado na forma de uma equação ou modelo, conectando a **variável de resposta** (ou dependente) com uma ou mais **variáveis preditoras** (ou explicativas).

- No exemplo anterior:
    - Variável de resposta: **consumo de cigarro**
    - Variáveis preditoras: **idade, educação, salário e preço do cigarro**

# Áreas de aplicação

- Possui inúmeras áreas de aplicação, como:
  - economia, finanças, negócios, direito, meteorologia, medicina, biologia, química, engenharia, física, educação, esportes, história, sociologia, psicologia, dentre outras.

- A Análise de Regressão é uma das mais utilizadas ferramentas estatísticas, por que provê métodos simples para **estabelecer uma relação funcional entre variáveis** e para **predição da variável de resposta**.

# Passos da Análise

- Definição do problema
- Seleção das variáveis potenciais e relevantes
- Coleta de dados
- Especificação do modelo
- Escolha do método e adequação do modelo
- Validação e análise crítica do modelo
- Uso dos modelos adotados para solucionar o problema exposto

# Passo 1: Definição do problema

- Inicia com a formulação do problema e determinação das questões a serem tratadas pela análise
- É o primeiro e um dos mais importantes passos

- **Um problema mal definido ou questões mal formuladas podem levar a**:
  - desperdício de tempo/esforço;
  - seleção de variáveis irrelevantes; e
  - escolha de um método estatístico ou modelo de análise não apropriados.

# Cenário 1

- Queremos determinar se uma empresa faz ou não discriminação entre os grupos de funcionários (homens e mulheres)
  - Os dados sobre salário, qualificações e sexo estão disponíveis nos registros da empresa.
- Para identificar ocorrência de discriminação, devemos responder:
  - Na média, as mulheres são menos pagas do que homens igualmente qualificados?
  - Na média, as mulheres são mais qualificadas do que homens igualmente pagos?

# Passo 2: Seleção das variáveis

- Neste passo são selecionadas o conjunto de variáveis que explicam ou predizem a variável de resposta.

- A variável de resposta é indicada por **Y** e a variável preditora ou explicativa por $\mathbf{X_1}$, $\mathbf{X_2}$,... $\mathbf{X_p}$, onde **p** indica o número de variáveis preditoras.

# Cenário 1

- Para responder a questão:
  "Na média, as mulheres são menos pagas que homens igualmente qualificados?"

Escolhemos **salário** como variável de resposta, e **qualificação** e **sexo** como variáveis preditoras.

- Para questão:
  "Na média, as mulheres são mais qualificadas que homens igualmente pagos?"

Escolhemos **qualificação** como variável de resposta, e **sexo** e **salário** como variáveis preditoras.

# Passo 3: Coleta de dados

- Este passo consiste na observação de **n** "alvos" e cada uma das **n** observações consiste em medições para cada uma das variáveis relevantes;
- As variáveis podem ser qualitativas e quantitativas

| Número da Observação | Resposta Y | Preditoras | | | |
|---|---|---|---|---|---|
| | | $X_1$ | $X_2$ | ... | $X_p$ |
| 1 | $Y_1$ | $X_{11}$ | $X_{12}$ | ... | $X_{1p}$ |
| 2 | $Y_2$ | $X_{21}$ | $X_{22}$ | ... | $X_{2p}$ |
| 3 | $Y_3$ | $X_{31}$ | $X_{32}$ | ... | $X_{3p}$ |
| ... | ... | ... | ... | ... | ... |
| n | $Y_n$ | $X_{n1}$ | $X_{n2}$ | ... | $X_{np}$ |

# Passo 4: Especificação do Modelo

- Neste passo é definido o **modelo de regressão** que será **utilizado** para **relacionar** a **variável de resposta** ao conjunto de **variáveis preditoras**

- Regressão ***Simples*** e ***Múltipla***:
    - Determinado pelo número de variáveis preditoras

- Regressão ***Univariável*** e *Multivariável*
    - Determinado pelo número de variáveis de resposta

# Passo 5: Adequação do modelo

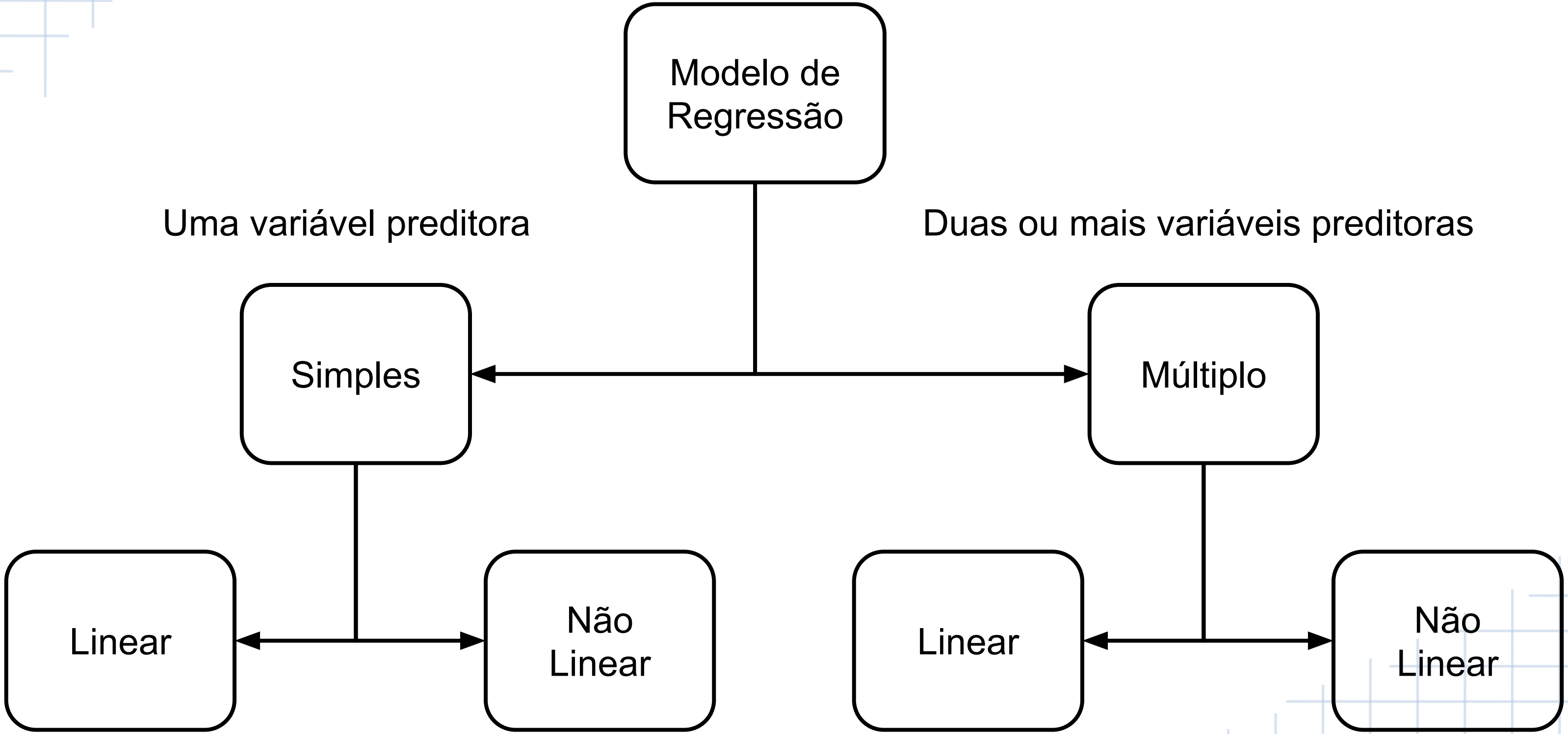
- Com base no modelo definido e nos dados coletados, o próximo passo é estimar os parâmetros de regressão utilizando algum método de estimativa.

- O mais utilizado é o **método dos mínimos quadrados**.

# Passo 6: Validação e Análise Crítica

- A validade do método estatístico depende de premissas feitas sobre os dados e modelo
- A análise de regressão é vista como um processo iterativo, onde as saídas são utilizadas para diagnosticar, validar, criticar, e possivelmente modificar as entradas
- O processo deve ser repetido até obter saídas satisfatórias

# Modelos de Regressão Univariável



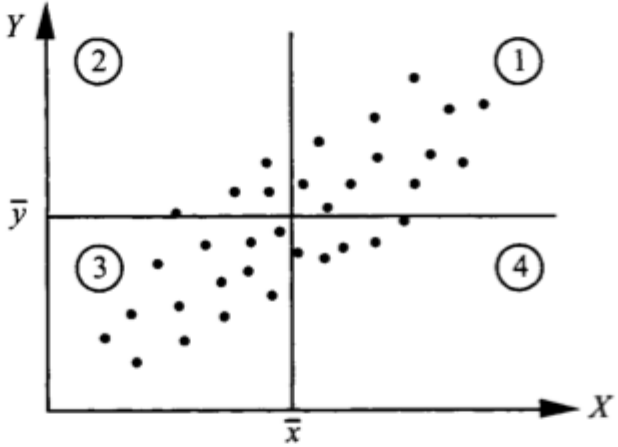
# Regressão Linear Simples

- Envolve o estudo do relacionamento entre uma **variável de resposta** Y e uma **variável preditora** $X_1$
- $\mathbf{X_1}$ passa a ser considerado apenas como **X**
- Também conhecida como *Straight Line Regression* (Regressão em Linha Reta) (GRAYBILL & IYER, 1994, p. 99)

# Covariância e Correlação

- São métricas que avaliam a **direção** e **força** de uma relação linear entre duas variáveis X e Y.

| Observation Number | Response $Y$ | Predictor $X$ |
|---|---|---|
| 1 | $y_1$ | $x_1$ |
| 2 | $y_2$ | $x_2$ |
| $\vdots$ | $\vdots$ | $\vdots$ |
| $n$ | $y_n$ | $x_n$ |




$y_i - \bar{y}$ Desvio entre cada observação y e a média da variável de resposta

$x_i - \bar{x}$ Desvio entre cada observação x e a média da variável preditora

# Covariância e Correlação

- Se a relação linear entre Y e X é positiva (quando X aumenta, Y também aumenta), temos mais pontos nos quadrantes 1 e 3;
- Quando é negativa (quando X aumenta, Y diminui), temos mais pontos nos quadrantes 2 e 4;
- Determina a direção da reta de regressão

# Covariância

- Para determinar a covariância entre Y e X, temos:

$$\text{Cov}(Y, X) = \frac{\sum_{i=1}^{n} (y_i - \bar{y})(x_i - \bar{x})}{n - 1}$$

- Se Cov(Y, X) > 0, então temos um relacionamento positivo

- Se Cov(Y, X) < 0, então temos um relacionamento negativo

# Correlação

- Para determinar a força da relação entre as variáveis Y e X, identificamos o **coeficiente de correlação** (valor entre -1 até 1):

$$\mathrm{Cor}(Y,X) = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{y_i - \bar{y}}{s_y}\right)\left(\frac{x_i - \bar{x}}{s_x}\right)$$

$S_y$ = Desvio padrão de y $\qquad$ $S_x$ = Desvio padrão de x

- Fórmula equivalente:

$$= \frac{\sum(y_i - \bar{y})(x_i - \bar{x})}{\sqrt{\sum(y_i - \bar{y})^2 \sum(x_i - \bar{x})^2}}$$

# Exemplo

- Considerando uma empresa que vende e conserta computadores. Qual a relação entre o tempo de uma chamada de serviço e o número de componentes a serem substituídos ou consertados?
  - tempo de uma chamada de serviço (variável de resposta)
  - número de componentes reparados (variável preditora)

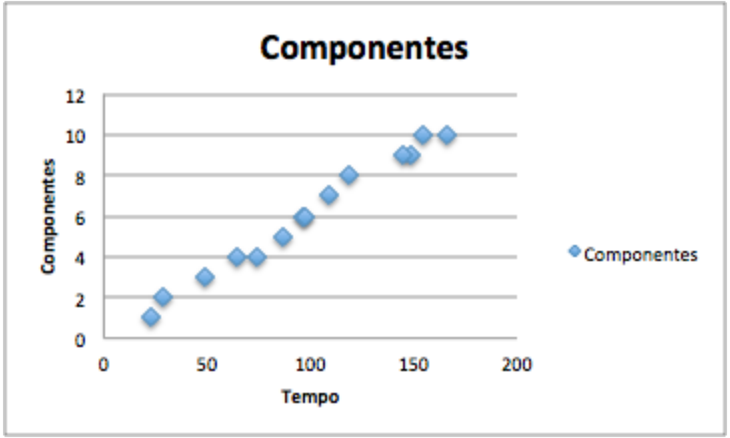
**Dados coletados:**

**Table 2.5** Length of Service Calls (in Minutes) and Number of Units Repaired

| Row | Minutes | Units | Row | Minutes | Units |
|---|---|---|---|---|---|
| 1 | 23 | 1 | 8 | 97 | 6 |
| 2 | 29 | 2 | 9 | 109 | 7 |
| 3 | 49 | 3 | 10 | 119 | 8 |
| 4 | 64 | 4 | 11 | 149 | 9 |
| 5 | 74 | 4 | 12 | 145 | 9 |
| 6 | 87 | 5 | 13 | 154 | 10 |
| 7 | 96 | 6 | 14 | 166 | 10 |

# Exemplo

## Gráfico de dispersão:




## Covariância:

$$\mathrm{Cov}(Y,X) = \frac{\sum_{i=1}^{n}(y_i-\bar{y})(x_i-\bar{x})}{n-1}$$

$$= \frac{1768}{13}$$

$$= 136$$

## Coeficiente de Correlação:

$$\mathrm{Cor}(Y,X) = \frac{\sum(y_i-\bar{y})(x_i-\bar{x})}{\sqrt{\sum(y_i-\bar{y})^2\sum(x_i-\bar{x})^2}}$$

$$= \frac{1768}{1779{,}2113}$$

$$= 0{,}993698746$$

# Modelo de Regressão Linear Simples

- O relacionamento entre a variável de resposta Y e a preditora X é estabelecido como um modelo linear:

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

- Onde:
    - $\square_0$ e $\square_1$ são constantes chamadas de coeficientes de regressão ou parâmetros
    - ε é uma margem de erro

# Modelo de Regressão Linear Simples

- □$_0$ -> intercepto
  - Valor previsto em Y quando X = 0.
- □$_1$ -> inclinação
  - Pode ser interpretada como a mudança ocorrida em Y para uma mudança de unidade em X.

- Cada observação pode ser escrita como:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad i = 1, 2, \ldots, n$$

# Modelo de Regressão Linear Simples

- Voltando ao exemplo:
  - Suponhamos que a empresa quer prever o número de engenheiros de serviço necessários ao longo dos próximos anos.
  - Um modo linear seria:

$$\textbf{Tempo}_i = \beta_0 + \beta_1 \textbf{ Componentes}_i + \varepsilon_i$$

# Modelo de Regressão Linear Simples

- A análise de regressão difere-se da análise de correlação em:
- Na correlação:
  - X e Y têm igual importância
  - O coeficiente de correlação é simétrico: Cor(Y,X) = Cor(X,Y)
- Na regressão:
  - A variável de resposta Y é de importância primordial
  - X tem habilidade de representar a variabilidade em Y

# Estimativa de Parâmetros

- Com base nos dados disponíveis podemos estimar os parâmetros $\square_0$ e $\square_1$
- Equivale encontrar a linha reta que dá o melhor ajuste dos pontos no gráfico de dispersão para variável de resposta vs. preditora.
- Para isso, utilizamos o método dos mínimos quadrados
  - Estabelece a linha que minimiza a soma dos quadrados das distâncias verticais, a partir de cada ponto até a linha.
- As distâncias verticais representam os erros na variável resposta.

# Estimativa de Parâmetros

- Os erros podem ser obtidos reescrevendo a fórmula:

$$\varepsilon_i = y_i - \beta_0 - \beta_1 x_i, \quad i = 1, 2, \ldots, n.$$

- A soma dos quadrados das distâncias podem ser obtidas por:

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

# Estimativa de Parâmetros

- Os valores ajustados de $\square_0$ e $\square_1$ que minimiza $S(\square_0,\square_1)$ é dado por:

$$\hat{\beta}_1 = \frac{\sum(y_i - \bar{y})(x_i - \bar{x})}{\sum(x_i - \bar{x})^2}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

- Estes valores determinam a **intercepção** e a **inclinação** da linha com a menor soma possível dos quadrados das distâncias verticais de cada ponto para a linha.
- São chamados de linha de regressão dos mínimos quadrados

# Estimativa de Parâmetros

- A linha de regressão dos mínimos quadrados é dada por:

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$$

  - Note que desprezamos o erro (ε) por já estarem considerados nos valores ajustados de ($\square_0$ e $\square_1$).

- Para encontrar os valores ajustados da variável de resposta, de cada observação, utilizamos:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i, \quad i = 1, 2, \ldots, n.$$

# Estimativa de Parâmetros

- Para calcular a distância vertical, temos:

$$e_i = y_i - \hat{y}_i, \quad i = 1, 2, \ldots, n$$

- Essa distância corresponde ao residual dos mínimos quadrados.

# Exemplo

- Utilizando os dados do exemplo anterior, temos:

$$\hat{\beta}_1 = \frac{\sum(y_i - \bar{y})(x_i - \bar{x})}{\sum(x_i - \bar{x})^2}$$

$$= \frac{1768}{114}$$

$$= 15{,}50877193$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

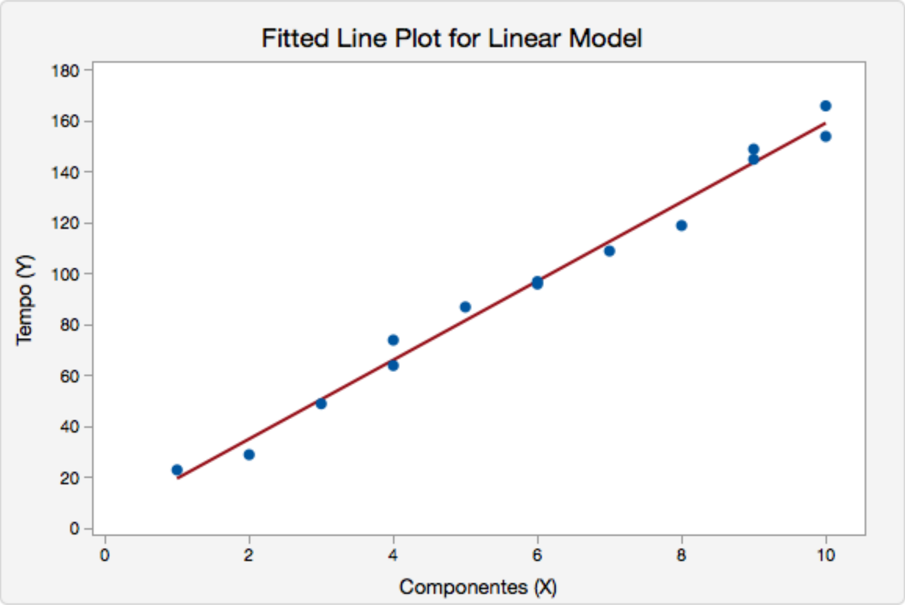
$$= 4{,}161654135$$

A equação da linha de regressão dos mínimos quadrados fica:

**Tempo = 4.162 + 15.509 * Componentes**

# Exemplo

- Usando o minitab para Análise de Regressão Simples



# Testes de Hipóteses

- É um método mais formal para medir a utilidade da variável X como uma preditora da variável Y, em comparação a análise do coeficiente de correlação e do gráfico de dispersão.
- Os testes de hipóteses são realizados para o parâmetro de regressão $\square_1$

# Testes de Hipóteses: Exemplo

- A empresa espera que o **aumento no tempo de serviço** para cada componente adicional a ser reparado seja de **12 minutos.**
- Teremos:
  - $\mathbf{H_0 : \square_1 = 12}$ minutos

    é chamada de **hipótese nula** e sempre abrange a igualdade.
  - $\mathbf{H_1 : \square_1 \neq 12}$ minutos

    é chamada de **hipótese alternativa** e é sempre o oposto da hipótese nula, que será apoiada caso a $\mathbf{H_0}$ seja rejeitada.

Esse é um tipo de hipótese **bilateral**.

# Testes de Hipóteses

- Em alguns casos podemos formular hipóteses **unilaterais:**
  - $\mathbf{H_0 : \square_1 = 12}$ minutos
  - $\mathbf{H_1 : \square_1 > 12}$ minutos

ou

  - $\mathbf{H_0 : \square_1 = 12}$ minutos
  - $\mathbf{H_1 : \square_1 < 12}$ minutos

# Testes de Hipóteses: Erros

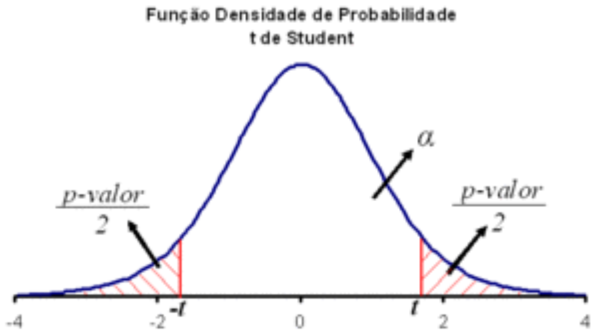
- **Erro do tipo I**:
  - Probabilidade de rejeitarmos a **hipótese nula** quando ela é efetivamente verdadeira (□ - nível de significância).
- **Erro do tipo II**:
  - Probabilidade de rejeitarmos a **hipótese alternativa** quando ela é efetivamente verdadeira.

# Testes de Hipóteses: Procedimentos

1. Avaliação do problema e escolha das hipóteses.
2. Decisão da estatística (estimador) p/ teste da hipótese nula (Propriedades: média, desvio padrão, distribuição estatística)
3. Estabelece um valor para Erro tipo I (□ - nível de significância)
4. Construção da região crítica (RC) que servirá de regra para rejeição ou não a hipótese nula.
5. Retira-se uma amostra da população e executa-se os cálculos para determinação do valor da estatística de teste (t).
6. Se $t \notin RC$, estabelecida □, não rejeite a hipótese nula. Se pertencer, rejeite a hipótese nula.

# Teste T de Student

- Um teste estatístico apropriado para testar a hipótese nula ($H_0$)
- O valor **t** é aplicado à função densidade de probabilidade da distribuição **t** para medir o tamanho da área abaixo dessa função para valores maiores ou iguais a **t**

# Teste T de Student: Fórmulas

$$\hat{\sigma}^2 = \frac{\sum e_i^2}{n-2} = \frac{\sum (y_i - \hat{y}_i)^2}{n-2} = \frac{\text{SSE}}{n-2}$$

$$\text{s.e.}(\hat{\beta}_0) = \hat{\sigma}\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\sum (x_i - \bar{x})^2}}$$

$$t_1 = \frac{\hat{\beta}_1 - \beta_1^0}{\text{s.e.}(\hat{\beta}_1)}$$

$$\text{s.e.}(\hat{\beta}_1) = \frac{\hat{\sigma}}{\sqrt{\sum (x_i - \bar{x})^2}}$$

- $\beta_1^0$ = Constante escolhida pelo avaliador. No exemplo: 12.
- s.e. = Erro padrão da estimativa
- $\sigma^2$ = Param. desconhecido estimado dos dados
- $\hat{\sigma}$ = Raiz quadrada de $\sigma^2$
- SSE = Soma dos quadrados residuais

# Teste T de Student

- Testando $H_0$: $\square_{\mathbf{1}}$ **= 12** contra $H_1$: $\square_{\mathbf{1}}$ **≠ 12**

# Tabela de Distribuição T de Student

- Com 12 graus de liberdade (n - 2)
- Nível de confiança de 95% bilateral
- Temos que:

$$t_1 = 6{,}948 \text{ é} > 2{,}179$$

**Resultado altamente significante levando a rejeição da hipótese nula**

| *Unilateral* | 75% | 80% | 85% | 90% | 95% | 97,50% | 99% | 99,50% | 99,75% | 99,90% | 99,95% |
|---|---|---|---|---|---|---|---|---|---|---|---|
| *Bilateral* | 50% | 60% | 70% | 80% | 90% | 95% | 98% | 99% | 99,50% | 99,80% | 99,90% |
| 1 | 1 | 1,376 | 1,963 | 3,078 | 6,314 | [illegible] | 31,82 | 63,66 | 127,3 | 318,3 | 636,6 |
| 2 | 0,816 | 1,061 | 1,386 | 1,886 | 2,92 | [illegible] | 6,965 | 9,925 | 14,09 | 22,33 | 31,6 |
| 3 | 0,765 | 0,978 | 1,25 | 1,638 | 2,353 | [illegible] | 4,541 | 5,841 | 7,453 | 10,21 | 12,92 |
| 4 | 0,741 | 0,941 | 1,19 | 1,533 | 2,132 | [illegible] | 3,747 | 4,604 | 5,598 | 7,173 | 8,61 |
| 5 | 0,727 | 0,92 | 1,156 | 1,476 | 2,015 | [illegible] | 3,365 | 4,032 | 4,773 | 5,893 | 6,869 |
| 6 | 0,718 | 0,906 | 1,134 | 1,44 | 1,943 | [illegible] | 3,143 | 3,707 | 4,317 | 5,208 | 5,959 |
| 7 | 0,711 | 0,896 | 1,119 | 1,415 | 1,895 | [illegible] | 2,998 | 3,499 | 4,029 | 4,785 | 5,408 |
| 8 | 0,706 | 0,889 | 1,108 | 1,397 | 1,86 | [illegible] | 2,896 | 3,355 | 3,833 | 4,501 | 5,041 |
| 9 | 0,703 | 0,883 | 1,1 | 1,383 | 1,833 | [illegible] | 2,821 | 3,25 | 3,69 | 4,297 | 4,781 |
| 10 | 0,7 | 0,879 | 1,093 | 1,372 | 1,812 | [illegible] | 2,764 | 3,169 | 3,581 | 4,144 | 4,587 |
| 11 | 0,697 | 0,876 | 1,088 | 1,363 | 1,796 | [illegible] | 2,718 | 3,106 | 3,497 | 4,025 | 4,437 |
| 12 | [illegible] | [illegible] | [illegible] | [illegible] | [illegible] | **2,179** | 2,681 | 3,055 | 3,428 | 3,93 | 4,318 |
| 13 | 0,694 | 0,87 | 1,079 | 1,35 | 1,771 | 2,16 | 2,65 | 3,012 | 3,372 | 3,852 | 4,221 |
| 14 | 0,692 | 0,868 | 1,076 | 1,345 | 1,761 | 2,145 | 2,624 | 2,977 | 3,326 | 3,787 | 4,14 |
| 15 | 0,691 | 0,866 | 1,074 | 1,341 | 1,753 | 2,131 | 2,602 | 2,947 | 3,286 | 3,733 | 4,073 |
| 16 | 0,69 | 0,865 | 1,071 | 1,337 | 1,746 | 2,12 | 2,583 | 2,921 | 3,252 | 3,686 | 4,015 |
| 17 | 0,689 | 0,863 | 1,069 | 1,333 | 1,74 | 2,11 | 2,567 | 2,898 | 3,222 | 3,646 | 3,965 |
| 18 | 0,688 | 0,862 | 1,067 | 1,33 | 1,734 | 2,101 | 2,552 | 2,878 | 3,197 | 3,61 | 3,922 |
| 19 | 0,688 | 0,861 | 1,066 | 1,328 | 1,729 | 2,093 | 2,539 | 2,861 | 3,174 | 3,579 | 3,883 |
| 20 | 0,687 | 0,86 | 1,064 | 1,325 | 1,725 | 2,086 | 2,528 | 2,845 | 3,153 | 3,552 | 3,85 |
| 21 | 0,686 | 0,859 | 1,063 | 1,323 | 1,721 | 2,08 | 2,518 | 2,831 | 3,135 | 3,527 | 3,819 |
| 22 | 0,686 | 0,858 | 1,061 | 1,321 | 1,717 | 2,074 | 2,508 | 2,819 | 3,119 | 3,505 | 3,792 |
| 23 | 0,685 | 0,858 | 1,06 | 1,319 | 1,714 | 2,069 | 2,5 | 2,807 | 3,104 | 3,485 | 3,767 |
| 24 | 0,685 | 0,857 | 1,059 | 1,318 | 1,711 | 2,064 | 2,492 | 2,797 | 3,091 | 3,467 | 3,745 |
| 25 | 0,684 | 0,856 | 1,058 | 1,316 | 1,708 | 2,06 | 2,485 | 2,787 | 3,078 | 3,45 | 3,725 |
| 26 | 0,684 | 0,856 | 1,058 | 1,315 | 1,706 | 2,056 | 2,479 | 2,779 | 3,067 | 3,435 | 3,707 |
| 27 | 0,684 | 0,855 | 1,057 | 1,314 | 1,703 | 2,052 | 2,473 | 2,771 | 3,057 | 3,421 | 3,69 |

# Regressão Linear Múltipla

- Representa uma generalização da regressão linear simples
- Considera um **conjunto *p* de variáveis preditoras**, expandindo o potencial analítico da regressão para **múltiplos fatores** de influência
- Para cada valor das variáveis preditoras ($X_p$), o modelo expresso através de uma equação linear fornece uma aproximação aceitável da verdadeira relação entre o valor de Y e os valores de X

# Regressão Linear Múltipla

- Desta forma, o relacionamento entre as variáveis passa a ser definido perante o seguinte modelo:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \varepsilon$$

  - $\beta_0, \beta_1, \beta_2, \ldots, \beta_p$ são os coeficientes de regressão
  - $\varepsilon$ representa o fator de erro (perturbação aleatória)
  - $X_1, X_2, \ldots, X_p$ são as variáveis preditoras

# Regressão Linear Múltipla

- Em uma representação tabular para o modelo expresso na equação anterior...

| Observation Number | Response $Y$ | Predictors $X_1$ | $X_2$ | $\ldots$ | $X_p$ |
|---|---|---|---|---|---|
| 1 | $y_1$ | $x_{11}$ | $x_{12}$ | $\ldots$ | $x_{1p}$ |
| 2 | $y_2$ | $x_{21}$ | $x_{22}$ | $\ldots$ | $x_{2p}$ |
| 3 | $y_3$ | $x_{31}$ | $x_{32}$ | $\ldots$ | $x_{3p}$ |
| $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ |
| $n$ | $y_n$ | $x_{n1}$ | $x_{n2}$ | $\ldots$ | $x_{np}$ |

# Estimativa de parâmetros

- Para efetuar a descoberta do valor para os parâmetros (coeficientes de regressão), é necessário **aplicar o método dos quadrados mínimos** (assim como na regressão linear simples)

$$\text{SSE} = \sum_{i=1}^{n} \varepsilon_i^2$$

- Neste caso, considerando a característica de múltiplas variáveis preditoras, temos:

$$\text{SSE} = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_{i1} - \ldots - \beta_p x_{ip})^2$$

# Interpretando coeficientes de regressão

- Utilizando soma dos quadrados mínimos (assim como na regressão linear simples, comparando cada preditora à variável de resposta)
- Resolvendo através de um sistema de equações e substituindo valores dos coeficientes
- Utilizando estratégia de equações normais (matrizes)

# Exemplo

Uma pesquisa foi aplicada considerando a satisfação dos funcionários em relação aos seus supervisores em uma determinada empresa. Com base em um estudo exploratório, seis itens em um questionário foram escolhidos como possíveis variáveis preditoras:

- Recebe as queixas de insatisfação que o funcionário faz
- Não permite privilégios especiais
- Dá oportunidade ao aprendizado de coisas novas
- Cresce com base no desempenho
- É bastante crítico quanto a níveis de desempenho baixos

# Exemplo

- Observações
  - As questões 1, 2, e 5 estão relacionadas ao relacionamento interpessoal entre o funcionário e o supervisor
  - As questões 3 e 4 são referentes à realização das atividades do trabalho como um todo
  - A questão 6 representa uma percepção do funcionário sobre o progresso do supervisor na empresa
  - As respostas para as questões variam em uma escala de 1 a 5, representando respectivamente de muito satisfeito para muito insatisfeito

# Exemplo

- Definição de variáveis na pesquisa de desempenho do supervisor

| | |
|---|---|
| $X_1$ | Recebe as queixas de insatisfação que o funcionário faz |
| $X_2$ | Não permite privilégios especiais |
| $X_3$ | Dá oportunidade ao aprendizado de coisas novas |
| $X_4$ | Cresce com base no desempenho |
| $X_5$ | É bastante crítico quanto a níveis de desempenho baixos |
| $X_6$ | Frequência com que avança para funções melhores |
| $Y$ | Taxa geral de trabalho sendo feito (satisfatoriamente) pelo supervisor |

# Exemplo

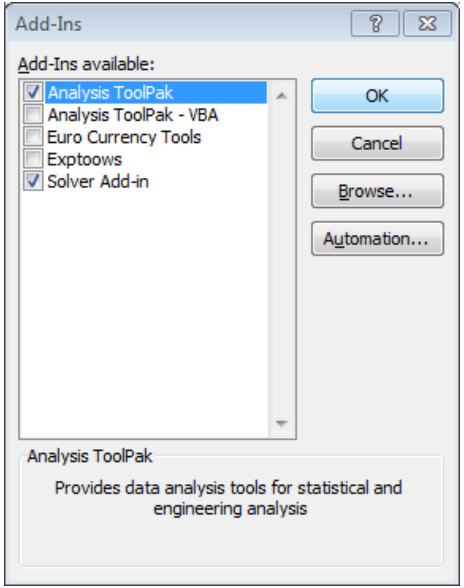
| Row | $Y$ | $X_1$ | $X_2$ | $X_3$ | $X_4$ | $X_5$ | $X_6$ |
|---|---|---|---|---|---|---|---|
| 1 | 43 | 51 | 30 | 39 | 61 | 92 | 45 |
| 2 | 63 | 64 | 51 | 54 | 63 | 73 | 47 |
| 3 | 71 | 70 | 68 | 69 | 76 | 86 | 48 |
| 4 | 61 | 63 | 45 | 47 | 54 | 84 | 35 |
| 5 | 81 | 78 | 56 | 66 | 71 | 83 | 47 |
| 6 | 43 | 55 | 49 | 44 | 54 | 49 | 34 |
| 7 | 58 | 67 | 42 | 56 | 66 | 68 | 35 |
| 8 | 71 | 75 | 50 | 55 | 70 | 66 | 41 |
| 9 | 72 | 82 | 72 | 67 | 71 | 83 | 31 |
| 10 | 67 | 61 | 45 | 47 | 62 | 80 | 41 |

. . .

http://www.ilr.cornell.edu/~hadi/RABE4/Data4/P056.txt

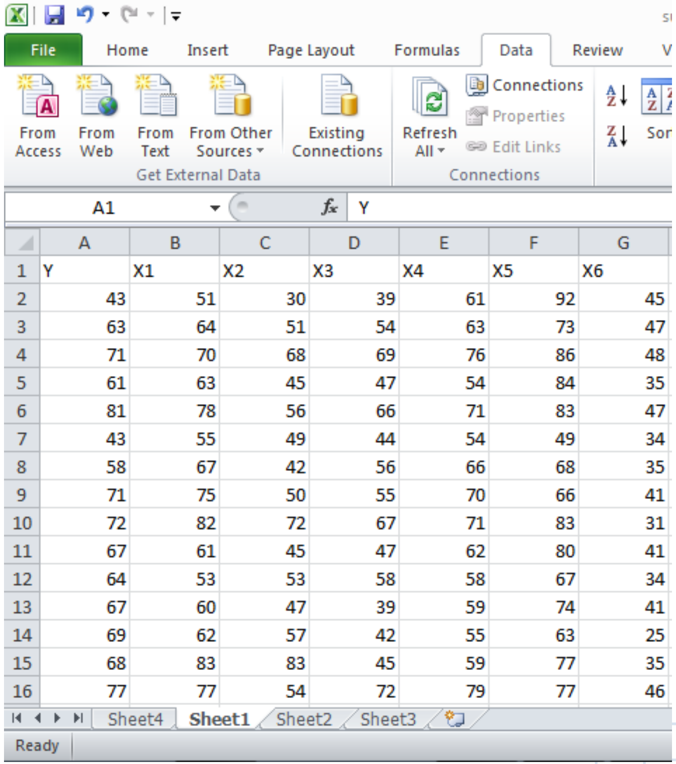
# Exemplo

- Utilizando o Excel
  - Certificar que o Add-in de Análise de Dados está ativo



# Exemplo

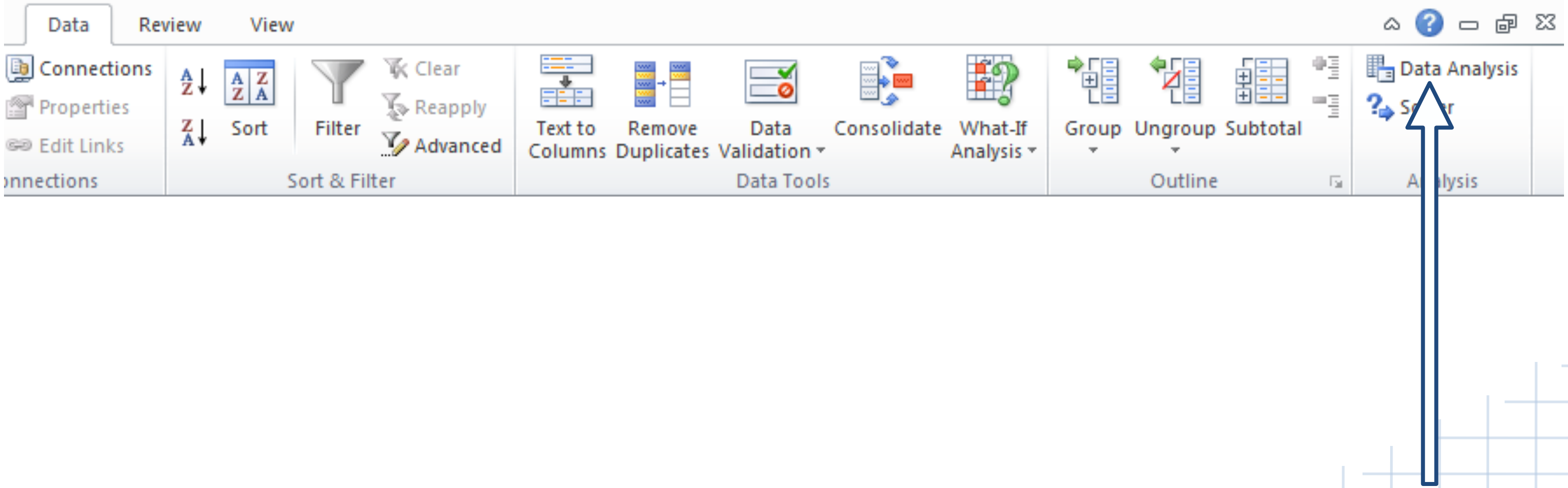
- Utilizando o Excel
  - Importar dados para planilha




| | A | B | C | D | E | F | G |
|---|---|---|---|---|---|---|---|
| 1 | Y | X1 | X2 | X3 | X4 | X5 | X6 |
| 2 | 43 | 51 | 30 | 39 | 61 | 92 | 45 |
| 3 | 63 | 64 | 51 | 54 | 63 | 73 | 47 |
| 4 | 71 | 70 | 68 | 69 | 76 | 86 | 48 |
| 5 | 61 | 63 | 45 | 47 | 54 | 84 | 35 |
| 6 | 81 | 78 | 56 | 66 | 71 | 83 | 47 |
| 7 | 43 | 55 | 49 | 44 | 54 | 49 | 34 |
| 8 | 58 | 67 | 42 | 56 | 66 | 68 | 35 |
| 9 | 71 | 75 | 50 | 55 | 70 | 66 | 41 |
| 10 | 72 | 82 | 72 | 67 | 71 | 83 | 31 |
| 11 | 67 | 61 | 45 | 47 | 62 | 80 | 41 |
| 12 | 64 | 53 | 53 | 58 | 58 | 67 | 34 |
| 13 | 67 | 60 | 47 | 39 | 59 | 74 | 41 |
| 14 | 69 | 62 | 57 | 42 | 55 | 63 | 25 |
| 15 | 68 | 83 | 83 | 45 | 59 | 77 | 35 |
| 16 | 77 | 77 | 54 | 72 | 79 | 77 | 46 |

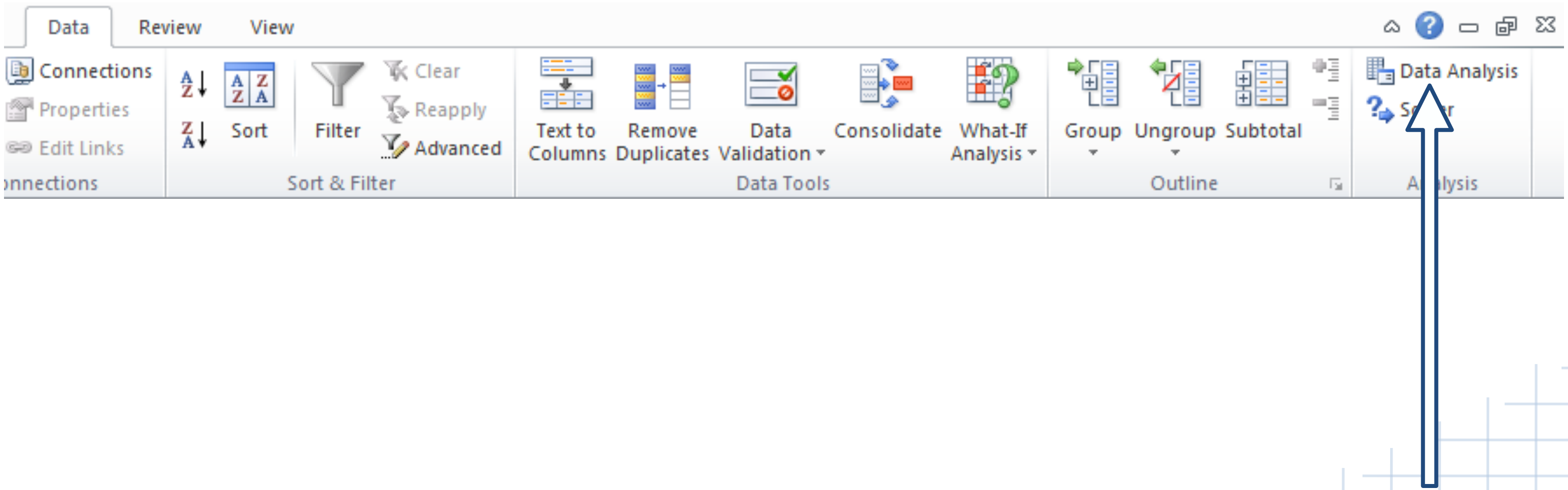
# Exemplo

- Utilizando o Excel
  - Na guia “Dados” da faixa de opções, selecionar “Análise de Dados”



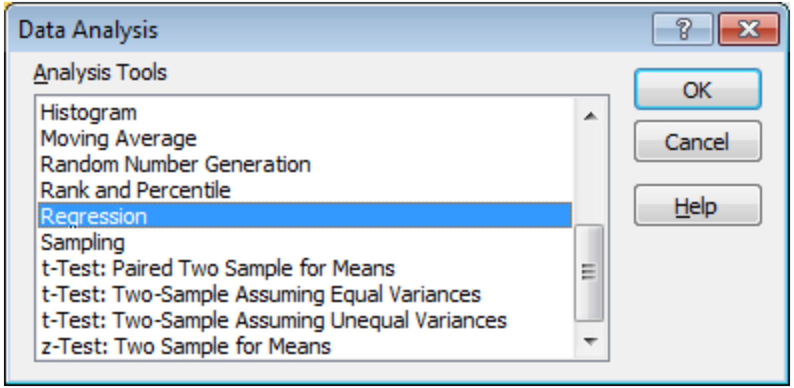
# Exemplo

- Utilizando o Excel
  - Na guia “Dados” da faixa de opções, selecionar “Análise de Dados”



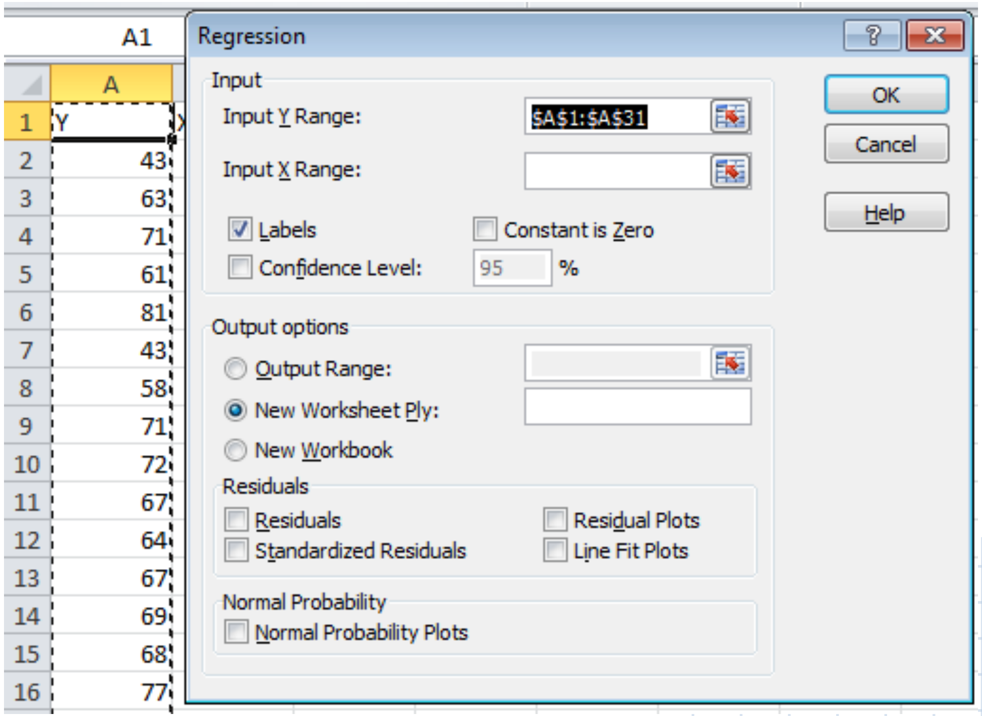
# Exemplo

- Utilizando o Excel
  - Na janela do add-in de análise de dados, selecionar a opção “Regressão”



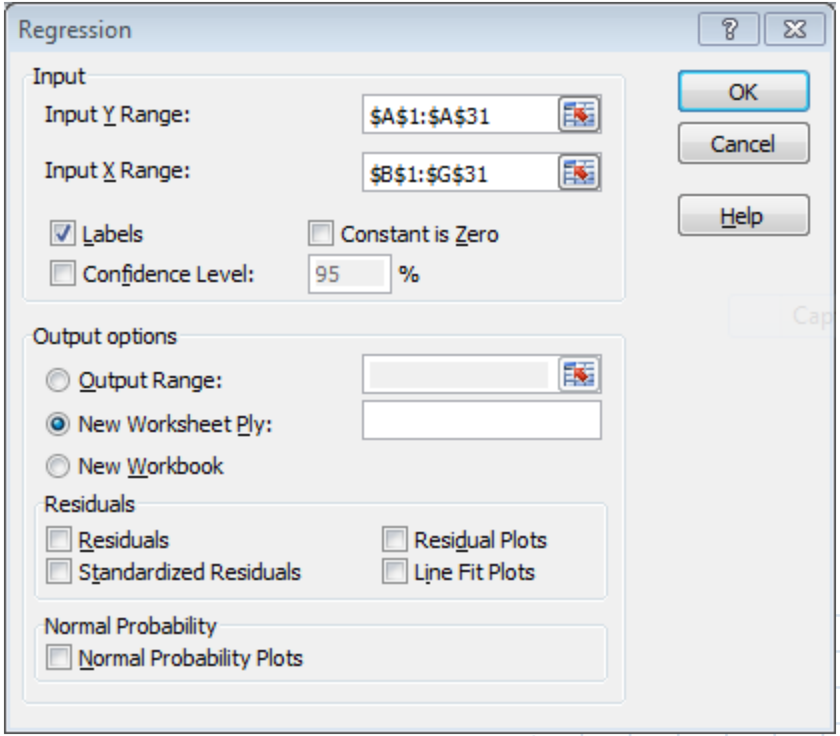
# Exemplo

- Utilizando o Excel
    - Na janela de configuração para a ferramenta de Regressão, selecionar os dados para a coluna que representa a variável de resposta (Y)



# Exemplo

- Utilizando o Excel
  - Fazer o mesmo para as variáveis preditoras ($X_p$)



# Exemplo

SUMMARY OUTPUT

| *Regression Statistics* | |
|---|---|
| Multiple R | 0,855921721 |
| R Square | 0,732601993 |
| Adjusted R Square | 0,662845991 |
| Standard Error | 7,067993765 |
| Observations | 30 |

ANOVA

| | *df* | *SS* | *MS* | *F* | *Significance F* |
|---|---|---|---|---|---|
| Regression | 6 | 3147,966342 | 524,661057 | 10,50235065 | 0,000012404 |
| Residual | 23 | 1149,000325 | 49,95653586 | | |
| Total | 29 | 4296,966667 | | | |

| | *Coefficients* | *Standard Error* | *t Stat* | *P-value* | *Lower 95%* | *Upper 95%* |
|---|---|---|---|---|---|---|
| Intercept | 10,78707639 | 11,58925724 | 0,930782375 | 0,361633721 | -13,18712881 | 34,76128158 |
| X1 | 0,613187608 | 0,160983115 | 3,809018158 | 0,000902868 | 0,280168662 | 0,946206554 |
| X2 | -0,073050143 | 0,13572469 | -0,53822295 | 0,595593921 | -0,353818056 | 0,20771777 |
| X3 | 0,320332116 | 0,168520319 | 1,900851595 | 0,069925346 | -0,028278723 | 0,668942956 |
| X4 | 0,081732134 | 0,221477677 | 0,369031022 | 0,715480088 | -0,376429349 | 0,539893616 |
| X5 | 0,038381447 | 0,146995442 | 0,261106377 | 0,796334264 | -0,265701793 | 0,342464688 |
| X6 | -0,217056682 | 0,178209471 | -1,217986229 | 0,235577049 | -0,58571106 | 0,151597697 |