Regressão Linear Simples

Frases

"Por serem mais precisos que as palavras, os números são particularmente adequados para transmitir conclusões científicas"

Pagano e Gauvre, 2004

Roteiro

1. Modelagem de Relação
2. Modelo Linear
3. Estimação dos Parâmetros
4. Inferência sobre os Parâmetros
5. Avaliação do Modelo
6. Aplicação
7. Referências

Modelagem de Relação

Regressão e Correlação

- Regressão:

  Usa variável(eis) explicativa(s) para explicar ou predizer comportamento de variável resposta (quando houver sentido).

- Correlação:

  Trata simetricamente duas variáveis

Regressão

- Variável resposta (Y):

  Variável resposta cujo comportamento se quer explicar

- Variável(eis) explicativa(s) (X_i):

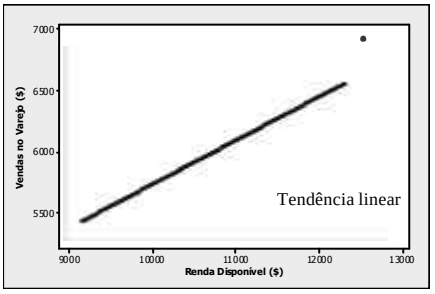
  São de interesse caso ajudem a entender, explicar ou predizer o comportamento de Y.

- O enfoque da regressão é natural quando Y é aleatória e X_i é controlada ou não-aleatória.

x	*Y*
• Variável explicativa • Variável independente • Regressor • Preditor • Variável exógena • Variável de controle ou estímulos	• Variável explicada • Variável dependente • Regredido • Predito • Variável endógena • Variável resposta

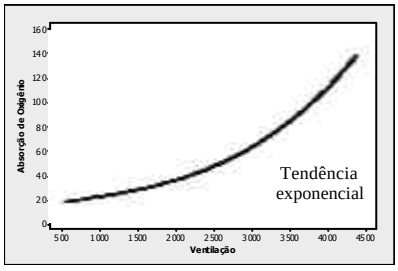
Exemplo 1 – Vendas/Renda

- Variável resposta: Vendas no varejo ($)
- Variável explicativa: Renda disponível ($)



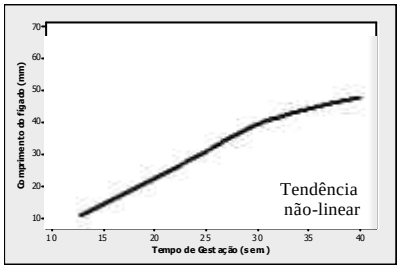
Exemplo 2 – Absorção de Oxigênio

- Variável resposta: Absorção de Oxigênio
- Variável explicativa: Ventilação



Exemplo 3 – Comprimentos de Fígados

- Variável resposta: Comprimento do fígado (mm)
- Variável explicativa: Tempo de gestação (sem.)



Modelo de Regressão

- Relação de regressão:

  Tendência + dispersão residual

- Tendência:
  - √ Suavização dos dados
  - √ Explica a maior parte das diferenças de Y
- Valores atípicos:

  Observações muito diferente do restante dos dados

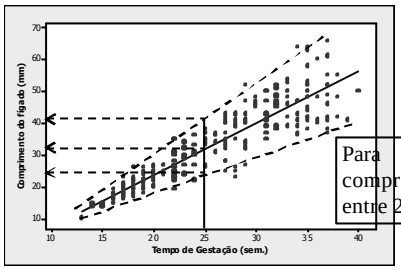
Relações Fortes e Fracas

- Relação Forte:

  A dispersão é pequena em relação à amplitude dos valores da curva de tendência

- Em dados observacionais, relações fortes não são necessariamente causais

Intervalo de Predição

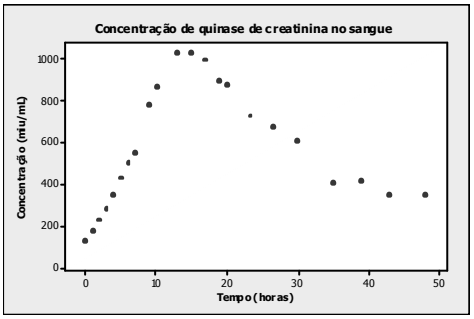


Para 25 semanas: comprimento fígado entre 23 e 41 mm

Grande dispersão em torno da tendência ⇨ Intervalos de predição amplos

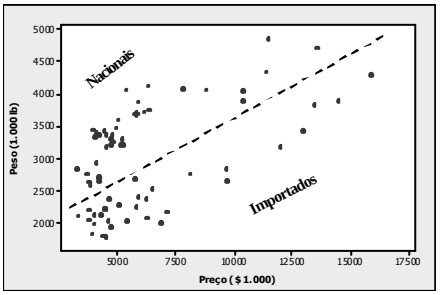
Relação fraca **Predição imprecisa**

Outros Padrões (1)



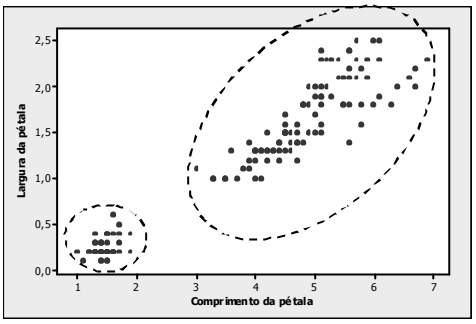
Os padrões da tendência variam

Outros Padrões (2)



Importante descobrir o que define os grupos

Outros Padrões (3)



Variedades diferentes de Flores

Resumo de Tendência – Abordagens

- Ajuste de funções matemáticas:

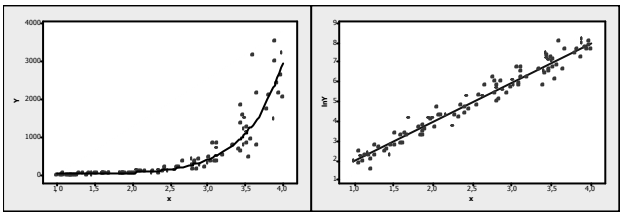
$$Y = f(X)$$

- Técnicas de suavização:
  'Lowess', núcleo-estimador, *'spline'*

Ajuste de Funções

- Tendência linear: $Y = b_0 + b_1 X$
  - √ Para cada mudança de uma unidade em X, Y muda uma quantidade fixa.
- Tendência quadrática: $Y = b_0 + b_1 X + b_2 X^2$
  - √ Tendência levemente curva

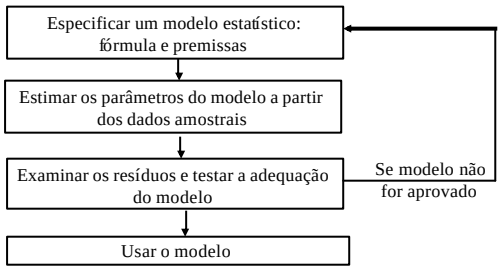
- Tendência exponencial: $Y = b_0 e^{b_1 X}$



√ Cada mudança de uma unidade em *X*, *Y* muda uma % fixa

√ Se a tendência é exponencial, o gráfico de *log(Y)* vs *X* têm tendência linear

Passos na Construção de um Modelo Estatístico



Modelo Linear

Tipos

- Simples:
  - √ Uma variável independente (explicativa)
- Múltipla:
  - √ Duas ou mais variáveis independentes

Objetivos

- Encontrar equação matemática que permita:
  - √ Descrever e compreender a relação entre 2 ou mais variáveis aleatórias
  - √ Projetar ou estimar uma nova observação
- Ajustar uma reta a partir dos dados amostrais

Utilidades

- Busca de relações de Causa e Efeito;
- Predição de valores;
- Estabelecer explicação sobre população a partir de uma amostra

Regressão Linear Simples

- Busca-se a equação de uma reta que permita:
  - √ Descrever e compreender a relação entre duas variáveis
  - √ Projetar e estimar uma das variáveis em função da outra.

Regressão Linear Simples (2)

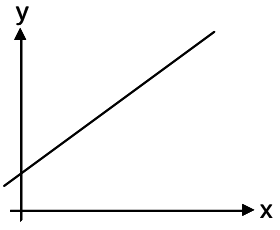
- A partir de valores observados de X e Y, modelar a tendência através de uma equação do tipo:

$$Y_i = b_0 + b_1 X_i$$

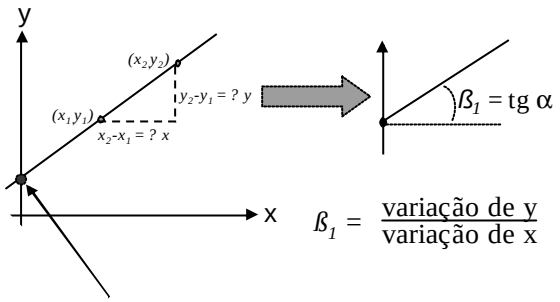
Função Linear

$$Y_i = b_0 + b_1 X_i$$

- f(x) se modifica a uma taxa constante em relação à sua variável independente
- $ß_0$ e $ß_1$ são constantes
- $ß_0$: intercepto-y
- $ß_1$: coeficiente angular



Intercepto e Coeficiente Angular



$ß_1 = \frac{\text{variação de y}}{\text{variação de x}}$

$ß_0$: intersecção da reta com o eixo y

Interpretação dos Parâmetros

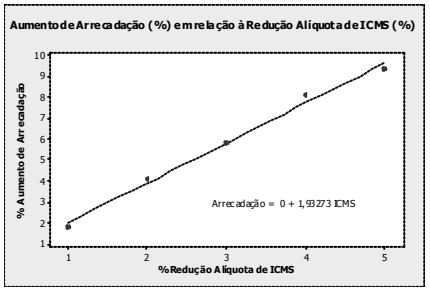
- $ß_1$: declividade da reta

  define o aumento ou diminuição da variável *Y* por unidade de variação de *X*

- $ß_0$ = intercepto em y

  define o valor médio de *Y* sem a interferência de *X* (com *X=0*).

Exemplo



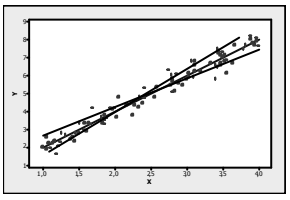
$ß_1$: a cada redução de 1% na alíquota há *1,9 %* de aumento de arrecadação

$ß_0$*=0:* Se não há redução na alíquota não há de aumento de arrecadação

A Reta de Regressão

- Para um mesmo valor X_i podem existir um ou mais valores de Y_i amostrados
- Para esse mesmo valor X_i haverá um valor ajustado $\hat{Y}_i$
- Para cada valor X_i existirá um dado desvio d_i dos valores de $\hat{Y}_i$
- Há observações que não são pontos da reta.

Ajuste da Reta



- Qual a reta que se ajusta melhor aos dados? ou seja quais os valores de $ß_0$ e $ß_1$?
- Escolher $ß_0$ e $ß_0$ de maneira a tornar mínima a distância entre a reta e os pontos

Método dos Mínimos Quadrados

- Critério:

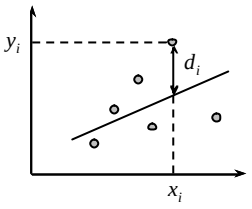
  Valores dos parâmetros que minimizam a soma dos quadrados dos desvios

$$\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$$

Método dos Mínimos Quadrados (2)

- Minimização em relação a $ß_0$ e $ß_1$:

$$S = \sum d_i^2 = \sum \{Y_i - (b_0 + b_1 x_i)\}^2$$



$$\frac{\partial S}{\partial b_0} = 0$$

$$\frac{\partial S}{\partial b_1} = 0$$

Método dos Mínimos Quadrados (3)

- Resultados das derivadas parciais:

$$\hat{b}_1 = \frac{n.\sum(x_i y_i) - (\sum x_i).(\sum y_i)}{n.\sum x_i^2 - (\sum x_i)^2} \qquad \hat{b}_1 = \frac{S_{xy}}{S_{xx}}$$

$$\hat{b}_0 = \frac{\sum y_i - b\sum x_i}{n} \qquad \hat{b}_0 = \overline{Y} - \hat{b}_1 \overline{X}$$

- Calculando por medidas estatísticas :

$$\hat{b}_1 = \frac{s_{XY}}{s_X^2} = r_{XY}\frac{s_Y}{s_X}$$

Estimativa de Mínimos Quadros

- Dadas observações (X_1, Y_1), ... , (X_n, Y_n) os coeficientes da reta que melhor se ajusta aos dados são:

$$b_0 = \hat{b}_0 \quad \text{e} \quad b_1 = \hat{b}_1$$

que são chamados estimativas de mínimos quadrados do intercepto e da declividade

- A reta de mínimos quadrados é dada por:

$$\hat{Y}_i = \hat{b}_0 + \hat{b}_1 X_i$$

Exemplo – Lei de Consumo Keynesiano

- Na média, há disposição do indivíduo aumentar seu consumo quando sua renda aumenta, mas não tanto quanto o aumento em sua renda.
- A propensão marginal para consumir é maior que *0* mas menor que *1*

Consumo Familiar ($)	Renda Familiar ($)			
Y	X	X^2	X.Y	Y^2
70	80	6.400	5.600	4.900
65	100	10.000	6.500	4.225
90	120	14.400	10.800	8.100
95	140	19.600	13.300	9.025
110	160	25.600	17.600	12.100
115	180	32.400	20.700	13.225
120	200	40.000	24.000	14.400
140	220	48.400	30.800	19.600
155	240	57.600	37.200	24.025
150	260	67.600	39.000	22.500
1.110	1.700	322.000	205.500	132.100

$\bar{y} = 111 \quad \bar{x} = 170$

$S_{xx} = 322.000 - 10(170)^2 = 33.000$

$S_{xy} = 205.500 - 10(170)(111) = 16.800$

$\hat{b}_1 = \frac{16.800}{33.000} = 0{,}5091$

$\hat{b}_0 = 111 - (0{,}5091)(170) = 24{,}45$

$\hat{Y} = 24{,}45 + 0{,}5091X$

Exemplo – HP

- Entrar cada par de dados:
  - √ Y_i ***Enter*** X_i $\boldsymbol{S}_+$
- Para determinar o intercepto-y:
  - √ Digite *0* ***g*** $\boxed{\hat{y}, r}$ $\hat{b}_0 = 24{,}45$
- Para determinar a inclinação:
  - √ Digite *1* ***g*** $\boxed{\hat{y}, r}$ *24,9636*
  - √ Subtraia o valor anterior do último valor obtido $\hat{b}_1 = 24{,}9636 - 24{,}4545 = 0{,}5091$

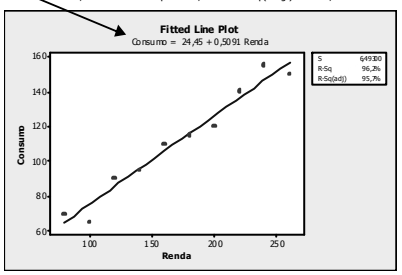
Exemplo –Minitab

Stat > Regression > Fitted Line Plot

Exemplo – Saída Minitab

Regression Analysis: Consumo versus Renda

```
The regression equation is
Consumo = 24,45 + 0,5091 Renda

S = 6,49300   R-Sq = 96,2%   R-Sq(adj) = 95,7%
```



Interpretação

$$\hat{Y} = 24{,}45 + 0{,}5091X$$

- Inclinação:
  Propensão marginal ao consumo
  Quando a renda aumenta $1, o aumento estimado no consumo médio é $ 0,51
  As estimativas são válidas dentro da classe amostrada (renda semanal entre $80 e3 $260)
- Intercepto-*y*
  A reta indica $24,45 como nível de consumo quando a renda é zero
  Esta interpretação não é válida já que não há pontos amostrais próximos à renda zero

Resíduo de uma Observação

- Diferença entre o valor observado e o valor estimado (ponto na reta)

$$\hat{e}_i = y_i - \hat{y}_i$$

Em que: $\hat{y}_i = \hat{b}_0 + \hat{b}_1 x_i$

Propriedades dos *EMQO*

- A soma dos resíduos é zero;
- A soma dos valores observados da resposta é igual à soma dos valores ajustados

$$\sum y_i = \sum \hat{y}_i$$

- A reta de regressão passa pelo ponto
- A soma dos resíduos ponderados pelos níveis da variável independente é zero

$$\sum x_i \hat{e}_i = 0$$

Propriedades dos *EMQO* (2)

- A soma dos resíduos ponderados pelos valores ajustados é zero;

$$\sum \hat{e}_i \hat{y}_i = 0$$

- Os resíduos não tem correlação com os valores ajustados

$$corr(\hat{e}_i, \hat{y}_i) = 0$$

- Os resíduos não tem correlação com a variável independente

$$corr(\hat{e}_i, x_i) = 0$$

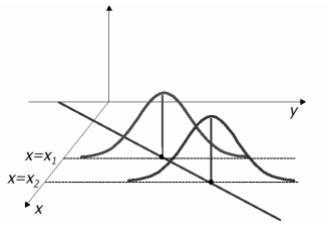
Estimação dos Parâmetros

Elaboração do Modelo

- Quer se encontrar um modelo para relações aproximadamente lineares, com dois aspectos:
  - √ tendência central linear
  - √ flutuações em torno desta tendência

Hipóteses sobre as Flutuações

- As observações de Y_i em $X = x_i$ são aleatórias, com alguma distribuição com:
  - √ média μ_Y pertencente a uma reta (explica padrão linear)
  - √ desvio padrão s (explica a dispersão)
- Suponha que esta distribuição seja normal:



- Quando $X = x_i$, $Y \sim N(\mu_Y, s^2)$, com $\mu_Y = ß_0 + ß_1 x_i$
- Apresentado de outra maneira:
  Quando $X = x_i$, $Y = ß_0 + ß_1 x_i + e$, $e \sim N(0, s^2)$

Modelo Clássico – Hipóteses Subjacentes

- Modelo de regressão linear nos parâmetros;
- Os valores de X são fixados em amostragem repetida (X não-estocástico);
- Perturbação e_i com média zero;
- Variância constante de e_i (homocedasticidade)
- Perturbações não correlacionadas:

$$cov(e_i , e_j)=0$$

Modelo Clássico – Hipóteses Subjacentes

- Covariância zero entre e_i e X_i;
- Número de observações maior que o número de parâmetros a serem estimados;
- Variabilidade nos valores de X $(Var(X)>0)$;
- Não há viés de especificação
  (o modelo está corretamente especificado)

Modelo Clássico – Hipóteses Subjacentes

- Não existe multicolinearidade perfeita:

  Não há relações lineares perfeitas entre as variáveis explicativas

Reta de Regressão – Estimação

Sob estas hipóteses:

- e são os melhores estimadores lineares não viciados de $ß_0$ e $ß_1$.

ou seja

é uma estimativa da relação populacional

$$Y_i = ß_0 + ß_1 x_i + e$$

onde e representa a dispersão na população.

Propriedades Importantes de um Bom Estimador

- Consistência:
  - √ Estimativa se aproxima do verdadeiro valor do parâmetro à medida que o tamanho da amostra aumenta

$$\hat{q} \to q$$

Propriedades Importantes de um Bom Estimador

- Exatidão:
  √ Relacionada com o vício do estimador

$$Bias(\hat{q}) = \hat{q} - q$$

- Precisão:
  √ Relacionada com a variabilidade do estimador

$$Var(\hat{q})$$

Quanto menor a variabilidade, mais preciso é o estimador

Estimadores de Mínimos Quadrados

- São lineares
- São não-viciados
- São estimadores eficientes (mínima variância)
- Têm variância mínima na classe dos estimadores lineares não-viciados
- Propriedades de amostras finitas (independem do tamanho da amostra)

Estimador de β_1

- É estimador não viciado:

$$E(\hat{b}_1) = b_1$$

- Variância:

$$Var(\hat{b}_1) = \frac{s^2}{S_{xx}}$$

Quanto maior a variabilidade de X, maior sua precisão

Estimador de $ß_0$

- É estimador não viciado:

$$E\left(\hat{b}_0\right) = b_0$$

- Variância:

$$Var\left(\hat{b}_0\right) = s^2\left[\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right]$$

Quanto mais afastado for o centróide, menor sua precisão

Covariância entre os Estimadores

- É dada por

$$Cov\left(\hat{b}_0, \hat{b}_1\right) = -\frac{\bar{x}s^2}{S_{xx}}$$

Se $ß_1$ for superestimado, o intercepto será subestimado (ou vice-versa)

Inferência sobre s^2

- Em geral, s^2 é desconhecido.
- A estimativa de mínimos quadrados ordinários é dada por:

$$\hat{s}^2 = \frac{1}{n-2}\sum_i \hat{e}_i^2 = \frac{SQ\,\mathrm{Res}}{gl}$$

√ *SQRes*: Soma dos Quadrados dos Resíduos

√ gl: graus de liberdade

Quantidade de observações – número parâmetros

Exemplo – Lei de Consumo

- Soma dos Quadrados dos Resíduos: *337,27*
- Estimativa de s^2:

$$\hat{s}^2 = \frac{337,27}{8} = 42,159$$

- Estimativa da Variância dos Estimadores:

$$Var(\hat{b}_0) = 42,159\left[\frac{1}{10} + \frac{(170)^2}{33.000}\right] = 41,137$$

$$Var(\hat{b}_1) = \frac{42,159}{33.000} = 0,001278$$

Erro Padrão

Definido como a raiz quadrada da variância do estimador

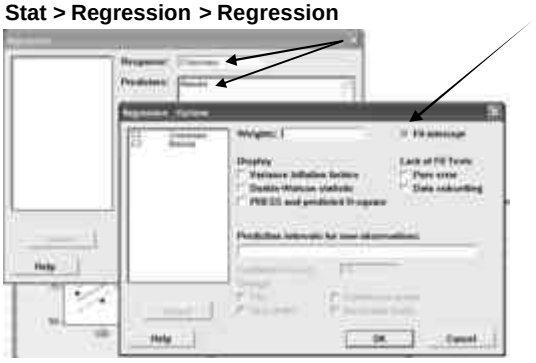
$$ep(\hat{b}_0) = \sqrt{41,137} = 6,414$$

$$ep(\hat{b}_1) = \sqrt{0,001278} = 0,0357$$

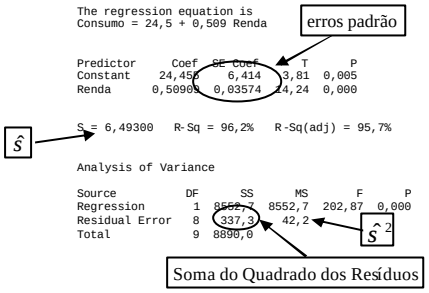
Importantes na construção dos intervalos de confiança

Exemplo –Minitab

Stat > Regression > Regression



Saída Minitab



Inferência sobre os Parâmetros

Hipótese de Normalidade

- Os termos de erro estocástico são independentes e identicamente distribuídos, com $e_i \sim N(0, s^2)$, $i = 1, 2, ..., n$
- Os estimadores têm distribuição normal e tornam-se os melhores estimadores não-viciados dos parâmetros

Normalidade Assintótica dos Estimadores

- Mesmo se os erros não forem normais, se a amostra for suficientemente grande, os estimadores terão uma distribuição aproximadamente normal

Inferência para β_1

- Distribuição do estimador:

$$\hat{b}_1 \sim N\left(b_1, ep\left(\hat{b}_1\right)\right)$$

- Intervalo de confiança

$$\hat{b}_1 - t_{a/2;(n-2)} ep\left(\hat{b}_1\right) \leq b_1 \leq \hat{b}_1 + t_{a/2;(n-2)} ep\left(\hat{b}_1\right)$$

Inferência para β_1 – continuação

- Teste de significância do parâmetro:

H_0: $\beta_1 = 0$ vs *regressão não é significativa*
H_1: $\beta_1 \neq 0$ *regressão significativa*

- Estatística de teste:

$$T = \frac{\hat{b}_1}{ep\left(\hat{b}_1\right)}$$

- Distribuição da estatística de teste:

$$T \sim t_{(n-2)}$$

Inferência para $ß_0$

- Distribuição do estimador:

$$\hat{b}_0 \sim N\left(b_0, ep\left(\hat{b}_0\right)\right)$$

- Intervalo de confiança

$$\hat{b}_0 - t_{a/2;(n-2)} ep\left(\hat{b}_0\right) \le b_0 \le \hat{b}_0 + t_{a/2;(n-2)} ep\left(\hat{b}_0\right)$$

Inferência para $ß_0$ – continuação

- Teste de significância do parâmetro:

H_0: $ß_0 = 0$ vs *intercepto não é significativo*
H_1: $ß_0 \neq 0$ *interecepto é significativo*

- Estatística de teste:

$$T = \frac{\hat{b}_0}{ep\left(\hat{b}_0\right)}$$

- Distribuição da estatística de teste:

$$T \sim t_{(n-2)}$$

Inferência para Parâmetros – Caso Geral

- Teste de Hipóteses de parâmetro:

H_0: $ß_i = ß^0$ vs
H_1: $ß_i \neq ß^0$, $i = 0,1$

- Estatística de teste:

$$T = \frac{\hat{b}_i - b^0}{ep\left(\hat{b}_i\right)}, i = 0,1$$

- Distribuição da estatística de teste:

$$T \sim t_{(n-2)}$$

Exemplo – Lei de Consumo

- Estimativas: $\hat{b}_0 = 24{,}454$, $\hat{b}_1 = 0{,}5091$
- Erros padrão: $ep(\hat{b}_0) = 6{,}4138$, $ep(\hat{b}_1) = 0{,}0357$
- Tamanho da amostra: *10*

Intervalos com *95%* de Confiança

- Do Parâmetro $ß_0$:

$$24{,}4540 - (2{,}306)(6{,}4138) \le b_0 \le 24{,}4540 + (2{,}306)(6{,}4138)$$
$$9{,}664 \le b_0 \le 39{,}244$$

- Do Parâmetro $ß_1$:

$$0{,}5091 - (2{,}306)(0{,}0357) \le b_1 \le 0{,}5091 + (2{,}306)(0{,}0357)$$
$$0{,}4268 \le b_0 \le 0{,}5914$$

Saída Minitab

```
The regression equation is
Consumo = 24,5 + 0,509 Renda

Predictor      Coef  SE Coef      T      P
Constant     24,455    6,414   3,81  0,005
Renda      0,50909  0,03574  14,24  0,000

S = 6,49300   R-Sq = 96,2%   R-Sq(adj) = 95,7%

Analysis of Variance

Source           DF      SS      MS       F      P
Regression        1  8552,7  8552,7  202,87  0,000
Residual Error    8   337,3    42,2
Total             9  8890,0
```

Estatística t

$$T_{b_0} = \frac{24{,}455}{6{,}414} = 3{,}813 \qquad T_{b_1} = \frac{0{,}50909}{0{,}03574} = 14{,}24$$

Avaliação do Modelo

Qualidade do Ajuste

- Ajustada uma equação de regressão entre *X* e *Y*, qual a qualidade do ajuste?
  - √ Análise de variância do modelo
  - √ Análise dos resíduos

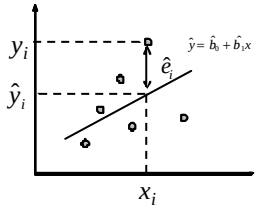
Reta de Regressão e Resíduos

Valores ajustados:

$$\hat{y}_i = b_0 + b_1 x_i$$

Resíduos:

$$\hat{e}_i = y_i - \hat{y}_i$$



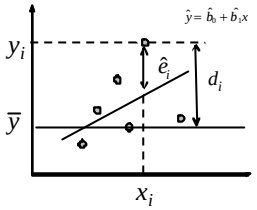
Desvios

Desvio em relação à média aritmética:

$$d_i = y_i - \bar{y}$$

Desvio em relação à reta de regressão:

$$\hat{e}_i = y_i - \hat{y}_i$$



$$y_i - \bar{y} = y_i - \hat{y}_i + \hat{y}_i - \bar{y}$$

Soma dos Quadrados

$$\underbrace{\sum (y_i - \bar{y})^2}_{\text{SQT}} = \underbrace{\sum (\hat{y}_i - \bar{y})^2}_{\text{SQReg}} + \underbrace{\sum (y_i - \hat{y}_i)^2}_{\text{SQRes}}$$

SQT	SQReg	SQRes
variação total	variação explicada pela regressão	variação não explicada

Somas dos Quadrados - Cálculos

$$SQT = \sum (y_i - \bar{y})^2 = S_{yy}$$

$$SQ\,\text{Re}\,g = \sum (y_i - \hat{y}_i)^2 = \hat{b_1}^2 S_{xx}$$

$$SQ\,\text{Re}\,s = SQT - SQ\,\text{Re}\,g$$

Análise de Variância

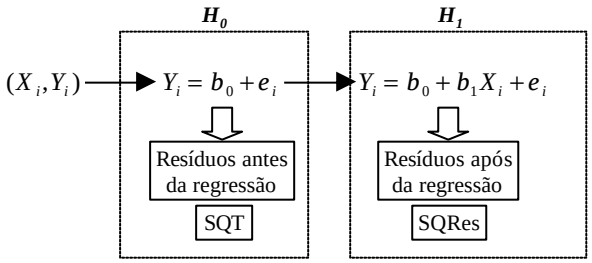
- Objetivo: Ajustar e comparar 2 modelos aos dados
- Modelo alternativo (H_0):

$$Y_i = b_0 + e_i \qquad \hat{b_0} = \overline{Y_i}$$

- Modelo de regressão simples (H_1):

$$Y_i = b_0 + b_1 X_i + e_i$$

Análise de Variância – Estrutura



- *SQReg*: redução na SQT devido ao modelo de regressão
- *SQT = SQReg + SQRes*

Graus de Liberdade

- SQReg: 1
- SQRes: n – 2
- SQT: n – 1

Análise de Variância do Modelo

Fonte de variação	***gl***	***SQ***	*QM*	*Razão*
Regressão	1	$SQReg$	$\frac{SQ\,\mathrm{Re}\,g}{1}$	$F = \frac{QM\,\mathrm{Re}\,g}{QM\,\mathrm{Re}\,s}$
Erro	$n-2$	$SQRes$	$\frac{SQ\,\mathrm{Re}\,s}{n-2}$	
Total	$n-1$	SQT		

Teste de Hipóteses

- Hipóteses:

$H_0: \; Y_i = b_0 + e_i \qquad H_1: \; Y_i = b_0 + b_1 X_i + e_i$

- Estatística de teste:

$F \sim F_{1,(n-2);a}$

No caso da regressão linear simples

$T = \sqrt{F} \sim t_{a/2;(n-2)}$

Exemplo – Lei de Consumo

```
The regression equation is
Consumo = 24,5 + 0,509 Renda

Predictor      Coef  SE Coef      T      P
Constant     24,455    6,414   3,81  0,005
Renda      0,50909  0,03574  14,24  0,000

S = 6,49300   R-Sq = 96,2%   R-Sq(adj) = 95,7%

Analysis of Variance

Source          DF      SS      MS       F      P
Regression       1  8552,7  8552,7  202,87  0,000
Residual Error   8   337,3    42,2
Total            9  8890,0
```

Estatística F

$SQRes = \frac{337,3}{10-2} = 42,2 \qquad F = \frac{85527}{42,2} = 202,87$

Análise de Variância – Interpretação

- Se *SQReg* é elevada, então o modelo de regressão é melhor que o modelo da média amostral;
- Equivale a ß1 ? 0

Coeficiente de Determinação (1)

$$R^2 = \frac{\text{Variação explicada}}{\text{Variação total}} = \frac{\Sigma (\hat{y}_i - \bar{y})^2}{\Sigma (y_i - \bar{y})^2}$$

- Indica a porcentagem de variabilidade que é explicada pelo modelo de regressão

Coeficiente de Determinação (2)

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

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

Coeficiente de Determinação (3)

$$R^2 = \frac{SQReg}{SQT} = \frac{\left(\frac{S_{xy}}{S_{xx}}\right)^2 S_{xx}}{S_{yy}} = r^2$$

- Matematicamente, o coeficiente de determinação é o quadrado do coeficiente de correlação de Pearson

Exemplo – Lei de Consumo

```
The regression equation is
Consumo = 24,5 + 0,509 Renda

Predictor     Coef  SE Coef      T      P
Constant    24,455    6,414   3,81  0,005
Renda      0,50909  0,03574  14,24  0,000

S = 6,49300   R-Sq = 96,2%   R-Sq(adj) = 95,7%

Analysis of Variance

Source          DF       SS       MS       F      P
Regression       1  8552,7  8552,7  202,87  0,000
Residual Error   8   337,3    42,2
Total            9  8890,0
```

R^2

$$R^2 = \frac{8.552,7}{8.890} = 0,962 \qquad R^2 = (0,981)^2 = 0,962$$

Considerações sobre R^2

- Não mede a adequação do modelo linear
- Para comparação entre modelos, é importante observar a variação do erro quadrático médio
- Sua grandeza depende também do intervalo de variação da variável regressora

  valor grande de R^2 pode ser resultado de variação irrealista de x

Abusos Comuns

- Deve-se tomar cuidado na forma do modelo e na seleção das variáveis que serão usadas.
  - √ Forte associação não implica relação causal entre variáveis
- Relações de regressão são válidas somente dentro da faixa dos dados originais de *x*.
  - √ Modelos de regressão não são necessariamente válidos para fins de extrapolação

Predição

Perigos da Predição

- Seja cauteloso ao predizer fora do domínio de variação dos dados.

Valor Ajustado

- É estimador não viciado da linha de regressão ou seja de $E(Y/x_0)$:

$$E\left(\hat{Y}_i\right) = b_0 + b_1 x_i$$

- Variância:

$$Var\left(\hat{Y}_i\right) = s^2\left[\frac{1}{n} + \frac{(x_i - \bar{x})^2}{S_{xx}}\right]$$

Quanto mais afastado do centróide dos dados, mais imprecisa será a estimativa do valor ajustado

Inferência para Y_o

- Intervalo de (1 – a) 100% de confiança em torno da linha de regressão:

$$\hat{Y}_i - t_{a/2;(n-2)}ep\left(\hat{Y}_i\right) \le Y_i \le \hat{Y}_i + t_{a/2;(n-2)}ep\left(\hat{Y}_i\right)$$

Exemplo – Lei de Consumo Keynesiano

- Deseja-se determinar o ponto da reta para o nível de renda de x_i = *$100*.
- O consumo médio estimado para nível de renda *$100* é:

$$\hat{Y}_i = 24{,}455 + 0{,}50909(100) = 75{,}364$$

- A variância da média de consumo para este nível de renda é:

$$Var(\hat{Y}_i) = 42{,}159\left[\frac{1}{10} + \frac{(100-170)^2}{33.000}\right] = 10{,}476$$

$$ep(\hat{Y}_i) = \sqrt{10{,}476} = 3{,}237$$

- Como $t_{0,025;10-2}=2{,}306$, então o intervalo de *95%* de confiança é:

$$75{,}364 - (2{,}306)(3{,}237) \le Y_i \le 75{,}364 + (2{,}306)(3{,}237)$$

$$67{,}899 \le Y_i \le 82{,}829$$

Cálculo do Valor Ajustado – HP

- Entrar cada par de dados:
  √ Y_i ***Enter*** X_i $\boldsymbol{S}_+$
- Para determinar o valor ajustado para nível de renda *$100*:
  √ Digite *100* ***g*** $\boxed{\hat{y},r}$ $\hat{Y}_i = 75{,}364$

Intervalo de Pontos da Reta – Minitab

Stat > Regression > Regression

Minitab – Saída

Regression Analysis: Consumo versus Renda

```
The regression equation is
Consumo = 24,5 + 0,509 Renda


Predictor      Coef  SE Coef      T      P
Constant     24,455    6,414   3,81  0,005
Renda       0,50909  0,03574  14,24  0,000


S = 6,49300   R-Sq = 96,2%   R-Sq(adj) = 95,7%


Analysis of Variance

Source          DF      SS      MS       F      P
Regression       1  8552,7  8552,7  202,87  0,000
Residual Error   8   337,3    42,2
Total            9  8890,0
```

Valor ajustado

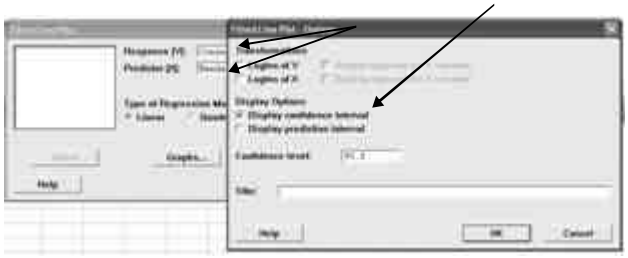
Erro padrão

Tabela de Ajustes e Resíduos

Obs	Renda	Consumo	Fit	SE Fit	Residual	St Resid
1	80	70,00	65,16	3,82	4,82	0,92
2	100	65,00	75,36	3,24	-10,36	-1,84
3	120	90,00	85,55	2,72	4,45	0,76
4	140	95,00	95,73	2,32	-0,73	-0,12
5	160	110,00	105,91	2,08	4,09	0,67
6	180	115,00	116,09	2,08	-1,09	-0,18
7	200	120,00	126,27	2,32	-6,27	-1,03
8	220	140,00	136,45	2,72	3,55	0,60
9	240	155,00	146,64	3,24	8,36	1,49
10	260	150,00	156,82	3,82	-6,82	-1,30

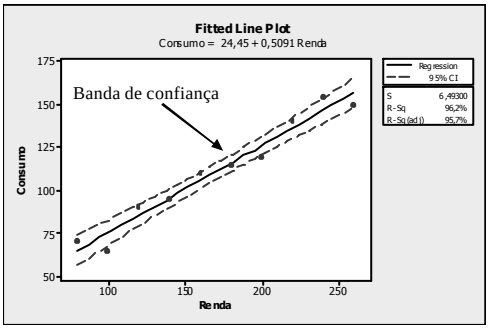
Intervalo de Confiança da Reta – Minitab

Stat > Regression > Fitted Line Plot



Fitted Line Plot

Consumo = 24,45 + 0,5091 Renda



Predição de Nova Observação

- Predição de resposta nova ou futura para um nível x_0 não utilizado na estimação dos parâmetros
- $\tilde{Y}_0$ estimação pontual de Y_0 (resposta não observada:

$$\tilde{Y}_0 = \hat{b}_0 + \hat{b}_1 x_0$$

Valor Ajustado

- É estimador não viciado da nova resposta:

$$E(\tilde{Y}_0) = b_0 + b_1 x_i$$

- Variância:

$$Var(\tilde{Y}_0) = s^2 + s^2\left[\frac{1}{n} + \frac{(x_i - \bar{x})^2}{S_{xx}}\right]$$

Fontes de variabilidade:

1. variação dos estimadores
2. variação natural de Y_0

Inferência para Y_o

- Intervalo de predição de (1 – a) 100% para uma nova observação:

$$\tilde{Y}_0 - t_{a/2;(n-2)}ep(\tilde{Y}_0) \le Y_0 \le \tilde{Y}_0 + t_{a/2;(n-2)}ep(\tilde{Y}_0)$$

- O intervalo de predição em x_0 é sempre mais largo que o intervalo de confiança em x_0.
- A largura do intervalo é mínima quando

Lei de Consumo Keynesiana – Predição

- Consumo para indivíduo não observado com renda mensal x_0 = *$100*

$$\tilde{Y}_0 = 24{,}455 + 0{,}50909(110) = 80{,}455$$

- Erro padrão da predição:

$$Var(\tilde{Y}_0) = 42{,}159\left[1 + \frac{1}{10} + \frac{(110-170)^2}{33.000}\right] = 50{,}974$$

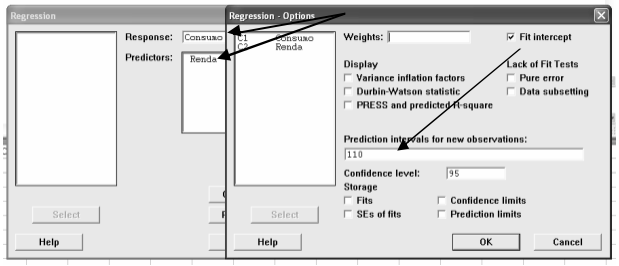
$$ep(\tilde{Y}_0) = \sqrt{50{,}974} = 7{,}140$$

- Intervalo de 95% de confiança da predição:

$$80{,}455 \pm 2{,}306(7{,}140) \qquad 63{,}99 \le \tilde{Y}_0 \le 96{,}92$$

Intervalo de Pontos da Reta – Minitab

Stat > Regression > Regression



Pode-se entrar uma coluna de valores

Saída Minitab

Regression Analysis: Consumo versus Renda

The regression equation is
Consumo = 24,5 + 0,509 Renda

Predictor	Coef	SE Coef	T	P
Constant	24,455	6,414	3,81	0,005
Renda	0,50909	0,03574	14,24	0,000

S = 6,49300 R-Sq = 96,2% R-Sq(adj) = 95,7%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	8552,7	8552,7	202,87	0,000
Residual Error	8	337,3	42,2		
Total	9	8890,0			

intervalo confiança

intervalo predição

Predicted Values for New Observations

New Obs	Fit	SE Fit	95% CI	95% PI
1	80,45	2,97	(73,61; 87,30)	(63,99; 96,92)

Values of Predictors for New Observations

New Obs	Renda
1	110

Aplicação

Um Modelo Econômico

- Objetivo: Estudar a relação entre renda familiar e despesas com alimentação.
- Experimento:

  Amostra aleatória de residências, com renda familiar semanal maior que $480
- Característica de interesse: Despesa semanal da residência com alimentação

  "Quanto foi gasto com alimentação na semana passada?"

Modelo Proposto

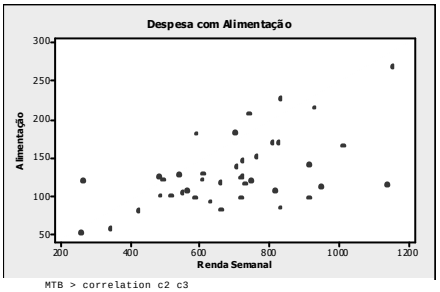
$$Y_i = ß_0 + ß_1 x_i + e_i$$

- Variáveis:

  Resposta (Y_i) : Despesa com alimentação

  Explicativa(x_i): Renda Familiar Semanal

  Erro (e_i): Todos os fatores que afetam Y, exceto renda
- Dados: *alimentacao*

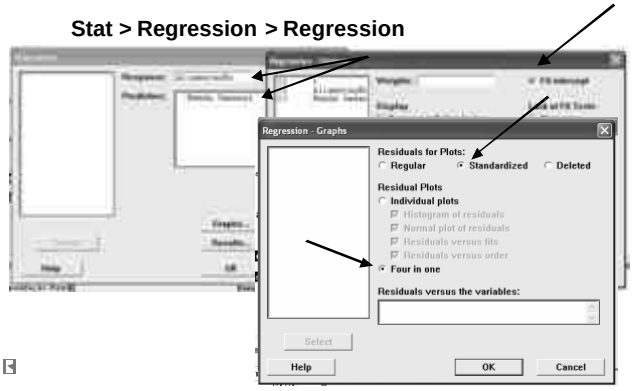



```
MTB > correlation c2 c3

Correlations: Alimentação; Renda Semanal

Pearson correlation of Alimentação and Renda Semanal = 0,546
P-Value = 0,000
```

- Há indícios de associação linear entre renda e despesas com alimentação

Ajuste de Mínimos Quadrados

Stat > Regression > Regression

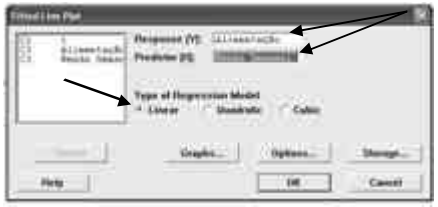


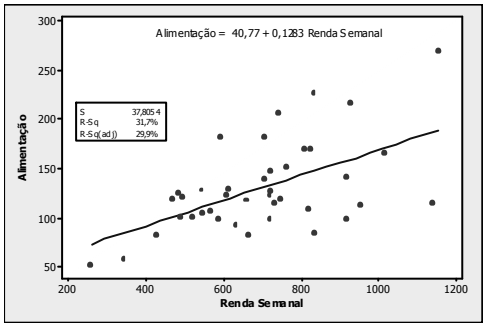

```
Regression Analysis: Alimentação versus Renda Semanal

The regression equation is
Alimentação = 40,8 + 0,128 Renda Semanal


Predictor         Coef  SE Coef     T      P
Constant         40,77    22,14  1,84  0,073
Renda Semanal  0,12829  0,03054  4,20  0,000


S = 37,8054   R-Sq = 31,7%   R-Sq(adj) = 29,9%


Analysis of Variance

Source          DF     SS     MS      F      P
Regression       1  25221  25221  17,65  0,000
Residual Error  38  54311   1429
Total           39  79533
```

- Para cada aumento de $1,00 na renda, estima-se que o consumo;
- Aumento de $100 na renda ➔ estimativa de aumento médio de $12,83 nos gastos com alimentação
- Intercepto sem interpretação, pois, não há observações próximas da renda nula

Reta de Regressão

Stat > Regression > Fitted Line Plot





Dados Importantes de Saída

$ep(\hat{b}_0)$	*22,14*
$ep(\hat{b}_1)$	*0,03054*
$\hat{s}^2$	*1.429*
R^2	*0,317*

32% da variação dos gastos de alimentação é explicada pela renda

Matriz de Variâncias e Covariâncias

- Covariância entre $ß_0$ e $ß_1$:

$$Cov(\hat{b}_0, \hat{b}_1) = -\frac{\bar{x}s^2}{S_{xx}} = -\frac{698(1.429)}{1.532.463} = -0,651$$

- Matriz de variâncias e covariâncias:

$$\begin{bmatrix} 490,18 & -0,651 \\ -0,651 & 0,0009 \end{bmatrix}$$

- Correlação entre os estimadores:

$$corr(\hat{b}_0, \hat{b}_1) = \frac{-0,651}{(22,14)(0,03054)} = -0,9628$$

- Cálculo de S_{xx} no Minitab

```
MTB > Let K1=ssq('Renda Semanal')-(mean('Renda Semanal')**2)*count('Renda Semanal')
MTB > Print K1

Data Display

K1     1532463
```

Intervalos de *95%* de Confiança

- $t_{0,025;40-2} = 2,024$
- $ß_1$: $0,1283 \pm 2,024(0,0305)$ $\quad 0,067 \leq b_1 \leq 0,190$
- $ß_0$: $40,77 \pm 2,024(22,14)$ $\quad -4,041 \leq b_0 \leq 85,581$

Testes de Hipóteses - Significância

- Nível de significância: $a = 5\%$
- Valor crítico: $t_{0,025;40-2} = 2,024$
- $H_0: ß_0 = 0$ vs $H_1: ß_0 \neq 0$
  - √ Estatística T (saída Minitab): *1,84*
  - √ Comparação com valor crítico: *1,84 < 2,024*
  - √ Não há evidências para rejeitar a hipótese de que o intercepto seja zero.

- $H_0: ß_1 = 0$ vs $H_1: ß_1 \neq 0$
  - √ Estatística T (saída Minitab): *4,20*
  - √ Comparação com valor crítico: *4,20 > 2,024*
  - √ Há evidências para considerar a regressão significante.

- $H_0: ß_1 = 0,10$ vs $H_1: ß_1 \neq 0,10$
  - √ Estatística T: $T = \frac{\hat{b}_1 - b^0}{ep(\hat{b}_1)} = \frac{0,12829 - 0,10}{0,03054} = 0,93$
  - √ Comparação com valor crítico: *0,93 < 2,024*
  - √ Não rejeitamos a hipótese que $ß_1 = 0,10$

- H_0: ß$_1$ = 0 vs H_1: ß$_1$ > 0
  - √ Valor crítico: $t_{0,05;40-2} = 1,686$
  - √ Estatística T (saída Minitab): *4,20*
  - √ Comparação com valor crítico: *4,20 > 1,686*
  - √ Há evidências para considerar a inclinação da reta regressão crescente.

Predição

- Despesa mensal com alimentação para residência com renda mensal x_0 = *$750*

$$\tilde{Y}_0 = 40,77 + 0,12829(750) = 136,98$$

- Erro padrão da predição:

$$Var(\tilde{Y}_0) = 1.429\left[1 + \frac{1}{40} + \frac{(750-698)^2}{1.532.463}\right] = 1.467,246$$

$$ep(\tilde{Y}_0) = \sqrt{1.467,246} = 38,305$$

- Intervalo de 95% de confiança da predição:

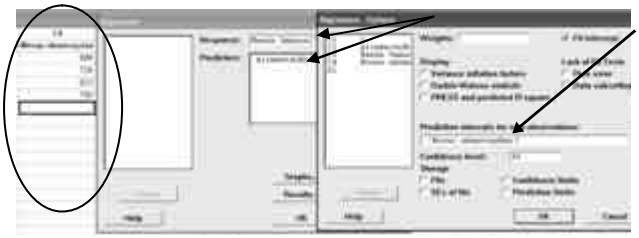
$$136,98 \pm 2,024(38,305) \qquad 59,45 \le \tilde{Y}_0 \le 214,51$$

Predição (2)

- O intervalo é grande. A predição não é confiável

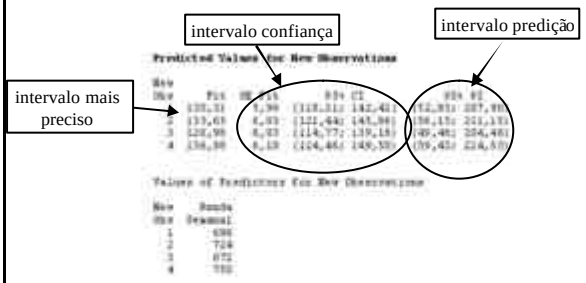
Intervalos de Confiança e de Predição – Minitab

Stat > Regression > Regression

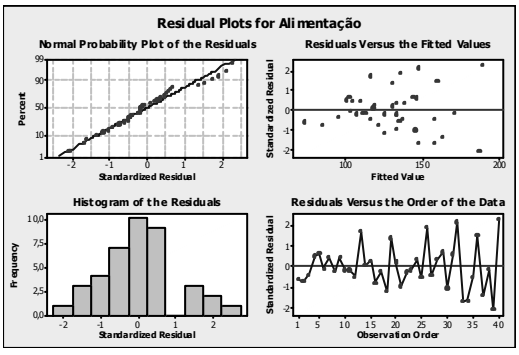


- Entrando com uma coluna de níveis não observados

Saída Minitab



Análise dos Resíduos



Referências

Bibliografia Recomendada

- Gujarati, D. N. (Pearson)
  Econometria Básica
- Hill, R. C., Griffiths, W. E. e Judge, G. (Saraiva)
  Econometria