# Unidade 2 - Regressão Linear Simples

## 1. Introdução

A regressão linear simples constitui uma tentativa de estabelecer uma equação matemática linear que descreva o relacionamento entre duas variáveis.

Há diversas maneiras em que as equações de regressão são utilizadas:

1. Em situações em que as duas variáveis medem aproximadamente a mesma coisa, mas uma delas é relativamente dispendiosa, ou difícil de lidar, enquanto a outra não. A finalidade de uma equação de regressão, nesse caso, seria *estimar* valores de uma variável, com base nos valores conhecidos da outra.
2. Explicar valores de uma variável em termos da outra. Isto é, a análise de regressão apenas indica qual relacionamento matemático pode existir entre duas variáveis, se existir algum.
3. *Predizer* valores futuros de uma variável.

Embora as relações entre variáveis possam assumir uma grande variedade de formas, nos limitaremos ao estudo das equações lineares simples. As equações lineares são importantes porque servem para aproximar muitas relações da vida real, e porque são relativamente fáceis de lidar e de interpretar. Outras formas de análise de regressão, tais como regressão múltipla (mais de duas variáveis) e regressão polinomial (não-linear) envolvem extensões dos mesmos conceitos usados na regressão linear simples.

É importante ter em mente que nem todas as situações são bem aproximadas por uma equação linear. Por isso, em geral, é necessário desenvolver um trabalho preliminar para determinar se um modelo linear é adequado. O processo mais simples consiste em construir um diagrama de dispersão e avaliar se a relação linear parece razoável. Quando os dados não podem ser aproximados por um modelo linear, as alternativas são procurar um modelo não-linear conveniente, ou transformar os dados para a forma linear. Por exemplo, a conversão de uma ou de ambas as variáveis para escala logarítmica. Observe as figuras abaixo:

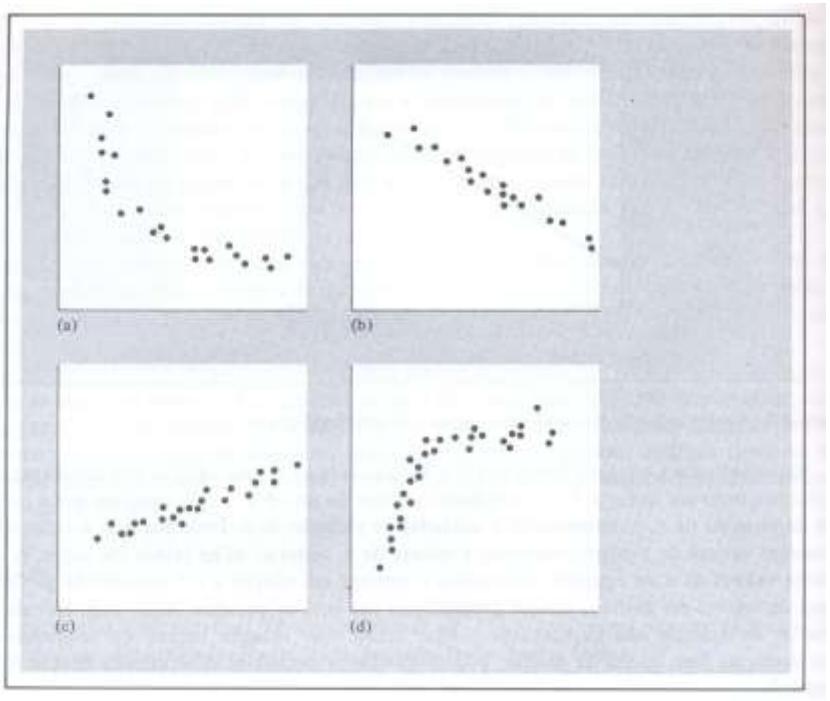



Figura 5 – Nem toda relação entre duas variáveis é linear. Os gráficos b e c parecem dispor-se segundo um padrão linear, o que não ocorre com a e d.
*Fonte: Stevenson, 2001*

# 2. Modelo de Regressão e Equação de Regressão

Uma vez que assumimos uma relação linear entre as variáveis, o **modelo de regressão linear simples** será:

$$y = \beta_0 + \beta_1 x + \varepsilon$$

onde:

y $\rightarrow$ É a variável dependente (ou resposta), é a variável que imaginamos depender de x.

x $\rightarrow$ É a variável independente (ou preditora).

$\beta_0 \rightarrow$ É o parâmetro que representa o coeficiente linear (ou intercepto) da reta.

$\beta_1 \rightarrow$ É o parâmetro que representa o coeficiente angular (ou inclinação) da reta.

$\varepsilon \rightarrow$ É o erro aleatório (vindo de erros de medidas e/ de ausência de outras variáveis independentes também consideradas importantes para explicar a variável resposta).

Mesmo que o modelo acima descrito seja um modelo simples, ainda assim devemos fazer algumas suposições à respeito do erro aleatório ($\varepsilon$). As suposições são necessárias para que possamos fazer inferências sobre alguma previsão e parâmetros do modelo (teste de hipóteses, construção de intervalos de confiança).

As suposições são:

- ✓ Os erros se distribuem normalmente com média zero e variância constante $\sigma^2$.
- ✓ Os erros não são correlacionados, ou seja, o fato de um erro ser maior não tende a elevar o valor de um outro erro.

Usando apenas os dados amostrais não podemos obter os valore exatos dos parâmetros $\beta_0$ e $\beta_1$. Esses parâmetros deverão ser estimados com base nos dados amostrais. A equação da reta com os parâmetros estimados é representada por:

$$\hat{y} = b_0 + b_1 x$$

Onde:

$b_0 \rightarrow$ é uma estimativa do $\beta_0$ e representa o coeficiente linear (ou intercepto) da reta estimada.

$b_1 \rightarrow$ é uma estimativa do $\beta_1$ e representa o coeficiente angular (ou inclinação) da reta estimada.

Como é impossível a reta passar por todos os pontos, sempre haverá diferença entre algum valor observado $y_i$ e o valor ajustado pela reta $\hat{y}_i$. Essa diferença $y_i - \hat{y}_i$ é denominada de resíduo (ou erro de estimação). Observe a figura abaixo:

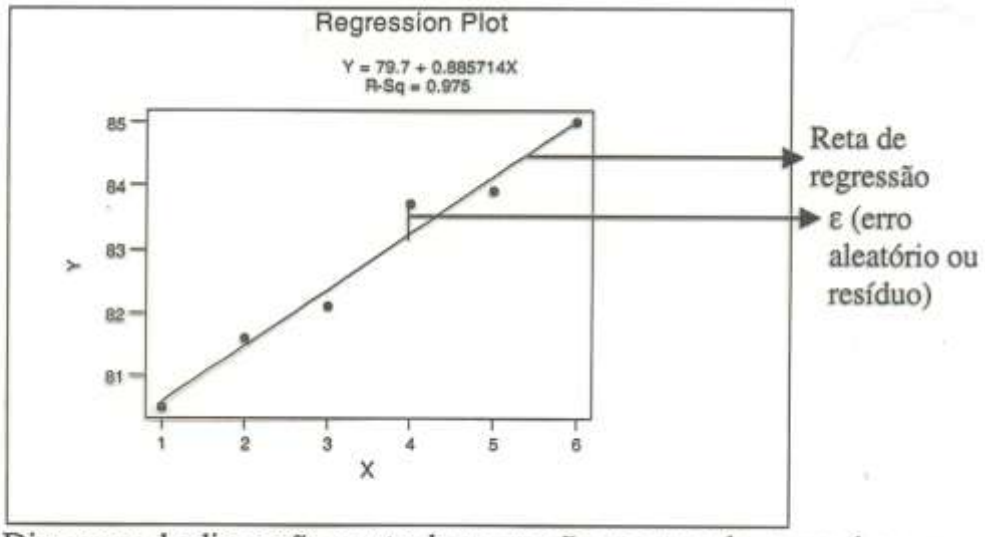



Figura 3 Diagrama de dispersão e reta de regressão mostrando o que é o erro aleatório ($\varepsilon$)

# 3. Método dos Mínimos Quadrados

O *Método dos Mínimos Quadrados* é um critério que utiliza os dados da amostra para obter os valores de $b_0$ e $b_1$ que minimiza a soma dos quadrados dos resíduos. Com essas estimativas conseguimos tornar os resultados tão menores quanto possível e ajustar a reta que chamamos de reta de regressão (ou *reta ótima* ou *reta de melhor ajuste* ou *reta de mínimos quadrados*).

As fórmulas do *Método de Mínimos Quadrados* utilizadas para obter os valores de $b_0$ e $b_1$ são:

$$b_1 = \frac{n \times \left(\sum x.y\right) - \left(\sum x\right) \times \left(\sum y\right)}{n \times \left(\sum x^2\right) - \left(\sum x\right)^2}$$

$$b_0 = \bar{y} - \left(b_1 \times \bar{x}\right)$$

Os valores de $b_0$ e $b_1$ podem ser interpretados da seguinte maneira:
$b_0 \rightarrow$ Normalmente, só é interpretado quando, na prática, a variável x puder assumir valor igual a zero (x = 0). Dessa maneira, $b_0$ representaria a média de y quando x = 0.
$b_1 \rightarrow$ Representa a variação de y por unidade de variação de x.

**Exemplo 1 –** (Stevenson, 2001) Um pesquisador está interessado em estimar o preço de venda de veículos de determinada marca, modelo e ano de fabricação em função da quilometragem do veículo. Os dados encontram-se a seguir:

| Carro | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Quilometragem (em 1000 km) | 40 | 30 | 30 | 25 | 50 | 60 | 65 | 10 | 15 | 20 | 55 | 40 | 35 | 30 |
| Preço de venda (em dólares) | 1000 | 1500 | 1200 | 1800 | 800 | 1000 | 500 | 3000 | 2500 | 2000 | 800 | 1500 | 2000 | 2000 |

Primeiramente vamos construir o diagrama de dispersão para avaliar se uma reta descreve adequadamente os dados:

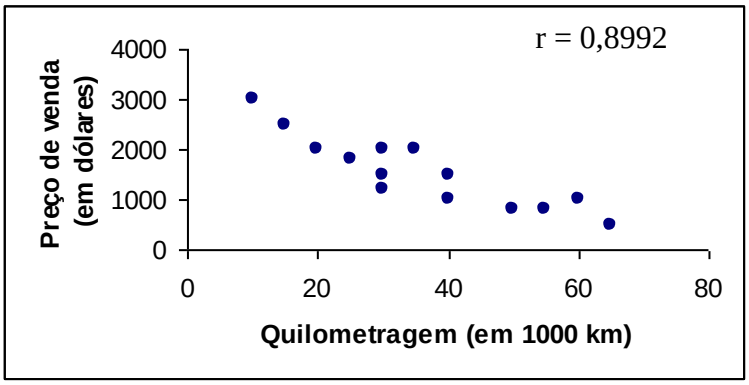



Figura 6 – Diagrama de dispersão

Parece que uma relação linear é razoavelmente consistente com os dados amostrais. Dessa maneira podemos prosseguir com a construção da equação de regressão pelo método de mínimos quadrados:

| Carro | Quilometragem (em 1000 km) x | Preço de vendas (em dólares) y | x.y | $x^2$ | $y^2$ |
|---|---|---|---|---|---|
| 1 | 40 | 1000 | 40000 | 1600 | 1000000 |
| 2 | 30 | 1500 | 45000 | 900 | 2250000 |
| 3 | 30 | 1200 | 36000 | 900 | 1440000 |
| 4 | 25 | 1800 | 45000 | 625 | 3240000 |
| 5 | 50 | 800 | 40000 | 2500 | 640000 |
| 6 | 60 | 1000 | 60000 | 3600 | 1000000 |
| 7 | 65 | 500 | 32500 | 4225 | 250000 |
| 8 | 10 | 3000 | 30000 | 100 | 9000000 |
| 9 | 15 | 2500 | 37500 | 225 | 6250000 |
| 10 | 20 | 2000 | 40000 | 400 | 4000000 |
| 11 | 55 | 800 | 44000 | 3025 | 640000 |
| 12 | 40 | 1500 | 60000 | 1600 | 2250000 |
| 13 | 35 | 2000 | 70000 | 1225 | 4000000 |
| 14 | 30 | 2000 | 60000 | 900 | 4000000 |
| Soma= | 505 | 21600 | 640000 | 21825 | 39960000 |

Utilizando as fórmulas do *Método de Mínimos Quadrados* temos:

$$b_1 = \frac{n \times \left(\sum x.y\right) - \left(\sum x\right) \times \left(\sum y\right)}{n \times \left(\sum x^2\right) - \left(\sum x\right)^2} = \frac{\left(14 \times 640000\right) - \left(505 \times 21600\right)}{\left(14 \times 21825\right) - \left(505\right)^2} = -38{,}555 \text{ dólares}$$

$b_1$ é o coeficiente angular (ou inclinação) da reta. Significa uma diminuição de 38,555 dólares no preço de venda do carro para cada aumento de 1000 km na quilometragem do veículo.

$$b_0 = \bar{y} - \left(b_1 \times \bar{x}\right) = \frac{21600}{14} - \left(-38{,}555 \times \frac{505}{14}\right) = 2933{,}6 \text{ dólares}$$

$b_0$ é o coeficiente linear (ou intercepto) da reta (valor de y para x = 0). Para um carro 'novo' (quilometragem = 0) o preço médio de venda seria de 2933,60 dólares.

Observe que tanto $b_0$ quanto $b_1$ têm a mesma unidade de medida de y.

O gráfico dos dados com o ajuste da reta de regressão é mostrado a seguir:

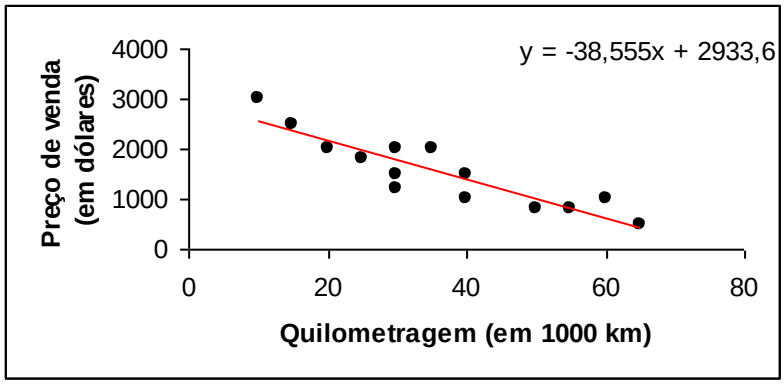



Figura 7 – Diagrama de dispersão com ajuste da reta de regressão

Se quiséssemos calcular os resíduos (ou erro aleatório) do modelo de regressão ajustado aos dados teríamos que calcular os valores preditos ($\hat{y}_i$) pela reta de regressão e posteriormente, calcular a diferença $y_i - \hat{y}_i$ que é denominada de resíduo (ou erro de estimação). Observe:

| Carro | Quilometragem (em 1000 km) x | Preço de vendas (em dólares) y | Valores preditos $\hat{y}_i$ =-38,555x+2933,6 | Resíduo $\varepsilon = y_i - \hat{y}_i$ |
|---|---|---|---|---|
| 1 | 40 | 1000 | 1391 | 391,4 |
| 2 | 30 | 1500 | 1777 | 277 |
| 3 | 30 | 1200 | 1777 | 577 |
| 4 | 25 | 1800 | 1970 | 169,7 |
| 5 | 50 | 800 | 1006 | 205,9 |
| 6 | 60 | 1000 | 620,3 | -379,7 |
| 7 | 65 | 500 | 427,5 | -72,47 |
| 8 | 10 | 3000 | 2548 | -452 |
| 9 | 15 | 2500 | 2355 | -144,7 |
| 10 | 20 | 2000 | 2163 | 162,5 |
| 11 | 55 | 800 | 813,1 | 13,07 |
| 12 | 40 | 1500 | 1391 | -108,6 |
| 13 | 35 | 2000 | 1584 | -415,8 |
| 14 | 30 | 2000 | 1777 | -223,1 |

# 4. Teste de Hipóteses sobre o Coeficiente Angular da Reta de Regressão

Ao realizarmos uma análise de regressão estimamos os parâmetros populacionais $\beta_0$ e $\beta_1$ utilizando dados amostrais por meio dos valores de $b_0$ e $b_1$. É, pois, importante testar os resultados de tais cálculos a fim de decidir se são significativos (isto é, se os verdadeiros parâmetros não são nulos). O que queremos é distinguir entre situações em que as variáveis são relacionadas, e situações em que não o são. Se não há relacionamento, é de se esperar um coeficiente angular igual a zero. Demos então testar as hipóteses:

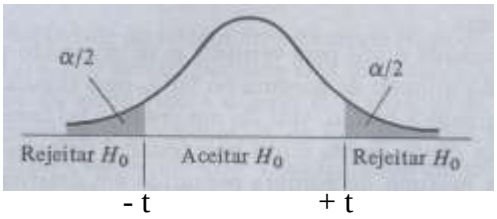
$H_0$: $\beta_1 = 0$ vs. $H_1$: $\beta_1 \neq 0$

A estatística de teste a ser utilizada é dada por

$$t = \frac{b_1}{\sqrt{\frac{\sum y^2 - \left(b_0 \times \sum y\right) - \left(b_1 \times \sum x.y\right)}{(n-2) \times \left(\sum x^2 - \frac{\left(\sum x\right)^2}{n}\right)}}}$$

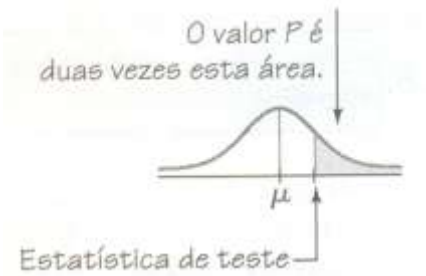
A distribuição a ser utilizada é a t-Student com n – 2 graus de liberdade.

A região crítica é definida como:




- t + t

O valor p para esse teste é obtido a partir de:




Lembrando que o valor da estatística de teste deverá ser utilizado em módulo para o cálculo do valor p!

**Exemplo 1 – (Continuação)** Utilizando os dados desse exemplo, teste a hipótese de que o coeficiente angular é igual a zero. Considere um nível de 5% de significância.

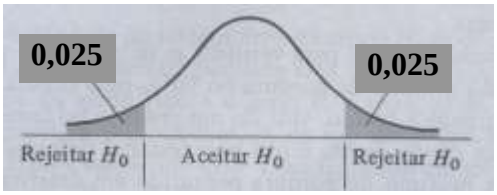
Definindo as hipóteses: $H_0$: $\beta_1 = 0$ vs. $H_1$: $\beta_1 \neq 0$

Calculando a estatística de teste:

$$t = \frac{-38{,}555}{\sqrt{\frac{3996000 - (2933{,}6 \times 21600) - (-38{,}555 \times 640000)}{(14-2) \times \left(21825 - \frac{(505)^2}{14}\right)}}} = 5{,}402$$

Grau de liberdade: n – 2 = 14 – 2 =12.

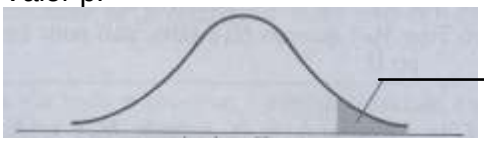
Região crítica:




Rejeita-se $H_0$

- 2,179 +2,179

Valor p:



Valor p = 2 x 0,0005 = 0,001

+5,402

Conclusão: Rejeita-se $H_0$ a um nível de 5% de significância, ou seja, existe uma relação significativa entre o preço de venda do veículo e sua quilometragem. A chance de errar ao afirmar que $\beta_1 \neq 0$, com base na amostra observada, é igual a 0,1%.

## 5. Coeficiente de Determinação

Uma medida útil, associada à reta de regressão, é o coeficiente de determinação ($R^2$). O coeficiente de determinação mede a proporção da variação em Y que é explicada pela equação de regressão estimada. Quanto maior for o valor de $R^2$, maior será a proporção da variação em Y explicada pela equação estimada. É muito usado para julgar a adequação de um modelo de regressão.

O coeficiente de determinação é obtido pela razão entre a 'variação explicada' e a 'variação total':

$$R^2 = \frac{Variação\ explicada}{Variação\ total}$$

Para compreendermos melhor o cálculo do $R^2$ é necessário entendermos os seguintes conceitos:

- ✓ Variação total: É a variação dos pontos em torno de $\bar{y}$ (média dos valores de y). É calculada como:

$$Variação\ total = \sum (y_i - \bar{y})^2$$

- ✓ Variação não-explicada: São os desvios verticais dos $y_i$'s em relação à reta de regressão. Essa variação é chamada de 'não-explicada' porque não pode ser explicada somente pelo valor de x, isto é, ainda há uma dispersão mesmo depois de se levar em conta a reta. É calculada como:

$$Variação\ não\text{-}explicada = \sum (y_i - \hat{y})^2$$

A figura a seguir mostra a diferença entre a 'variação total' e a 'variação não-explicada':

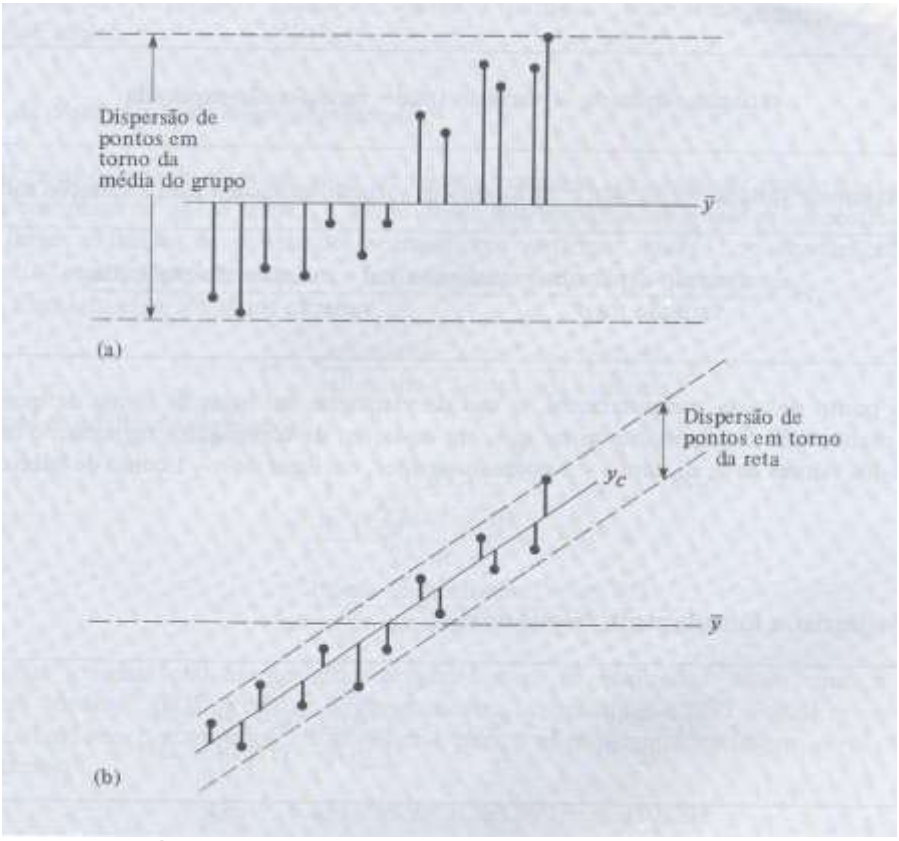



Figura 8 – O gráfico **a** mostra a variação total, ou seja, as distâncias entre os pontos e a média de y. O gráfico **b** mostra a variação não explicada, ou seja, as distâncias entre os pontos e a reta de regressão.
*Fonte: Stevenson, 2001*

- ✓ Variação explicada: É a diferença entre a variação total e a variação não-explicada:

$$Variação\ explicada = Variação\ total - Variação\ não\ explicada$$

O valor de $R^2$ pode variar de 0 a 1, sendo que, quanto mais próximo de 1 estiver o valor do coeficiente de determinação, maior é a proporção da variação de y que está sendo explicada pela reta de regressão.

**Exemplo 1 – (Continuação)** Voltando aos dados do exemplo do preço de venda dos carros em função da quilometragem. Calcule e interprete o coeficiente de determinação.

Para calcularmos o $R^2$ devemos calcular as variações total, não explicada e explicada:

| Carro | Quilometragem (em 1000 km) x | Preço de vendas (em dólares) y | Valores preditos $\hat{y}_i$ =-38,555x+2933,6 | Resíduo $\varepsilon = y_i - \hat{y}_i$ | Variação não explicada $(y_i - \hat{y}_i)^2$ | $(y - \bar{y})$ | Variação total $(y - \bar{y})^2$ |
|---|---|---|---|---|---|---|---|
| 1 | 40 | 1000 | 1391,40 | 391,40 | 153193,96 | -542,86 | 294693,88 |
| 2 | 30 | 1500 | 1776,95 | 276,95 | 76701,30 | -42,86 | 1836,73 |
| 3 | 30 | 1200 | 1776,95 | 576,95 | 332871,30 | -342,86 | 117551,02 |
| 4 | 25 | 1800 | 1969,73 | 169,73 | 28806,58 | 257,14 | 66122,45 |
| 5 | 50 | 800 | 1005,85 | 205,85 | 42374,22 | -742,86 | 551836,73 |
| 6 | 60 | 1000 | 620,30 | -379,70 | 144172,09 | -542,86 | 294693,88 |
| 7 | 65 | 500 | 427,53 | -72,47 | 5252,63 | -1042,86 | 1087551,02 |
| 8 | 10 | 3000 | 2548,05 | -451,95 | 204258,80 | 1457,14 | 2123265,31 |
| 9 | 15 | 2500 | 2355,28 | -144,73 | 20945,33 | 957,14 | 916122,45 |
| 10 | 20 | 2000 | 2162,50 | 162,50 | 26406,25 | 457,14 | 208979,59 |
| 11 | 55 | 800 | 813,08 | 13,07 | 170,96 | -742,86 | 551836,73 |
| 12 | 40 | 1500 | 1391,40 | -108,60 | 11793,96 | -42,86 | 1836,73 |
| 13 | 35 | 2000 | 1584,18 | -415,83 | 172910,43 | 457,14 | 208979,59 |
| 14 | 30 | 2000 | 1776,95 | -223,05 | 49751,30 | 457,14 | 208979,59 |
| Soma= | 505 | 21600 | | | 1269609,11 | | 6634285,71 |

- ✓ $Variação\ total = \sum (y_i - \bar{y})^2 = 6634285{,}71$
- ✓ $Variação\ não\text{-}explicada = \sum (y_i - \hat{y})^2 = 1269609{,}11$
- ✓ $Variação\ explicada = Variação\ total - Variação\ não\ explicada = 5364676{,}6$

Calculando o coeficiente de determinação:

$$R^2 = \frac{Variação\ explicada}{Variação\ total} = \frac{5364676{,}6}{6634285{,}71} = 0{,}8086$$

Podemos interpretar o coeficiente de determinação da seguinte maneira: 80,86% da variação total no preço de venda dos veículos pode ser explicada pela variação na quilometragem dos mesmos através da equação de regressão estimada. Os outros 19,14% restantes são explicados por outros fatores além da quilometragem do veículo e que não foram incluídos no modelo.

## 6. Intervalo de Previsão para um Valor Individual de y

No início dessa unidade vimos que a análise de regressão pode ser utilizada para predizer valores futuros. À partir de uma equação de regressão do tipo: $\hat{y} = b_0 + b_1 x$, podemos predizer o valor de $\hat{y}$ para um dado valor de $x_0$. Porém, o resultado dessa predição será um valor único, ou seja, uma estimativa pontual. Já sabemos que as estimativas pontuais têm a desvantagem de não dar qualquer idéia de sua precisão. Para superar essa desvantagem podemos construir um intervalo de predição para o valor de $\hat{y}$, que é uma estimativa intervalar de confiança de um valor predito.

Para estabelecermos um intervalo de predição vamos trabalhar basicamente com a mesma idéia da construção de intervalos de confiança, ou seja, para um dado valor fixo $x_i$, o intervalo de predição para um determinado $\hat{y}$ é: $\hat{y} \pm$ E, onde a margem de erro (E) é obtida da seguinte maneira:

$$E = t_{\frac{\alpha}{2}} \times s_e \times \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum x^2 - \left[\frac{(\sum x)^2}{n}\right]}}$$

Onde:

$t_{\frac{\alpha}{2}} \rightarrow$ É o valor da distribuição t-Student com n – 2 graus de liberdade.

$s_e \rightarrow$ É o *erro padrão da estimativa* sendo calculado da seguinte maneira:

$$s_e = \sqrt{\frac{\left(\sum y^2\right) - \left(b_0 \times \sum y\right) - \left(b_1 \times \sum xy\right)}{n-2}}$$

**Exemplo 1 – (Continuação)** Utilizando a equação de regressão encontrada para os dados desse exemplo:

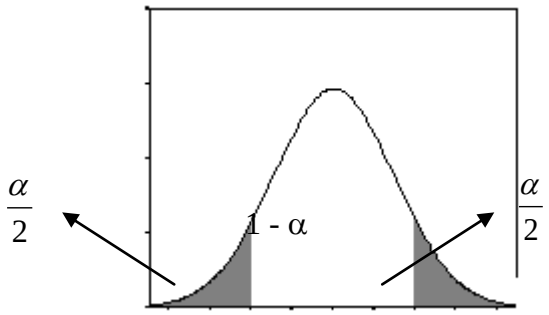
a) Encontre o valor predito de venda para um carro que tem 45 (em 1000 km rodados).
b) Construa o intervalo de predição com 95% de confiança para o valor predito em a).

A equação de regressão encontrada foi: $\hat{y}$ = -38,555x + 2933,6

a) Para uma valor de $x_0$ = 45 o valor predito é: $\hat{y} = (-38{,}555 \times 45) + 2933{,}6 = 1198{,}625$, ou seja, para um carro com 45000 km rodados, o valor predito de venda é de US$1198,62.

b) Para um intervalo de 95% de confiança, temos que nos lembrar que:

$1 - \alpha = 0{,}95$
$\alpha = 0{,}05$
$\frac{\alpha}{2} = 0{,}025$




O número de graus de liberdade é: n – 2 = 14 – 2 = 12.
O valor crítico é: $t_{\frac{\alpha}{2}}$ = 2,179

O erro padrão da estimativa é:

$$s_e = \sqrt{\frac{\left(\sum y^2\right) - \left(b_0 \times \sum y\right) - \left(b_1 \times \sum xy\right)}{n-2}}$$

$$s_e = \sqrt{\frac{39960000 - (2933{,}6 \times 21600) - (-38{,}555 \times 640000)}{14-2}}$$

$$s_e = 325{,}249$$

A margem de erro é:

$$E = t_{\frac{\alpha}{2}} \times s_e \times \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum x^2 - \left[\frac{(\sum x)^2}{n}\right]}}$$

$$E = 2{,}179 \times 325{,}249 \times \sqrt{1 + \frac{1}{14} + \frac{\left(45 - \frac{505}{14}\right)^2}{21825 - \left[\frac{(505)^2}{14}\right]}}$$

$$E = 741{,}116$$

Finalmente, o intervalo de predição é: $\hat{y} \pm$ E = 1198,62 $\pm$ 741,116 = [457,50 ; 1939,74]

Assim, para uma carro com 45000 km rodados, temos 95% de confiança de que seu verdadeiro valor de venda esteja entre US$457,50 e US$1939,74. Trata-se de um intervalo extremamente grande. Um fator que contribui para isso é o pequeno tamanho da amostra.

Além de sabermos que o preço de venda predito é de US$1198,62, temos agora uma idéia da confiabilidade real daquela estimativa. O intervalo de predição de 95% mostra que a estimativa de US$1198,62 pode variar substancialmente.

## 7. Diretrizes para o uso da Equação de Regressão

**a)** Se não há correlação linear significativa, não use a equação de regressão para fazer predições.
**b)** Ao aplicar a equação de regressão para predições mantenha-se dentro do âmbito dos dados amostrais. Se acharmos uma equação de regressão relacionando as alturas das mulheres com os números de seus sapatos, é absurdo predizer o número do sapato de uma mulher que tenha 3 metros de altura!
**c)** Uma equação de regressão baseada em dados passados não é necessariamente válida hoje. A equação de regressão que relaciona preços de carros usados e idades de carros não é mais válida, se se baseia em dados da década de 1970.
**d)** Não devemos fazer predições sobre uma população diferente daquela de onde provêm os dados amostrais. Coletam-se dados amostrais sobre homens e estabelecemos uma equação de regressão relacionando idade e uso de controle remoto de TV, os resultados não se aplicam necessariamente às mulheres.

## 8. Exercício de Revisão

A tabela mostrada a seguir relaciona o tempo (em horas) que 11 alunos gastaram estudando e as respectivas notas num teste:

| Horas | 2,5 | 3 | 4 | 4,5 | 5 | 5,5 | 6 | 6 | 7 | 8,5 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Notas | 70 | 74 | 75 | 80 | 82 | 85 | 89 | 90 | 91 | 93 | 95 |

a) Identifique as variáveis *x* e *y*.
b) Calcule e interprete o coeficiente de correlação para os dados acima.
c) Teste a hipótese de que o coeficiente de correlação é diferente de zero. Considere um nível de significância de 2%. Qual conclusão você obteve?
d) Encontra a reta de mínimos quadrados para esse conjunto de dados.
e) Interprete os coeficientes linear e angular da reta de regressão.
f) Encontre os valores dos resíduos do modelo para os alunos que gastaram 3, 6 e 10 horas de estudo.
g) Teste a hipótese de que o coeficiente angular é igual a zero. Considere um nível de significância de 2%. Qual conclusão você obteve?
h) Calcule e interprete o coeficiente de determinação para os dados do problema.
i) Encontre o valor predito de nota para um aluno que ficou 9 horas estudando para o teste. Calcule e interprete o intervalo de predição para o valor predito com 98% de confiança.