ESTATÍSTICA II

Michelle Hanne Soares de Andrade
michellehanne@cefetmg.br
1º. SEMESTRE 2018

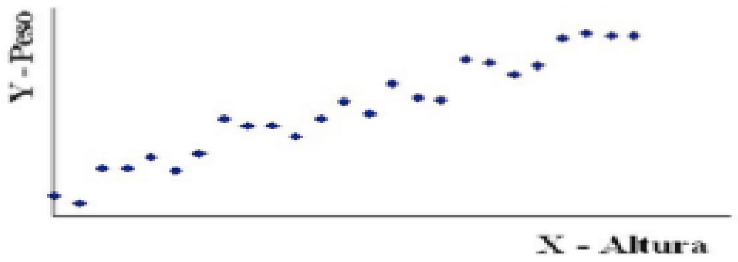
REGRESSÃO LINEAR SIMPLES

Introdução

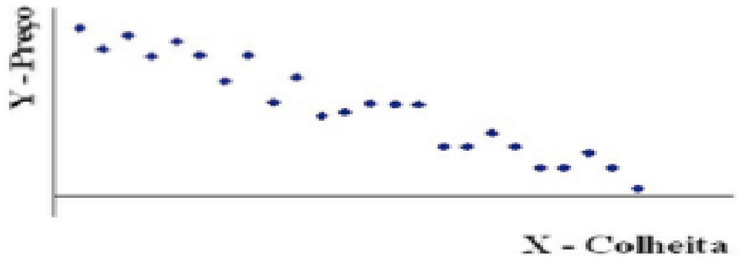
- A análise de regressão estuda a relação entre uma variável chamada variável dependente e outras variáveis chamadas variáveis independentes.
- A relação entre elas é representada por um modelo matemático, que associa a variável dependente com as variáveis independentes.

Introdução

Diagramas de dispersão que sugerem uma regressão linear entre as variáveis



Existência de correlação positiva (em média, quanto maior for a altura maior será o peso)



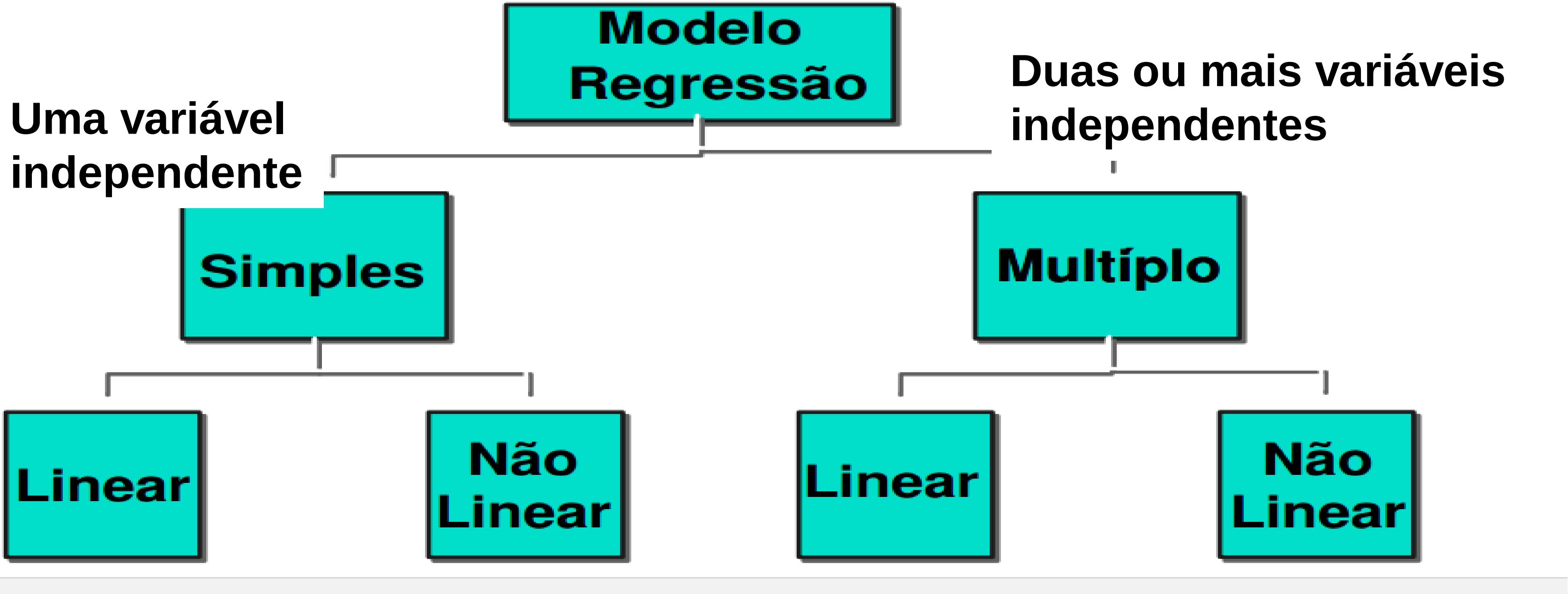
Existência de correlação negativa (em média, quanto maior for for a colheita menor será o preço

Introdução

- Este modelo é designado por modelo de regressão linear simples (MRLS) se define uma relação linear entre a variável dependente e uma variável independente.

- Se em vez de uma, forem incorporadas várias variáveis independentes, o modelo passa a denominar-se **modelo de regressão linear múltipla**.

Introdução



Modelo Regressão
Uma variável independente
Duas ou mais variáveis independentes
Simples
Multíplo
Linear
Não Linear
Linear
Não Linear

Introdução

- A presença ou ausência de **relação linear** pode ser investigada sob dois pontos de vista:

- Quantificando a força dessa relação: **correlação**.
- Explicitando a forma dessa relação: **regressão**.

Introdução

- No MRLS vamos estudar a relação linear entre duas variáveis quantitativas.

- **Exemplos:**
- Altura dos pais e altura dos filhos;
- Renda semanal e despesas de consumo;
- Variação dos salários e taxa de desemprego;
- Demanda dos produtos de uma firma e publicidade

Diagrama de dispersão

- Os dados para a análise de regressão e correlação simples são da forma:

$$(x_1, y_1), (x_2, y_2), \ldots, (x_i, y_i), \ldots, (x_n, y_n)$$

- Com base nos dados constrói-se o diagrama de dispersão, que deve exibir uma tendência linear para que se possa usar a regressão linear.

Diagrama de dispersão

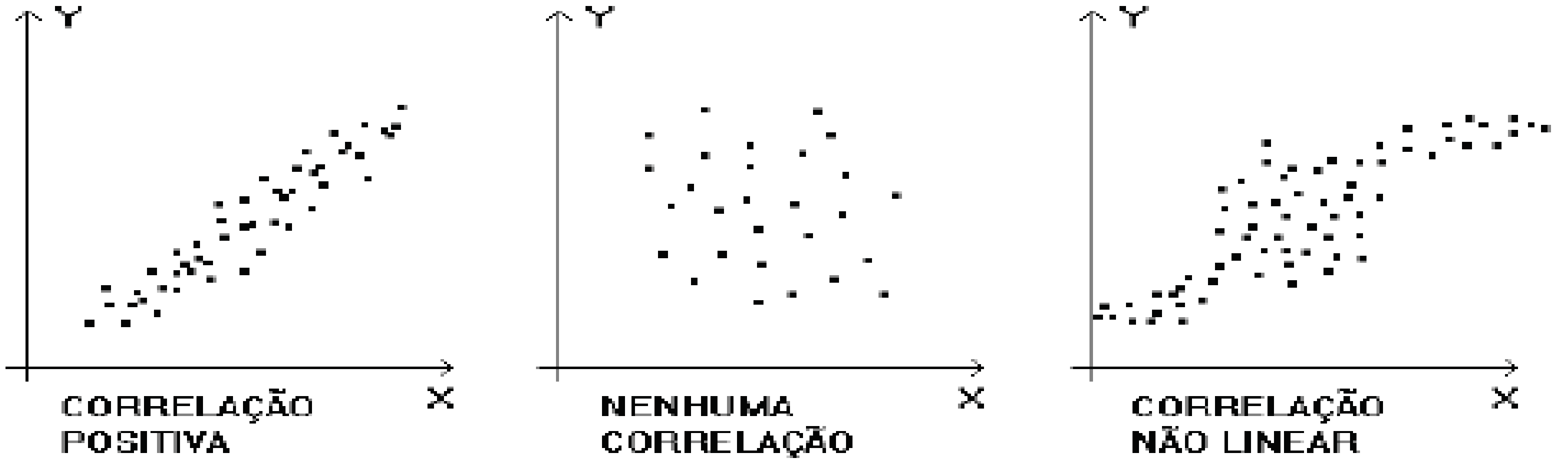


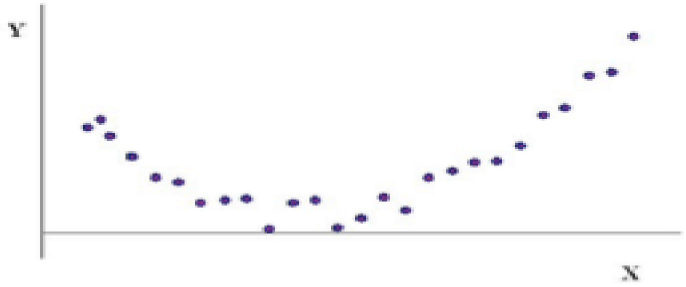
Diagrama de dispersão

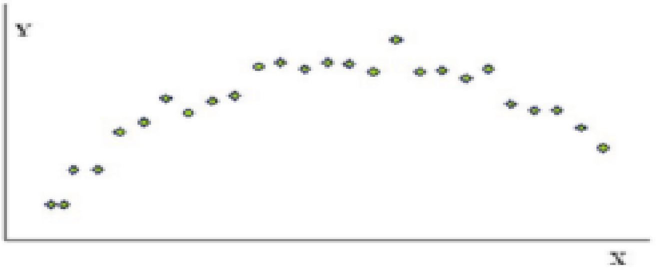
- Este diagrama permite decidir empiricamente:

- se um relacionamento linear entre as variáveis X e Y deve ser assumido
- se o grau de relacionamento linear entre as variáveis é forte ou fraco, conforme o modo como se situam os pontos em redor de uma reta imaginária que passa através do enxame de pontos.

Diagrama de dispersão

Diagramas de dispersão que sugerem uma regressão não linear entre as variáveis





Coeficiente de Correlação Linear

Designamos Coeficiente de Correlação Linear:

$$r = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sqrt{\left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right)\left(\sum_{i=1}^{n} y_i^2 - n\bar{y}^2\right)}}$$

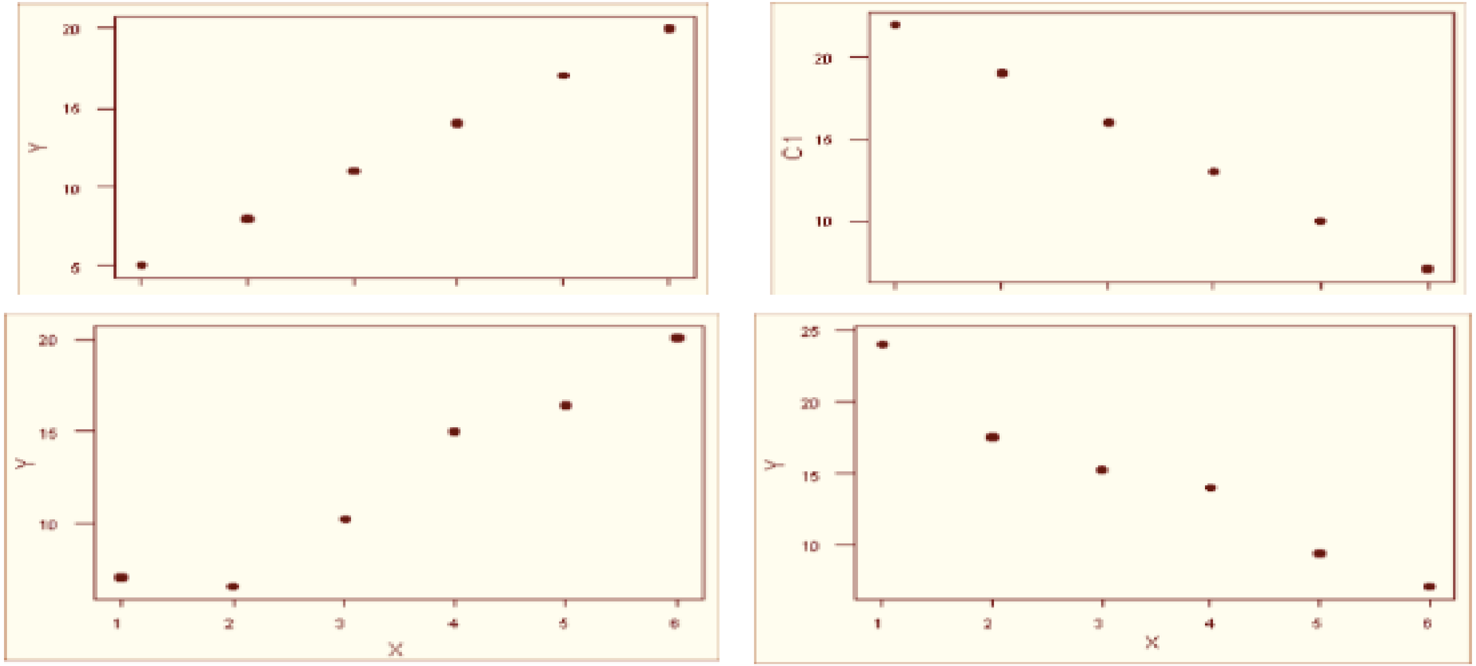
- Este coeficiente é uma medida do grau de dependência linear entre as duas variáveis, X e Y.

$$-1 \leq r \leq 1;$$

- $r = 1$: relação linear perfeita (e positiva) entre X e Y;
- $r = 0$: inexistência de relação linear entre X e Y;
- $r = -1$: relação linear perfeita (e negativa) entre X e Y;
- $r > 0$: relação linear positiva entre X e Y;
- $r < 0$: relação linear negativa entre X e Y.

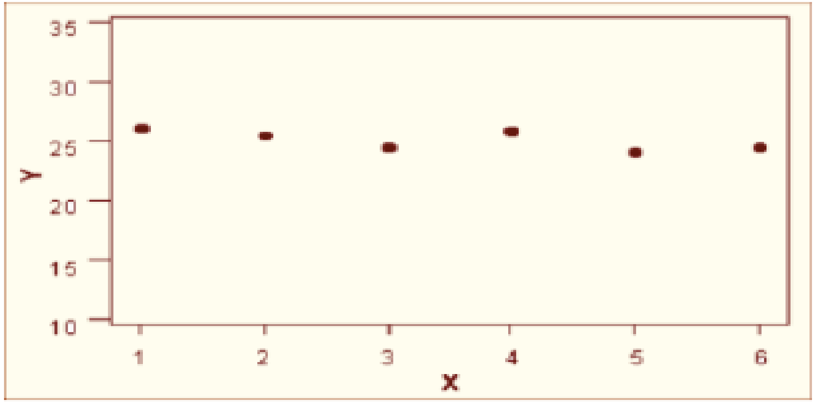
Coeficiente de Correlação Linear

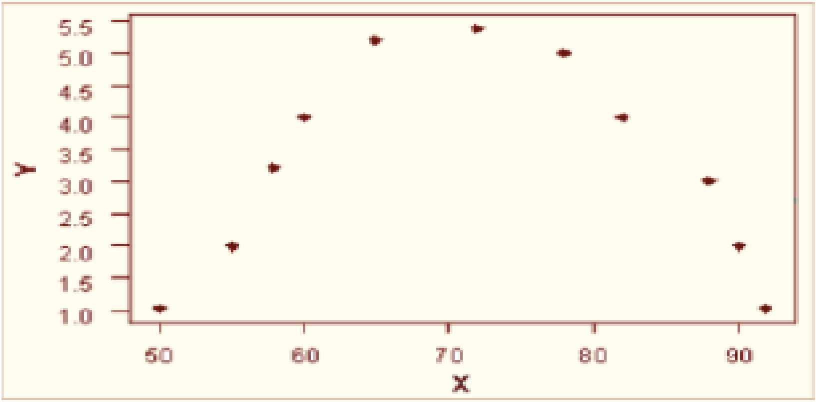
$r = 1$ $\qquad$ $r = -1$



$0 < r < 1$

Coeficiente de Correlação Linear





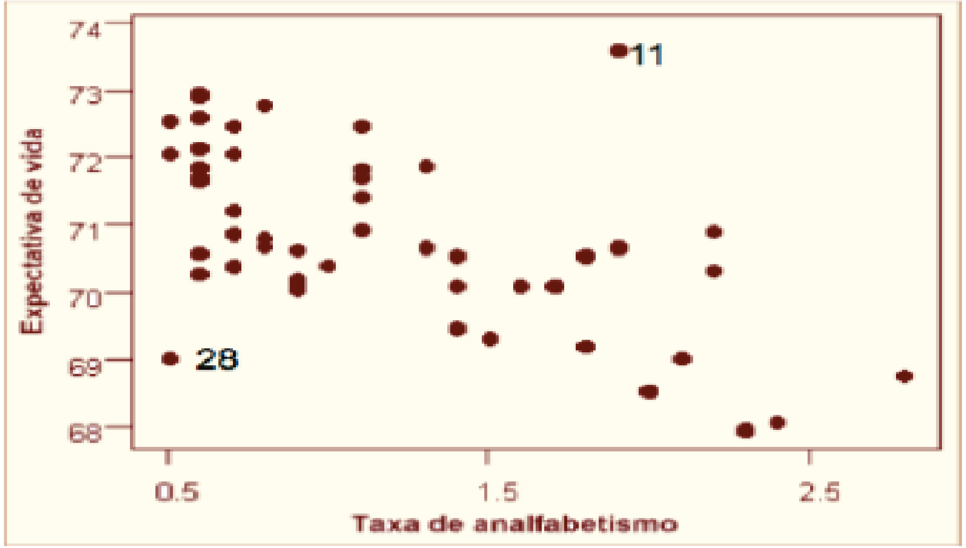
$$r = 0$$

Exemplo

- Considere as duas variáveis abaixo observadas em 50 estados norte-americanos.

Y: expectativa de vida

X: taxa de analfabetismo



Quanto maior é a taxa de analfabetismo, menor é a Expectativa de vida, e observamos ainda a existência de uma tendência linear entre as variáveis.

Exemplo

- Calcule o coeficiente de correlação entre X e Y, sabendo que:

$$\boxed{r = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sqrt{\left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right)\left(\sum_{i=1}^{n} y_i^2 - n\bar{y}^2\right)}}}$$

$$\bar{y} = 70,88; \quad \bar{x} = 1,17; \quad \sum_{i=1}^{n} x_i y_i = 4122,8$$

$$\sum_{i=1}^{n} x_i^2 - n\bar{x}^2 = 88,247; \quad \sum_{i=1}^{n} y_i^2 - n\bar{y}^2 = 18,173$$

$$r = \frac{4122,8 - 50 \times 1,17 \times 70,88}{\sqrt{88,247 \times 18,173}} = \frac{-23,68}{40,047} = -0,59$$

O Modelo de regressão linear simples (MRLS)

$$Y = E(Y|X = x) + \epsilon = \alpha + \beta x + \epsilon$$

- Y - variável explicada ou dependente (aleatória)
- X - variável explicativa ou independente medida sem erro (não aleatória)
- α - coeficiente de regressão, que representa o intercepto (parâmetro desconhecido do modelo -> a estimar)
- β - coeficiente de regressão, que representa o declive (inclinação) (parâmetro desconhecido do modelo -> a estimar)
- ϵ - erro aleatório ou estocástico, onde se procuram incluir todas as influências no comportamento da variável Y que não podem ser explicadas linearmente pelo comportamento da variável X;

O Modelo de regressão linear simples (MRLS)

Dadas n observações da variável X: $x_1, x_2, \ldots, x_n$, obtemos n v.a.'s $Y_1, Y_2, \ldots, Y_n$ satisfazendo a equação,

$$Y_i = E(Y|X = x_i) + \epsilon_i = \alpha + \beta x_i + \epsilon_i, \quad i = 1, 2, \ldots, n$$

Assume-se que as v.a.'s ϵ_i são v.a.'s independentes com média zero, $E(\epsilon_i|x) = 0$, e variância σ^2, $Var(\epsilon_i|x) = \sigma^2$.

Logo,

$$E(Y_i|X = x_i)) = \mu_{Y_i} = \alpha + \beta x_i \quad \text{e} \quad Var(Y_i|X = x_i) = \sigma^2$$

Recolhida uma amostra de n indivíduos, teremos n pares de valores (x_i, y_i), $i = 1, 2, \ldots, n$, que devem satisfazer o seguinte modelo,

$$y_i = \alpha + \beta x_i + \epsilon_i, \quad i = 1, 2, \ldots, n.$$

Temos n equações e $n + 2$ incógnitas $(\alpha, \beta, \epsilon_1, \ldots, \epsilon_n)$, por isso precisamos de introduzir um critério que permita encontrar α e β .

Método dos Mínimos Quadrados (MMQ)

- Encontrar os valores de α e β que minimizam a soma dos quadrados dos erros (ou desvios ou resíduos), dados por:

$$\epsilon_i = y_i - (\alpha + \beta x_i)$$

- Obtemos então, a quantidade de informação perdida pelo modelo ou soma dos quadrados dos resíduos

$$SQ(\alpha, \beta) = \sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} (y_i - \alpha - \beta x_i)^2$$

- Derivando em relação a α e β obtemos o sistema:

Método dos mínimos Quadrados (MMQ)

$$\begin{cases} \left.\dfrac{\partial SQ(\alpha,\beta)}{\partial \alpha}\right|_{\alpha=\hat{\alpha}} = 0 \\ \left.\dfrac{\partial SQ(\alpha,\beta)}{\partial \beta}\right|_{\beta=\hat{\beta}} = 0 \end{cases} \Leftrightarrow \begin{cases} \sum_{i=1}^{n}(y_i - \hat{\alpha} - \hat{\beta}x_i) = 0 \\ \sum_{i=1}^{n} x_i(y_i - \hat{\alpha} - \hat{\beta}x_i) = 0 \end{cases}$$

$$\Leftrightarrow \begin{cases} \sum_{i=1}^{n} y_i = n\hat{\alpha} + \hat{\beta}\sum_{i=1}^{n} x_i \\ \sum_{i=1}^{n} x_i y_i = \hat{\alpha}\sum_{i=1}^{n} x_i + \hat{\beta}\sum_{i=1}^{n} x_i^2 \end{cases} \Leftrightarrow \begin{cases} \hat{\alpha} = \bar{y} - \hat{\beta}\bar{x} \\ \hat{\beta} = \dfrac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2} \end{cases},$$

onde $\overline{x} = \dfrac{\sum_{i=1}^{n} x_i}{n}$ e $\overline{y} = \dfrac{\sum_{i=1}^{n} y_i}{n}$

Reta de Regressão Estimada

- Uma vez obtidas as estimativas, podemos escrever a equação estimada:

$$\hat{y} = \hat{\alpha} + \hat{\beta}x$$

- Definindo

$$S_{xy} = \sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y} \qquad S_{xx} = \sum_{i=1}^{n} x_i^2 - n\bar{x}^2 \qquad S_{yy} = \sum_{i=1}^{n} y_i^2 - n\bar{y}^2$$

- Obtemos:

$$\hat{\alpha} = \overline{y} - \hat{\beta}\overline{x} \quad \text{e} \quad \hat{\beta} = \frac{S_{xy}}{S_{xx}}$$

Interpretação das estimativas $\hat{\alpha}$ e $\hat{\beta}$

$$\hat{y} = \hat{\alpha} + \hat{\beta}x$$

$x = 0$: $\hat{y} = \hat{\alpha}$;

$x \rightarrow x + 1$: $\Delta\hat{y} = \hat{\alpha} + \hat{\beta}(x+1) - (\hat{\alpha} + \hat{\beta}x) = \hat{\beta}$

Logo, $\hat{\alpha}$ é o ponto onde a reta corta o eixo das ordenadas e pode ser interpretável ou não.

- $\hat{\beta}$ é o coeficiente angular, e representa o quanto varia a média de Y para um aumento de uma unidade da variável X.

Interpretação das estimativas $\hat{\alpha}$ e $\hat{\beta}$

- Tendo em conta a notação apresentada, o coeficiente de correlação simples pode ser escrito do seguinte modo

$$r = \frac{S_{xy}}{\sqrt{S_{xx}\,S_{yy}}}$$

Previsão

- Uma aplicação muito importante de um modelo de regressão é a previsão de novas ou futuras observações de $\boldsymbol{Y}, (\boldsymbol{Y}_f(\boldsymbol{x}))$ correspondente a um dado valor da variável explicativa X, x_f, então o estimador será

$$\hat{Y}_f = \hat{y}_f = \hat{\alpha} + \hat{\beta} x_f .$$

Exemplo 1

- Um psicólogo está investigando a existência de uma relação linear entre o tempo que um indivíduo leva a reagir a um certo estímulo visual (Y) e a respectiva idade (X), para indivíduos com idades compreendidas no intervalo [20,40]. Os resultados observados permitiram obter os dados abaixo:

$$n = 20 \quad \sum y_i = 2150 \quad \sum x_i = 600 \quad \sum x_i y_i = 65400$$

$$\overline{y} = 107,50 \quad \overline{x} = 30 \quad \sum x_i^2 = 19000$$

Exemplo 1

- Obtenha a equação do modelo ajustado, interprete as estimativas obtidas e estime o tempo médio de reação para um indivíduo de 25 anos, e para 45anos?

$$\hat{y} = \hat{\alpha} + \hat{\beta}x$$

Definindo

$$S_{xy} = \sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y} \qquad S_{xx} = \sum_{i=1}^{n} x_i^2 - n\bar{x}^2 \qquad S_{yy} = \sum_{i=1}^{n} y_i^2 - n\bar{y}^2$$

obtemos

$$\hat{\alpha} = \overline{y} - \hat{\beta}\overline{x} \quad \text{e} \quad \hat{\beta} = \frac{S_{xy}}{S_{xx}}$$

Exemplo 1

Resolução:

$S_{xy} = 65400 - 20 \times 30 \times 107,50 = 900$

$S_{xx} = 19000 - 20 \times 30^2 = 1000$

logo

$\hat{\beta} = \frac{900}{1000} = 0,9$

$\hat{\alpha} = 107,50 - 0,9 \times 30 = 80,50$

Equação do modelo ajustado: $\hat{y}_i = 80,50 + 0,9x_i,\ \ i = 1,2,\ldots 20$

Interpretação: $\hat{\alpha} = 80,50$ – tempo de reação para um recém-nascido (inadequação do modelo)
$\hat{\beta} = 0,9$ – por cada ano de envelhecimento das pessoas, o tempo médio de reação aumenta 0,9 unidades.

Previsão: $\hat{y}(25) = 80,50 + 0,9 \times 25 = 103$
$\hat{y}(45)$ – $45 \notin [20,40]$, logo não é possível determinar $\hat{y}(45)$.

Exemplo 2

- O gerente de uma cadeia de supermercados deseja desenvolver um modelo com a finalidade de estimar as vendas médias semanais (em milhares de dólares)
- Y - Vendas semanais; e
- X - Número de clientes.
- Estas variáveis foram observadas em 20 supermercados escolhidos aleatoriamente.

X	907	926	506	741	789	889	874	510	529	420
Y	11,20	11,05	6,84	9,21	9,42	10,08	9,45	6,73	7,24	6,12
X	679	872	924	607	452	729	794	844	1010	621
Y	7,63	9,43	9,46	7,64	6,92	8,95	9,33	10,23	11,77	7,41

Exemplo 2

- Considerando os dados:

$$
\begin{aligned}
n &= 20 \\
\sum_{i=1}^{n} x_i &= 907 + 926 + \ldots + 621 = 14.623; \quad \bar{x} = 731,15 \\
\sum_{i=1}^{n} y_i &= 11,20 + 11,05 + \ldots + 7,41 = 176,11; \quad \bar{y} = 8,8055 \\
\sum_{i=1}^{n} x_i^2 &= (907)^2 + (926)^2 + \ldots + (621)^2 = 11.306.209 \\
\sum_{i=1}^{n} y_i^2 &= (11,20)^2 + (11,05)^2 + \ldots + (7,41)^2 = 1.602,0971 \\
\sum_{i=1}^{n} x_i y_i &= (907)(11,20) + (11,05)(926) \ldots + (7,41)(621) = 134.127,90
\end{aligned}
$$

Exemplo 2

- Considerandos os dados:

$$
\begin{aligned}
S_{xx} &= \sum_{i=1}^{n} x_i^2 - n(\bar{x})^2 = 11.306.209 - 20(731,15)^2 = 614.603 \\
S_{xy} &= \sum_{i=1}^{n} x_i y_i - n(\bar{x})(\bar{y}) = 134.127,90 - 20(8,8055)(731,15) = 5.365,08 \\
S_{yy} &= \sum_{i=1}^{n} y_i^2 - n(\bar{y})^2 = 1.609,0971 - 20(8,8055) = 51,3605.
\end{aligned}
$$

Exemplo 2

- As estimativas dos parâmetros do MRLS são:

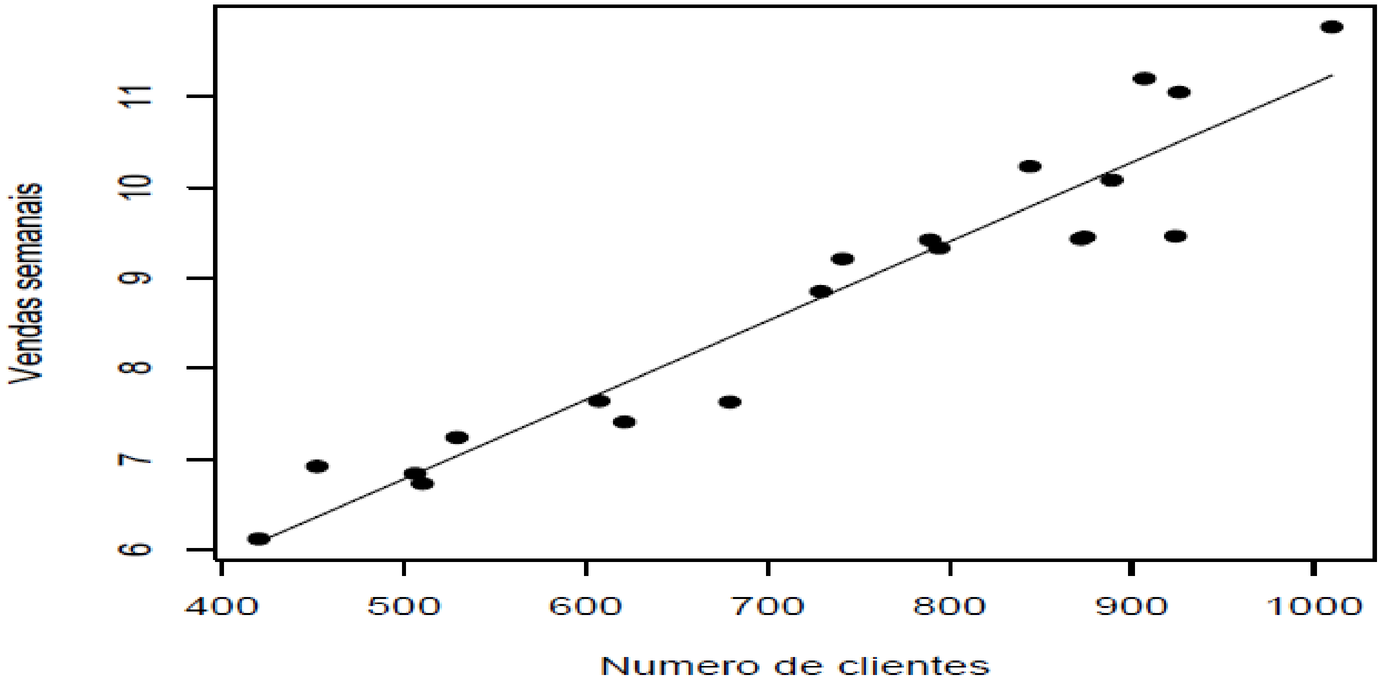
$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \frac{5.365,08}{614.603} = 0,00873;$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = 8,8055 - (0,00873)(731,15) = 2,423$$

- Portanto, a linha de regressão ajustada ou estimada para esses dados são:

$$\hat{y} = 2,423 + 0,00873x.$$

Exemplo 2



Vendas semanais
6
7
8
9
10
11
400
500
600
700
800
900
1000
Numero de clientes

Exemplo 2

- Suponha que tem-se interesse em prever as vendas semanais para um supermercado com 600 clientes.
- No modelo de regressão ajustado basta substituir X = 600, isto é,

$$\hat{y} = 2,423 + (0,00873)(600) = 7,661.$$

- A venda semanal de 7,661 mil dólares pode ser interpretada com uma estimação da venda média semanal verdadeira dos supermercados com X = 600 clientes,

Coeficiente de Determinação (r^2)

- Uma das formas de avaliar a qualidade do ajuste do modelo é através do coeficiente de determinação. Basicamente, este coeficiente indica quanto o modelo foi capaz de explicar os dados coletados. O coeficiente de determinação é dado pela expressão

$$R^2 = \frac{SQR}{SQT} = 1 - \frac{SQE}{SQT} = \frac{\widehat{\beta}_1 \sum_{i=1}^{n}(x_i - \bar{x})Y_i}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2},$$

Coeficiente de Determinação (r^2)

- O quociente entre SQReg e SQTot dá-nos uma medida da proporção da variação total que é explicada pelo MRLS.

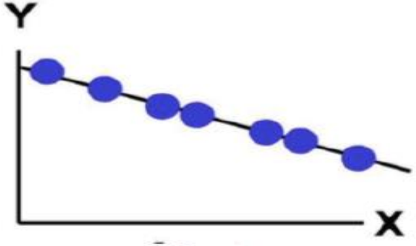
- Este coeficiente pode ser utilizado como uma medida da qualidade do ajustamento, ou como medida da confiança depositada na equação de regressão como instrumento de previsão, e representa a porcentagem da variação total que é explicada pelo MRLS. Note-se que o ajustamento será tanto melhor quanto mais **pequeno for SQRes (e portanto, maior for SQReg) relativamente a SQTot.**

Coeficiente de Determinação (r^2)

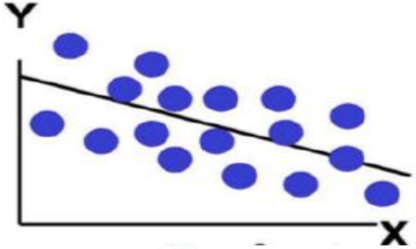
$0 \leq r^2 \leq 1$;

$r^2 \approx 0$ – modelo linear muito pouco adequado;

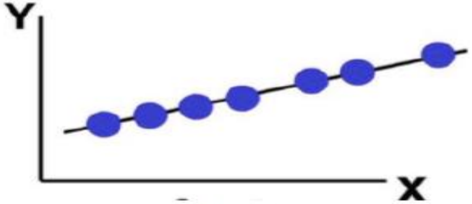
$r^2 \approx 1$ – modelo linear bastante adequado.



$r^2 = 1$



$0 < r^2 < 1$



$r^2 = 1$

Exempo (r^2)

- Calcule e interprete o coeficiente de determinação no Exemplo 1, sabendo que

$$\sum y_i^2 = 232498.$$

$$r^2 = \frac{\hat{\beta} S_{xy}}{S_{yy}} = \frac{0,9 \times 900}{232498 - 20 \times 107,50^2} = \frac{810}{1373} = 0,59 \rightarrow 59\%$$

- **Interpretação:** 59% da variação no tempo de reação está relacionada linearmente com a idade do indivíduo, sendo os restantes 41% da variação resultantes de outros fatores não considerados (sexo, acuidade visual,...).

Resíduos

- A diferença entre o valor observado Y_i e o correspondente valor ajustado $\widehat{Y}_i$, dado pela expressão abaixo, é chamada de resíduo:

$$e_i = y_i - \hat{y}_i = y_i - \hat{\alpha} - \hat{\beta} x_i$$

- Se os resíduos forem pequenos temos uma indicação de que o modelo está produzindo bons resultados.

Estimador da Variância Residual

- Para obtermos um estimador não enviesado de σ^2, analisamos a dispersão em torno da reta de regressão - **Variação não explicada/Residual**

$$SQRes = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y})^2 \quad \text{(Soma dos quadrados dos resíduos).}$$

Como $E(SQRes) = (n-2)\sigma^2$, então um estimador não enviesado de σ^2 é

$$\hat{\sigma}^2 = QMRes = \frac{SQRes}{n-2}$$

Estimador da Variância Residual

- Definindo a **Variação Total**, como sendo a dispersão em torno de $\bar{Y}$

$$SQTot = \sum_{i=1}^{n}(y_i - \overline{y})^2 = S_{yy} \quad \text{(Soma de quadrados totais)}$$

- **Prova-se que:**

$$SQRes = S_{yy} - \hat{\beta} S_{xy}$$

Exemplo σ^2 Residual

- **Calcular a σ^2 Residual para o Exemplo 2**

$$
\begin{aligned}
\hat{\sigma^2} &= \frac{SQR}{n-2} = \frac{S_{yy} - \hat{\beta}_1 S_{xy}}{n-2} \\
&= \frac{51,3605 - (0,00873)(5.365,08)}{20-2} = 0,2513.
\end{aligned}
$$