3 - ANÁLISE BIDIMENSIONAL

É comum haver interesse em saber se duas variáveis quaisquer estão relacionadas, e o quanto estão relacionadas, seja na vida prática, seja em trabalhos de pesquisa, por exemplo:
- se o sexo dos funcionários de uma empresa está relacionado com a função exercida;
- o quanto o a temperatura ambiente em uma região influencia as vendas de refrigerante;
- se o nível de escolaridade de um grupo de empreendedores está relacionado com o grau de sucesso por eles alcançado.

Muitas vezes queremos verificar se há uma relação de causa e efeito entre as duas variáveis (se as variáveis são dependentes ou não), se é possível estudar uma das variáveis através da outra (que é mais fácil de medir)- prever os valores de uma através dos valores da outra, ou calcular uma medida de correlação ou de dependência entre as variáveis.

A Análise Bidimensional[1] propõe-se a tentar responder as perguntas do parágrafo anterior. As duas variáveis abordadas podem ser qualitativas ou quantitativas, e para cada tipo haverá técnicas apropriadas.

Para variáveis qualitativas vamos estudar: tabelas de contingência (já vistas na seção 2.2), estatística Qui-Quadrado e o Coeficiente de Contingência Modificado[2]. Para variáveis quantitativas vamos abordar: diagramas de dispersão, análise de correlação, análise de regressão linear simples, coeficiente de determinação e análise de resíduos. As próximas seções tratarão de cada tópico.

3.1 - Análise Bidimensional de Variáveis Qualitativas

A análise bidimensional de variáveis qualitativas foi vista na seção 2.2, mas seria interessante relembrar alguns pontos.

Variáveis Qualitativas são as variáveis cujas realizações são atributos, categorias. Como exemplo de variáveis qualitativas tem-se: sexo de uma pessoa (duas categorias, masculino e feminino), grau de instrução (analfabeto, primeiro grau incompleto, etc.), opinião sobre um assunto (favorável, desfavorável, indiferente), etc.

Em estudos sobre variáveis qualitativas é extremamente comum registrar as frequências de ocorrência de cada valor que as variáveis podem assumir, e quando há duas variáveis envolvidas é comum registrar-se a frequência de ocorrência dos cruzamentos entre valores: por exemplo, quantas pessoas do sexo masculino são favoráveis a uma certa proposta de lei, quantas são desfavoráveis, quantas pessoas do sexo feminino são favoráveis, etc. E, para facilitar a análise dos resultados estes resultados costumam ser dispostos em uma Tabela de Contingências (fazendo uma dupla classificação). A Tabela de Contingências relaciona os possíveis valores de uma variável qualitativa com os possíveis valores da outra, registrando quantas ocorrências foram verificadas de cada cruzamento.

[1] Se mais de duas variáveis estiverem envolvidas será necessário empregar técnicas de análise multidimensional, ou ANÁLISE MULTIVARIADA.
[2] No Capítulo 6 iremos estudar o teste de independência do Qui-Quadrado, uma outra forma de avaliar a associação entre duas variáveis qualitativas.

Exemplo 3.1 - Vamos analisar novamente a tabela de contingências para as variáveis Sexo e Função dos funcionários da empresa Escolástica Ltda., construída no Exemplo 2.3.

Sexo	Função			
	Escritório	Serviços gerais	Gerência	Total
Masculino	157	27	74	258
Feminino	206	0	10	216
Total	363	27	84	474

Fonte: hipotética

As conclusões são as mesmas a que chegamos no Exemplo 2.3. Podemos apresentar os percentuais calculados em relação aos totais das colunas:

	Função			
Sexo	Escritório	Serviços gerais	Gerência	Total
Masculino	43,25%	100%	88,10%	54%
Feminino	56,75%	0%	11,90%	46%
Total	100%	100%	100%	100%

Fonte: hipotética

Seria interessante saber se as duas variáveis são estatisticamente dependentes, e o quão forte é esta associação. Repare que os percentuais de homens e mulheres em cada função são diferentes dos percentuais marginais (de homens e mulheres no total de funcionários), sendo que em duas funções as diferenças são bem grandes.

A tabela de contingências também é chamada de distribuição conjunta das duas variáveis. Permite descrever o grau de associação existente entre as duas variáveis: é possível avaliar a "força" do relacionamento, e caso haja uma associação forte pode-se prever os valores de uma variável através dos da outra. Se as variáveis forem ***independentes*** (ou seja, a associação entre elas for fraca), as frequências na tabela de contingências devem distribuir-se de forma a seguir o padrão dos ***totais marginais***. Se, porém, houver uma associação entre as variáveis, elas forem ***dependentes***, as frequências deverão seguir algum padrão ***diferente*** daquele apresentado pelos totais marginais.

Precisamos de uma estatística que relacione as frequências OBSERVADAS na tabela de contingências com as frequências ESPERADAS se as duas variáveis fossem independentes (se as frequências nos cruzamentos dos valores das variáveis seguissem os padrões dos totais marginais). E quais serão os valores das frequências esperadas?

Exemplo 3.2 - Calcule as frequências esperadas sob a condição de independência entre Sexo e Função para a tabela de contingências do Exemplo 3.1.

Se as variáveis são independentes as frequências de homens e mulheres em cada função devem ter a mesma proporção que homens e mulheres têm no total de funcionários. Lembrando que há 54% de homens e 46% de mulheres, esperamos que esses percentuais mantenham-se em cada função, se as variáveis são independentes.

- Em Escritório, há 363 pessoas nesta função, sob a condição de independência deveriam haver:
Homens => 54% de 363 = 197,58 Mulheres => 46% de 363 = 165,42
- Em Serviços Gerais, há 27 pessoas, sob a condição de independência deveriam haver:
Homens => 54% de 27 = 14,70 Mulheres => 46% de 27 = 12,30
- Em Gerência, há 84 pessoas, sob a condição de independência deveriam haver:
Homens => 54% de 84 = 45,72 Mulheres => 46% de 84 = 38,28

Um rápido exame da tabela do Exemplo 3.1 mostra que as frequências observadas estão razoavelmente distantes das esperadas sob a condição de independência. Há indícios de que as duas variáveis estão relacionadas.

Podemos calcular as frequências esperadas para todas as células da tabela de contingências diretamente, utilizando a seguinte fórmula:

$$E_{ij} = \frac{\text{total da linha } i \times \text{total da coluna } j}{\text{total geral}}$$

Onde E_{ij} é a frequência esperada, sob a condição de independência entre as variáveis, em uma célula qualquer da tabela de contingências. As frequências esperadas são necessárias para que possamos compará-las com as observadas, sendo essa comparação materializada em uma estatística, chamada de Qui-Quadrado: $\chi 2$. A expressão está descrita abaixo:

$$\chi^2 = \sum_{i=1}^{L}\sum_{j=1}^{C}\left[\frac{(O_{ij} - E_{ij})^2}{E_{ij}}\right]$$

Onde L é o número total de linhas da tabela de contingências (número de valores que uma das variáveis pode assumir), C é o número total de colunas da tabela (número de valores que a outra variável pode assumir), e O_{ij} é a frequência observada em uma célula qualquer da tabela de contingências. Então, para cada célula da tabela de contingências calcula-se a diferença entre a frequência observada e a esperada. Para evitar que as diferenças positivas anulem as negativas as diferenças são elevadas ao quadrado. E para evitar que uma diferença grande em termos absolutos, mas pequena em termos relativos, "inflacione" a estatística, ou que uma diferença pequena em termos absolutos, mas grande em termos relativos, tenha sua influência reduzida, divide-se o quadrado da diferença pela frequência esperada. Somam-se os valores de todas as células e obtêm-se o valor da estatística.

Exemplo 3.3 - Calcule a estatística Qui-Quadrado para a tabela de contingências do Exemplo 3.1.

Sexo	Função			
	Escritório	Serviços gerais	Gerência	Total
Masculino	157	27	74	258
Feminino	206	0	10	216
Total	363	27	84	474

Fonte: hipotética

Calculando as frequências esperadas de acordo com a fórmula vista anteriormente:

Masculino - Escritório $E = (258 \times 363)/474 = 197{,}58$

Masculino - Serviços Gerais $E = (258 \times 27)/474 = 14{,}70$

Masculino - Gerência $E = (258 \times 84)/474 = 45{,}72$

Feminino - Escritório $E = (216 \times 363)/474 = 165{,}42$

Feminino - Serviços Gerais $E = (216 \times 27)/474 = 12{,}30$

Feminino - Gerência $E = (216 \times 84)/474 = 38{,}28$

Agora podemos calcular as diferenças entre as frequências e as demais operações, que serão mostradas nas tabelas a seguir.

O - E	Função		
Sexo	Escritório	Serviços gerais	Gerência
Masculino	157 - 197,58	27 - 14,70	74 - 45,72
Feminino	206 - 165,42	0 - 12,30	10 - 38,28

$(O-E)^2$	Função		
Sexo	Escritório	Serviços gerais	Gerência
Masculino	1646,921	151,383	799,672
Feminino	1646,921	151,383	799,672

Finalmente:

$(O-E)^2/E$	Função		
Sexo	Escritório	Serviços gerais	Gerência
Masculino	8,336	10,301	17,490
Feminino	9,956	12,304	20,891

Agora podemos somar os valores:

$$\chi 2 = 8{,}336 + 10{,}301 + 17{,}490 + 9{,}956 + 12{,}304 + 20{,}891 = 79{,}227$$

Quanto maior for o valor de $\chi 2$ maior será o grau de associação entre as variáveis. No Capítulo 9 aprenderemos a usar esta estatística em um teste sobre a independência entre as variáveis. Neste Capítulo vamos utilizar outra estatística, a partir do $\chi 2$ para mensurar a força do relacionamento entre as variáveis: o Coeficiente de Contingência Modificado.

3.1.1 - Coeficiente de Contingência Modificado

O Coeficiente de Contingência Modificado permite quantificar a associação (grau de dependência) entre duas variáveis QUALITATIVAS, a partir da estatística $\chi 2$ vista anteriormente. Sua equação:

$$C^* = \sqrt{\frac{\chi^2}{\chi^2 + N}} \times \sqrt{\frac{k}{k-1}}$$

Onde:
- $\chi 2$ é a estatística Qui-Quadrado, calculada a partir das frequências observadas e esperadas (sob a condição de independência) a partir da tabela de contingências.
- N é o número total de observações da tabela de contingências.
- k é o menor número entre o número de linhas e colunas da tabela de contingências.

O Coeficiente de Contingência Modificado varia de zero (completa independência) até 1 (associação perfeita).

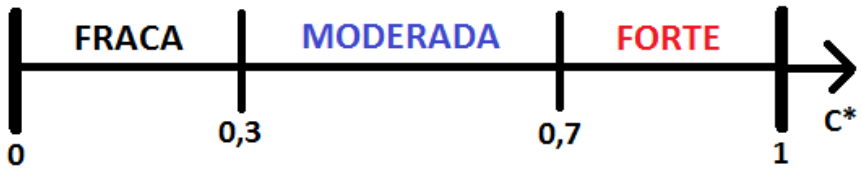


Figura 1 - Variação da associação entre as variáveis pelo Coeficiente de Contingência Modificado C*

Valores de C* entre 0 e 0,29 indicam associação fraca entre as variáveis, as frequências dos valores de uma das variáveis aparentemente não são influenciadas pelos valores da outra. Valores de C* entre 0,3 e 0,69 indicam uma associação moderada. Valores acima de C* acima de 0,7 indicam uma associação forte entre as variáveis, há evidência que as frequências dos valores de uma das variáveis foi influenciada pelos valores da outra. CUIDADO, porém, com as generalizações, associação estatística não significa relação de causa e efeito!

Exemplo 3.4 - Calcule o Coeficiente de Contingência Modificado para os dados do Exemplo 3.3.
O valor de $\chi 2$ foi calculado no Exemplo 3.3, a variável Sexo pode assumir 2 valores, e Função pode assumir 3. O total de observações é igual a 474.
Então: $\chi 2 = 79{,}227 \quad N = 474 \quad k = 2$ *(porque é o menor valor entre 2 e 3).*

$$C^* = \sqrt{\frac{\chi^2}{\chi^2 + N}} \times \sqrt{\frac{k}{k-1}} = \sqrt{\frac{79{,}227}{79{,}227 + 474}} \times \sqrt{\frac{2}{2-1}} \cong 0{,}54$$

Então a associação pode ser considerada moderada. O resultado é coerente com a tabela de contingências, pois há grandes diferenças entre as frequências esperadas e observadas.

3.2 - Análise Bidimensional de Variáveis Quantitativas

Muitas vezes também estamos interessados em avaliar o relacionamento entre variáveis QUANTITATIVAS, sejam elas discretas ou contínuas. Basicamente dois tipos de análise podem ser realizados: Análise de Correlação e Análise de Regressão.

Na análise de correlação e regressão há interesse em, a partir de dados de uma amostra aleatória, verificar SE e COMO duas ou mais variáveis quantitativas[3] relacionam-se entre si em uma população.

A Análise de Correlação fornece um número que resume o relacionamento entre as variáveis, indicando a ***força*** e a ***direção*** do relacionamento.

A Análise de Regressão fornece uma equação matemática que descreve a ***natureza*** do relacionamento entre as duas variáveis, permitindo inclusive que sejam feitas previsões dos valores de uma delas em função dos valores das outras.

Quando há apenas duas variáveis envolvidas a Análise de Regressão é chamada Simples. Quando há mais de duas variáveis temos a Análise de Regressão Múltipla.

Uma das suposições básicas da Análise de Correlação e Regressão é que há alguma teoria (ou evidência empírica) que permita levantar hipóteses sobre a relação de dependência entre as variáveis, ou seja, que permita identificar variáveis dependente e independente(s)[4]. A teoria deve mostrar se esperamos associação positiva ou negativa e em que grau. Por exemplo, ao avaliarmos o relacionamento entre renda mensal em reais e área em m^2 da residência de uma família, esperamos um relacionamento positivo entre ambas: para maior renda (independente) esperamos maior área (dependente).

Uma ou mais das variáveis são chamadas de ***Independente(s)***: podem ser uma ou mais variáveis que o pesquisador manipulou para observar o efeito em outra, ou mesmo variáveis cuja medição possa ser feita de maneira mais fácil ou precisa, sendo então suposta sem erro.

Há uma outra variável, chamada de ***Dependente***, seus valores seriam resultado da variação dos valores das variáveis Independentes[5]. Esta denominação costuma levar a má interpretação do significado da "correlação" entre variáveis: *se há correlação entre variáveis significa que os seus valores variam em uma mesma direção, ou em direções opostas, com uma certa "força",* ***não***

[3] Há possibilidade de avaliar o relacionamento entre duas variáveis qualitativas nominais (através do Coeficiente de Contingência Modificado, que foi visto anteriormente) e entre duas variáveis qualitativas ordinais (através dos coeficientes de correlação por postos, que não serão abordados nesta disciplina).
[4] Na Análise de Regressão Múltipla podem haver várias variáveis independentes mas apenas UMA dependente.
[5] Veja as definições de variáveis na seção 2.1.

significando necessariamente que uma variável depende das outras. Para tal conclusão seria necessário a existência de evidências "não estatísticas" dessa dependência, ou que os valores fossem o resultado de um experimento estatístico (adequadamente planejado e executado) em que todas as outras causas da variação tivessem sido eliminadas.

Para que seja possível realizar uma Análise de Correlação e/ou Regressão os dados devem provir de observações **emparelhadas** e em condições semelhantes. Se estamos avaliando a correlação existente entre a altura e o peso de um determinado grupo de crianças, por exemplo, o peso de uma determinada criança deve ser medido e registrado no mesmo instante em que é medida e registrada a sua altura. Renda e área da residência da mesma família, no mesmo momento. Se houver mais de duas variáveis todas devem ser medidas no mesmo instante.

Outro aspecto às vezes negligenciado é a **quantidade suficiente** de dados. Se apenas alguns poucos dados foram coletados podemos chegar a algumas conclusões errôneas:
- podemos descartar a correlação entre as variáveis, embora ela realmente exista, porque os dados foram insuficientes para mostrá-la;
- podemos concluir que há correlação, que na realidade não é significativa, porque os dados mostraram apenas uma pequena parte do conjunto total, onde, talvez por acaso, a correlação exista.

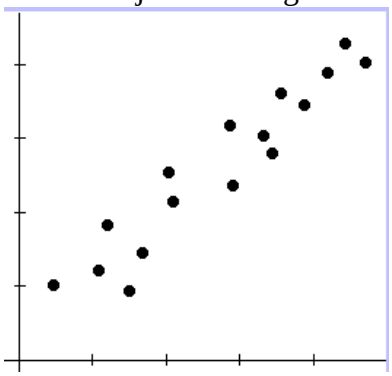
Por razões didáticas vamos limitar nosso estudo ao relacionamento entre duas variáveis apenas, e aos casos de relacionamento linear (em que o relacionamento pode ser descrito por uma equação de reta[6]). Se estamos trabalhando com apenas duas variáveis nosso primeiro passo é construir um gráfico que mostre o relacionamento entre as variáveis, um diagrama de dispersão.

3.2.1 - Diagrama de Dispersão

Se estamos analisando duas variáveis quantitativas, cujas observações constituem pares ordenados, chamando estas variáveis de **X** (independente) e **Y** (dependente), podemos plotar o conjunto de pares ordenados (x,y) em um diagrama cartesiano, que é chamado de ***Diagrama de Dispersão***.

Através do diagrama de dispersão é possível ter uma idéia inicial de como as variáveis estão relacionadas: a direção da correlação (isto é, o que ocorre com os valores de **Y** quando os valores de **X** aumentam, eles aumentam também ou diminuem), a força da correlação (em que "taxa" os valores de **Y** aumentam ou diminuem em função de **X**) e a natureza da correlação (se é possível ajustar uma reta, parábola, exponencial, etc., aos pontos).

Vejamos os seguintes diagramas de dispersão:

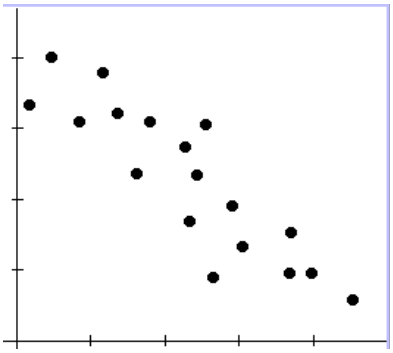


No diagrama ao lado percebemos dois aspectos básicos:
- à medida que a variável X aumenta, os valores de Y tendem a aumentar também.
- seria perfeitamente possível ajustar uma reta *crescente* que passasse por entre os pontos (obviamente a reta não poderia passar por todos eles).

Concluímos então que há correlação **linear** (porque é possível ajustar uma reta aos dados) **positiva** (porque as duas variáveis aumentam seus valores conjuntamente).

Figura 2 - Diagrama de dispersão 1º caso

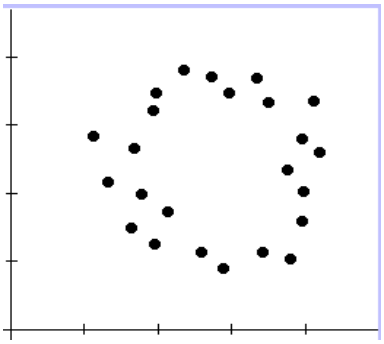
[6] Ou linearizável, que através de transformações apropriadas transforme-se em uma reta.



No diagrama ao lado percebemos dois aspectos básicos:
- à medida que a variável X aumenta, os valores de Y tendem a diminuir.
- seria perfeitamente possível ajustar uma reta *decrescente* que passasse por entre os pontos.

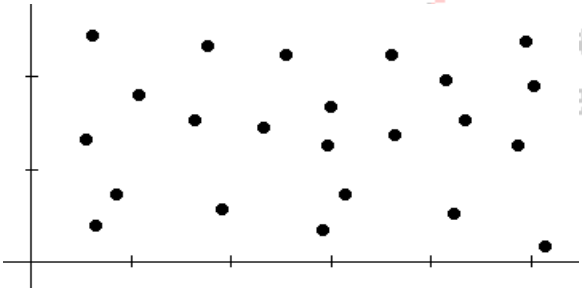
Concluímos então que há correlação **linear** (porque é possível ajustar uma reta aos dados) **negativa** (porque quando uma das variáveis aumenta seus valores e a outra diminui).

Figura 3 - Diagrama de dispersão 2º caso



No caso do diagrama ao lado é óbvio que há alguma espécie de correlação entre as variáveis: os pontos apresentam claramente um padrão, semelhante a um círculo. Contudo, não se trata de uma relação linear, pois seria totalmente inadequado ajustar uma reta aos dados (os resíduos seriam muito grandes). Assim, há correlação, mas **não é linear**.

Figura 4 - Diagrama de dispersão 3º caso



No caso do diagrama ao lado é óbvio temos uma situação totalmente diversa dos casos anteriores. NÃO HÁ padrão nos pontos, linear ou não linear, os pontos parecem distribuir-se de forma aleatória. Então, conclui-se que NÃO HÁ CORRELAÇÃO entre as duas variáveis.

Figura 5 - Diagrama de dispersão 4º caso

3.2.2 - Coeficiente de Correlação Linear de Pearson

Através do diagrama de dispersão é possível identificar se há correlação linear, e se a correlação linear é positiva ou negativa. Quanto mais o diagrama de dispersão aproximar-se de uma reta mais forte será a correlação linear.

É interessante notar que alguns erroneamente confundem "inexistência de correlação linear" com inexistência de correlação entre as duas variáveis. Duas variáveis podem apresentar uma forte correlação não-linear, conforme visto na seção anterior.

Se após observar o diagrama de dispersão decidir-se que é razoável considerar que as variáveis possuem um relacionamento linear é possível mensurar a direção e a força desse relacionamento através de um *coeficiente de correlação*: **o coeficiente de correlação linear de Pearson**. Este coeficiente é chamado de ρ quando são usados dados da população, e de **r** quando usados dados de uma amostra (mais comum).

Trata-se de um coeficiente adimensional, amostral, que pode ser expresso por:

$$r = \frac{Cov(X,Y)}{s_X \times s_Y} = \frac{\dfrac{\sum_{i=1}^{n}(x_i - \bar{x}) \times (y_i - \bar{y})}{n-1}}{s_X \times s_Y} \quad (1)$$

O numerador da expressão (1) é chamado de Covariância de **X** e **Y**, que permite mensurar o relacionamento entre as variáveis. A Covariância é dividida pelos desvios padrões de **X** e **Y** para que seja eliminado o efeito que uma variável com maiores valores numéricos causaria no resultado.

A covariância permite mensurar o relacionamento entre X e Y:
- quando os valores de X e Y são ambos grandes ou ambos pequenos (as distâncias em relação às médias têm o mesmo sinal) a covariância será grande e positiva.
- quando o valor de X é alto e o de Y é baixo (ou vice-versa) a covariância será grande e negativa.
dividindo-a por n-1 o seu valor não será mais afetado pelo tamanho da amostra.

Apesar de válida, a expressão (1) costuma levar a resultados que apresentam substanciais erros de arredondamento. A forma do coeficiente de correlação linear de Pearson mais utilizada (inclusive em calculadoras, programas estatísticos e planilhas eletrônicas) é:

$$r = \frac{n \times \sum_{i=1}^{n}(x_i \times y_i) - \sum_{i=1}^{n} x_i \times \sum_{i=1}^{n} y_i}{\sqrt{\left[n \times \sum_{i=1}^{n}(x_i^2) - \left(\sum_{i=1}^{n} x_i\right)^2\right]} \times \sqrt{\left[n \times \sum_{i=1}^{n}(y_i^2) - \left(\sum_{i=1}^{n} y_i\right)^2\right]}} \quad (2)$$

Para fazer os cálculos é preciso calcular a soma dos valores de **X**, a soma dos valores de **Y**, a soma dos valores do produto **XY**, a soma dos quadrados dos valores de **X**, a soma dos quadrados dos valores de **Y** e o número de valores da amostra (**n**).

O coeficiente de correlação linear de Pearson pode variar de -1 a +1 (passando por zero), e é adimensional[7]: se r = -1 significa que há uma correlação linear negativa perfeita entre as variáveis; se r = +1 significa que há uma correlação linear positiva perfeita entre as variáveis; e se r = 0 significa que não há correlação linear entre as variáveis. Admite-se que se |r| > 0,7 a correlação linear pode ser considerada forte.

Novamente, um alto coeficiente de correlação linear de Pearson (próximo a +1 ou a -1) não significa uma relação de causa e efeito entre as variáveis, apenas que as duas variáveis apresentam aquela tendência de variação conjunta.

Exemplo 3.5 - Estamos avaliando as médias de 15 estudantes no ensino médio, relacionando-as com os índices dos mesmos estudantes no seus cursos universitários. As médias no ensino médio podem variar de 0 a 100, e os índices na universidade de 0 a 4. Construa um diagrama de dispersão e calcule o coeficiente de correlação linear de Pearson para os dados a seguir. Interprete os resultados encontrados.

[7] Sem unidade.

Média no ensino médio	Índice na Universidade
80,0	1,0
82,0	1,0
84,0	2,1
85,0	1,4
87,0	2,1
88,0	1,7
88,0	2,0
89,0	3,5
90,0	3,1
91,0	2,4
91,0	2,7
92,0	3,0
94,0	3,9
96,0	3,6
98,0	4,0

O primeiro passo é definir qual variável é independente (X) e qual é a dependente (Y). Quem pode ter influenciado quem? É razoável imaginar que a média no ensino médio dos estudantes tenha influenciado de algum modo o índice por eles obtidos na universidade, simplesmente pelo fato de que é preciso cursar o ensino médio ***antes*** *da universidade. Assim sendo, X será a média no ensino médio (variável* ***independente****) e Y será o índice na universidade (variável* ***dependente****).*

Como será o relacionamento entre estas variáveis? Novamente, o bom senso indica que a valores altos de médias no ensino médio devam corresponder índices altos na universidade: espera-se uma correlação positiva.

Construindo o diagrama de dispersão (há várias planilhas eletrônicas e programas estatísticos que podem fazer isso) obtemos:

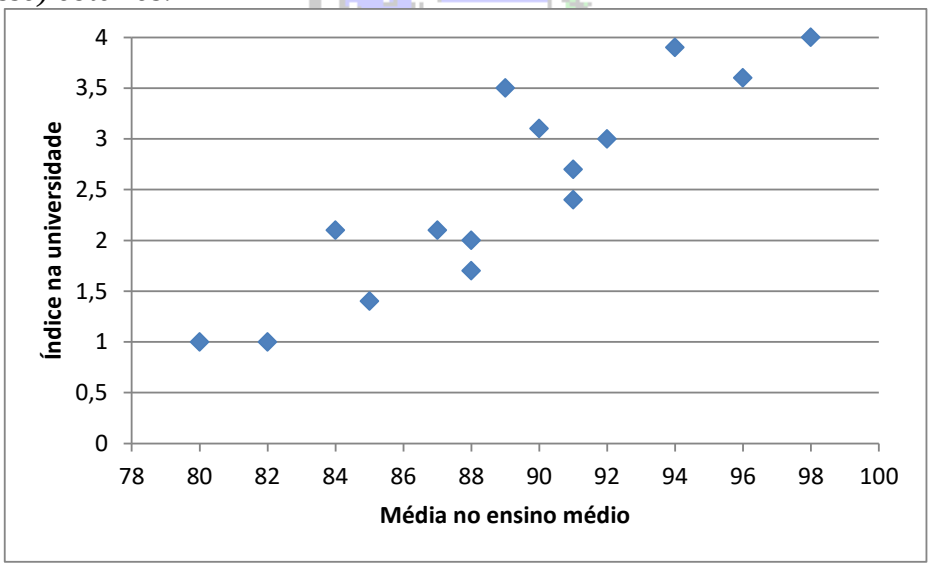


Figura 6 - Diagrama de dispersão: médias no ensino médio e índices na universidade

Observando o diagrama da Figura 6 *consegue-se vislumbrar que há uma correlação positiva entre as duas variáveis: de uma maneira geral, quanto maior o valor da média no ensino médio maior o índice na universidade. Além disso, é possível pensar em ajustar uma reta aos dados, que passasse por entre os pontos, e tal reta seria crescente (pois a correlação é positiva). Então, por ser possível ajustar uma reta aos dados, e os valores das variáveis caminham na mesma direção, há uma correlação* ***linear positiva*** *entre média no ensino médio e índice na universidade, ao menos para este conjunto de dados: ou seja, para valores MAIORES de média no ensino médio esperam-se valores também MAIORES de índice na universidade, e vice-versa, permitindo usar uma equação de reta para prever o índice na universidade de um aluno a partir da sua média no ensino médio.*

A correlação linear é forte? Quanto mais os pontos estiverem próximos da reta hipotética ajustada aos dados mais forte será a correlação. No diagrama da Figura 6 *os pontos estão próximos uns dos outros, estariam a pouca distância de uma reta que passasse entre eles. Conclui-se, então, que a correlação linear deve ser forte, o que resultará em um coeficiente de correlação linear de Pearson próximo de 1. Para calcular o coeficiente é preciso obter alguns somatórios.*

Média no ensino médio X	Índice na Universidade Y	X^2	Y^2	X×Y
80,0	1,0	6400	1,0	80,0
82,0	1,0	6724	1,0	82,0
84,0	2,1	7056	4,41	176,4
85,0	1,4	7225	1,96	119,0
87,0	2,1	7569	4,41	182,7
88,0	1,7	7744	2,89	149,6
88,0	2,0	7744	4,0	176,0
89,0	3,5	7921	12,25	311,5
90,0	3,1	8100	9,61	279,0
91,0	2,4	8281	5,76	218,4
91,0	2,7	8281	7,29	245,7
92,0	3,0	8464	9,0	276,0
94,0	3,9	8836	15,21	366,6
96,0	3,6	9216	12,96	345,6
98,0	4,0	9604	16,0	392,0

Sabe-se que n = 15 (há 15 alunos).

$$\sum_{i=1}^{15} x_i = 1335,0 \qquad \sum_{i=1}^{15} y_i = 37,5 \qquad \sum_{i=1}^{15} (x_i^2) = 119165,0 \qquad \sum_{i=1}^{15} (y_i^2) = 107,8 \qquad \sum_{i=1}^{15} (x_i \times y_i) = 3400,5$$

Substituindo os valores na equação do coeficiente de correlação linear de Pearson:

$$r = \frac{n \times \sum_{i=1}^{n}(x_i \times y_i) - \sum_{i=1}^{n} x_i \times \sum_{i=1}^{n} y_i}{\sqrt{\left[n \times \sum_{i=1}^{n}(x_i^2) - \left(\sum_{i=1}^{n} x_i\right)^2\right]} \times \sqrt{\left[n \times \sum_{i=1}^{n}(y_i^2) - \left(\sum_{i=1}^{n} y_i\right)^2\right]}} = \frac{15 \times 3400,5 - (1335 \times 37,5)}{\sqrt{[15 \times 119165] - (1335)^2} \times \sqrt{[15 \times 107,8] - (37,5)^2}}$$

r = 0,9

Corroborando as conclusões anteriores, o coeficiente de correlação linear de Pearson teve resultado positivo, e próximo de 1, indicando forte correlação linear positiva entre a média no 2º grau e o índice na universidade ao menos para estes estudantes[8].

AVISO IMPORTANTE

É preciso avaliar a coerência entre o valor do coeficiente de correlação linear de Pearson e a disposição dos pontos no diagrama de dispersão. Às vezes valores de r próximos a -1 (correlação linear negativa) ou +1 (correlação linear positiva) não necessariamente indicam que as variáveis realmente tem correlação linear. Para o caso do Exemplo 3.5 isso realmente acontece: r > 0,9 (portanto, maior do que 0,7, indicando forte correlação linear positiva), e os pontos estão dispostos no diagrama de dispersão da Figura 6 de tal forma que seria possível ajustar uma reta crescente aos dados. Mas isso nem sempre acontece.

Exemplo 3.6 - Estamos avaliando as médias de 15 estudantes no ensino médio, relacionando-as com os índices dos mesmos estudantes no seus cursos universitários. As médias no ensino médio podem variar de 0 a 100, e os índices na universidade de 0 a 4. Construa um diagrama de dispersão e calcule o coeficiente de correlação linear de Pearson para os dados a seguir. Interprete os resultados encontrados.

[8] Na prática não se deve utilizar uma quantidade de dados tão pequena.

Média no ensino médio X	Índice na Universidade Y	X^2	Y^2	X×Y
80	0,72	6400	0,5184	57,6
82	0,36	6724	0,1296	29,52
84	0,48	7056	0,2304	40,32
85	0,79	7225	0,6241	67,15
87	1	7569	1	87
87	0,83	7569	0,6889	72,21
88	1,06	7744	1,1236	93,28
89	1,18	7921	1,3924	105,02
90	1,63	8100	2,6569	146,7
91	1,67	8281	2,7889	151,97
91	1,77	8281	3,1329	161,07
93	2,34	8649	5,4756	217,62
94	2,56	8836	6,5536	240,64
95	3,43	9025	11,7649	325,85
96	3,55	9216	12,6025	340,8

Construindo o diagrama de dispersão (há várias planilhas eletrônicas e programas estatísticos que podem fazer isso) obtemos:

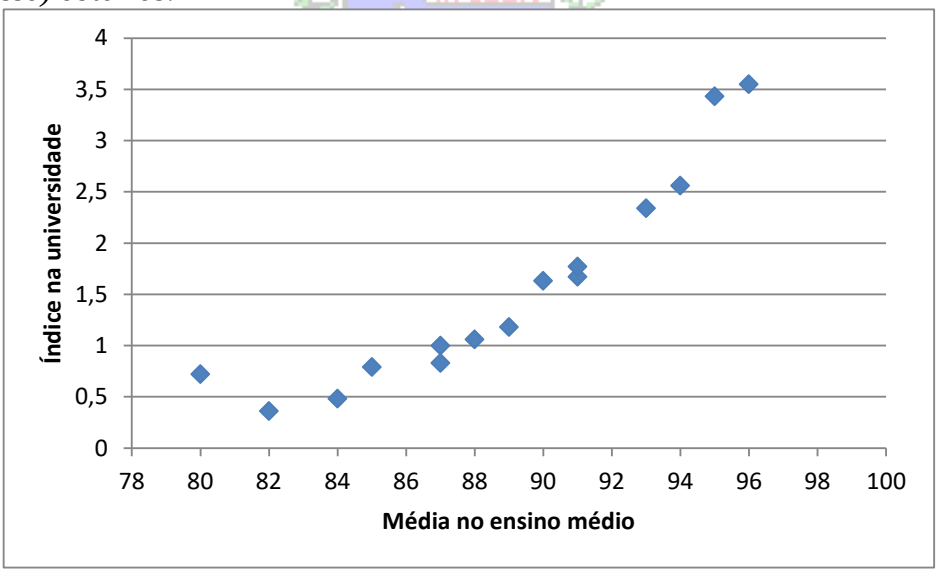


Figura 7 - Diagrama de dispersão: médias no ensino médio e índices na universidade

Observando o diagrama da Figura 7 *consegue-se vislumbrar que há uma correlação positiva entre as duas variáveis: de uma maneira geral, quanto maior o valor da média no ensino médio maior o índice na universidade. Mas, ao contrário do caso da* Figura 6, *talvez uma curva (um polinômio de 2º grau, por exemplo) fosse mais apropriada para ajustar aos dados. Então, apesar da correlação entre as variáveis parecer ser positiva, para valores MAIORES de média no ensino médio esperam-se valores também MAIORES de índice na universidade, e vice-versa, ela não parece ser linear, pois uma reta não acompanharia os dados. Como os pontos estão próximos entre si, pode-se afirmar que a correlação é forte. Ao calcular o coeficiente de correlação linear de Pearson para os dados acima será possível avaliar se há coerência ou não com a disposição dos pontos no diagrama de dispersão.*

Sabe-se que n = 15 (há 15 alunos).

$$\sum_{i=1}^{15} x_i = 1332,0 \quad \sum_{i=1}^{15} y_i = 23,27 \quad \sum_{i=1}^{15} \left(x_i^{\,2}\right) = 118596 \quad \sum_{i=1}^{15} \left(y_i^{\,2}\right) = 50,6827 \quad \sum_{i=1}^{15} \left(x_i \times y_i\right) = 2136,75$$

Substituindo os valores na equação do coeficiente de correlação linear de Pearson:

$$r = \frac{n \times \sum_{i=1}^{n}(x_i \times y_i) - \sum_{i=1}^{n} x_i \times \sum_{i=1}^{n} y_i}{\sqrt{\left[n \times \sum_{i=1}^{n}(x_i^2) - \left(\sum_{i=1}^{n} x_i\right)^2\right]} \times \sqrt{\left[n \times \sum_{i=1}^{n}(y_i^2) - \left(\sum_{i=1}^{n} y_i\right)^2\right]}} = \frac{15 \times 2136,75 - (1332 \times 23,27)}{\sqrt{[15 \times 118596] - (1332)^2} \times \sqrt{[15 \times 50,6827] - (23,27)^2}}$$

r = 0,918

Observe que o valor de r é igual a 0,918, maior do que 0,7, indicando uma forte correlação linear positiva, mas o valor está INCOERENTE com a disposição dos pontos no diagrama de dispersão da Figura 7. *Ou seja, não é possível confiar apenas no valor de r, é necessário inspecionar o diagrama de dispersão que mostra o relacionamento entre as variáveis.*

Para os dados do Exemplo 3.6 é possível ver o diagrama de dispersão da Figura 8.

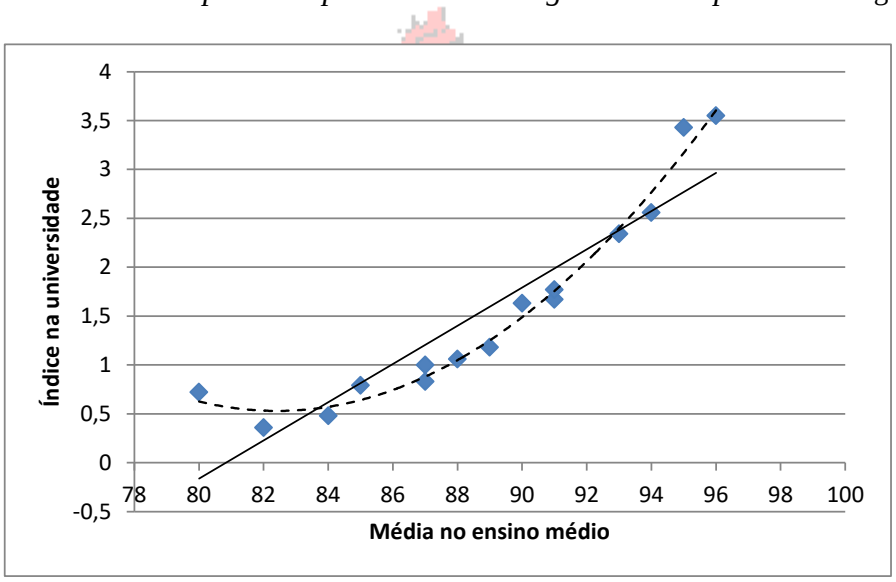


Figura 8 - Diagrama de dispersão: médias no ensino médio e índices na universidade – com duas curvas ajustadas aos pontos

Na Figura 8 *é possível ver a reta ajustada aos pontos (o processo para obter a equação da reta será mostrado adiante), a linha contínua, e o polinômio de segundo grau*[9] *(a linha interrompida). Observe como este último parece "seguir" melhor os dados do que a reta, não obstante o valor de r ser próximo a +1.*

O passo lógico seria obter uma equação que permitisse expressar o relacionamento das variáveis, de maneira que seja possível fazer previsões sobre a variável dependente a partir dos valores da variável independente.

3.2.3 - Análise de Regressão

A Análise de Regressão tem por finalidade obter uma função de regressão: uma função matemática que exprima o relacionamento entre duas ou mais variáveis. Se apenas duas variáveis estão envolvidas chama-se de regressão **simples**, se há mais de uma variável independente (e apenas uma dependente) chama-se de regressão **múltipla**.

[9] Ajustado usando o Microsoft Excel ®.

"A função de regressão 'explica' grande parte da variação de **Y** com **X**. Uma parcela da variação permanece sem ser explicada, e é atribuída ao acaso". As mesmas suposições gerais utilizadas na análise de correlação são necessárias: a existência de uma teoria que "explique" o relacionamento entre as variáveis, o pareamento dos dados, a quantidade suficiente de dados, etc.

Além desses, para realizar a Análise de Regressão, seja linear (reta), exponencial, logarítmica, polinomial, etc., alguns pressupostos básicos são necessários:
- supõe-se que há uma função que justifica **em média**, a variação de uma variável em função da variação da outra;
- os pontos experimentais (os pares x,y) terão uma variação em torno da linha representativa desta função, devido a uma variação aleatória adicional, chamada de **variância residual** ou **resíduo**;
- a variável **X** (variável INDEPENDENTE) é suposta **sem erro**.
- a variável **Y** (variável DEPENDENTE) terá uma variação nos seus valores "dependente[10]" de **X** se houver regressão.
- a função de regressão será: $\mathbf{Y} = \varphi(\mathbf{X}) + \Psi$ onde $\varphi(\mathbf{X})$ é a função de regressão propriamente dita e Ψ é a componente aleatória de **Y**, devida ao acaso (e que SEMPRE existirá).
- a variação residual de **Y** em torno da linha teórica de regressão segue uma distribuição normal com média zero e desvio padrão constante (independente dos valores de **X**).

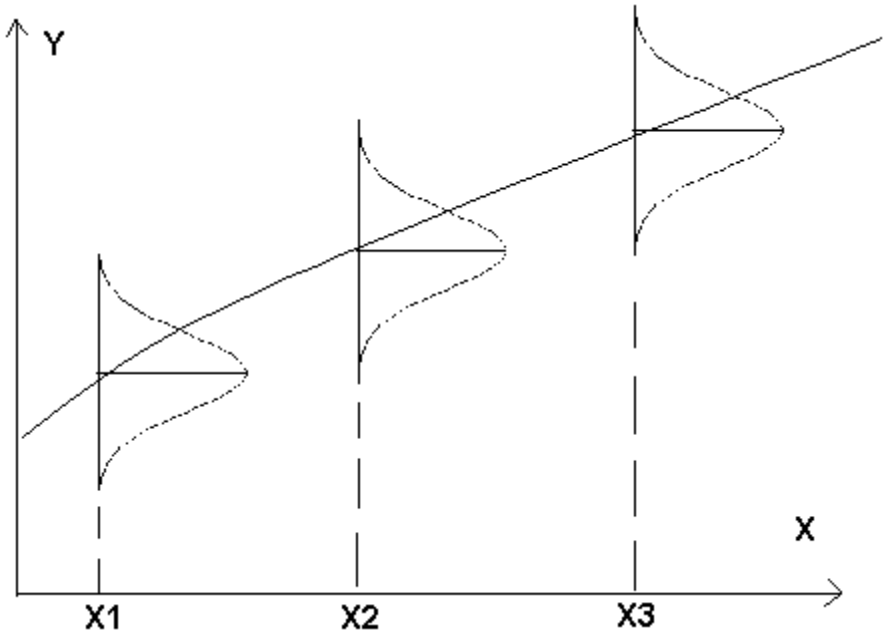


Figura 9 - Variação residual em torno da linha teórica de regressão

- para se decidir pela utilização de um modelo de regressão devem existir evidências NÃO ESTATÍSTICAS que indiquem relação causal entre as variáveis (alguma lei da física por exemplo, como a Lei de Hook).

Uma vez conhecida a forma da linha de regressão o problema resume-se a **estimar seus parâmetros**.

[10] Foi colocado entre aspas porque a existência de regressão NÃO IMPLICA necessariamente em que Y depende de X, apenas que elas têm uma variação relacionada, que pode ser causada por uma outra variável.

3.2.4 - Análise de Regressão Linear Simples

Restringe-se a análise a apenas DUAS variáveis, e supõe-se que a linha teórica de regressão é uma reta. Este modelo é bastante difundido porque muitos relacionamentos entre variáveis podem ser descritos através de uma reta, seja utilizando os dados originais, seja após aplicar alguma transformação (logarítmica, exponencial, etc.) a eles que cause a *linearização* da curva.

A reta teórica será $\mathbf{Y} = \alpha + \beta\,\mathbf{X}$ e os coeficientes α e β serão estimados através dos valores amostrais **a** e **b** respectivamente: $\hat{Y} = a + bX$, onde $\hat{Y}$ é a estimativa de **Y**, b é o coeficiente angular da reta (a sua inclinação), e a é o coeficiente linear (o ponto onde a reta toca o eixo Y).

A "melhor reta" será encontrada pelo método dos mínimos quadrados: são encontrados os coeficientes a e b que minimizam os quadrados dos desvios[11] de cada ponto do diagrama de dispersão em relação a uma reta teórica.

$$Desvio_i = Y_i - \hat{Y}_i$$

O Desvio em um ponto i qualquer é a diferença entre o valor OBSERVADO da variável Y naquele ponto e o valor PREDITO da variável Y pelo modelo de regressão (qualquer modelo) naquele mesmo ponto. O método dos mínimos quadrados procura encontrar os valores dos coeficientes do modelo de regressão, no caso da reta a e b tal que a soma dos quadrados dos desvios (considerando todos os pontos sob análise) seja a menor possível.

Para o caso específico da reta tem-se os seguintes valores de **a** e **b**:

$$b = \frac{n \times \sum_{i=1}^{n}(x_i \times y_i) - \sum_{i=1}^{n} x_i \times \sum_{i=1}^{n} y_i}{n \times \sum_{i=1}^{n}(x_i^{\,2}) - \left(\sum_{i=1}^{n} x_i\right)^2} \qquad a = \frac{\sum_{i=1}^{n} y_i - b \times \sum_{i=1}^{n} x_i}{n}$$

Muitas calculadoras já têm estas fórmulas programadas em um módulo estatístico (juntamente com a fórmula do coeficiente de correlação linear de Pearson). Além disso, planilhas eletrônicas e programas estatísticos também fazem tais cálculos.

Exemplo 3.7 - Calcule os coeficientes da reta de mínimos quadrados para os dados do Exemplo 3.5.

Conforme visto no Exemplo 3.5 as variáveis média no ensino médio e índice na universidade apresentam alta correlação linear positiva, o que é mostrado pelo diagrama de dispersão e pelo coeficiente de correlação linear de Pearson. Ajustar uma reta aos dados parece ser uma boa idéia, e todos os somatórios necessários foram calculados no Exemplo 3.5, a saber:

$$\sum_{i=1}^{15} x_i = 1335{,}0 \qquad \sum_{i=1}^{15} y_i = 37{,}5 \qquad \sum_{i=1}^{15}(x_i^{\,2}) = 119165{,}0 \qquad \sum_{i=1}^{15}(x_i \times y_i) = 3400{,}5 \qquad n = 15$$

Substituindo os valores nas equações de b e a:

$$b = \frac{n \times \sum_{i=1}^{n}(x_i \times y_i) - \sum_{i=1}^{n} x_i \times \sum_{i=1}^{n} y_i}{n \times \sum_{i=1}^{n}(x_i^{\,2}) - \left(\sum_{i=1}^{n} x_i\right)^2} = \frac{15 \times 3400{,}5 - (1335 \times 37{,}5)}{15 \times 119165 - (1335)^2} = 0{,}18$$

[11] Também chamados de resíduos.

$$a = \frac{\sum_{i=1}^{n} y_i - b \times \sum_{i=1}^{n} x_i}{n} = \frac{37{,}5 - 0{,}18 \times 1335}{15} = -13{,}52$$

A equação da reta será então: $\hat{Y} = -13{,}52 + 0{,}18 \times X$ *ou* $\hat{Y} = 0{,}18 \times X - 13{,}52$

Vejamos como ficaria o diagrama de dispersão com a reta acima traçada sobre ele.

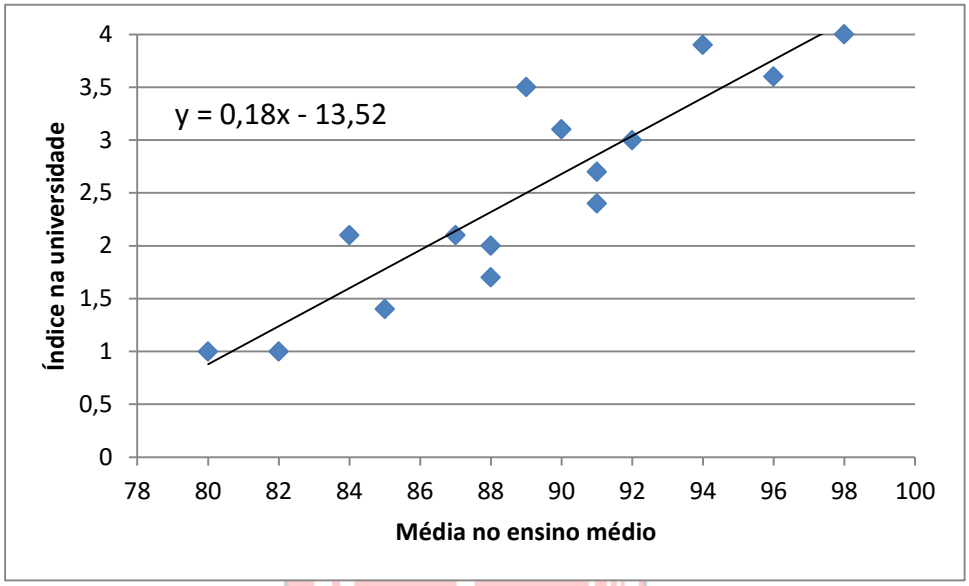


Figura 10 - Diagrama de dispersão: média no ensino médio e índice na universidade - reta ajustada

Diversos aplicativos computacionais (como o Microsoft Excel) permitem obter os coeficientes de mínimos quadrados para vários modelos de regressão: linear, polinômios de vários graus, logarítmico, exponencial, potência, etc. Abaixo são mostrados alguns deles, permitindo estimar os valores de Y através dos valores de X (a estimativa de Y é denotada como $\hat{Y}$):

- linear (reta) - $\hat{Y} = b \times X + a$;
- polinômio de segundo grau - $\hat{Y} = c \times X^2 + b \times X + a$
- logarítmico - $\hat{Y} = b \times Ln(X) + a$;
- potência - $\hat{Y} = b \times X^a$;
- exponencial[12] - $\hat{Y} = b \times e^{a \times X}$

Se houver mais de uma variável independente (preditora), a regressão é chamada de múltipla, e devido à complexidade matemática para obtenção dos coeficientes do modelo, mesmo para o caso linear, é recomendável a utilização de aplicativos computacionais.

Os modelos de regressão obtidos pelo método dos mínimos quadrados[13] podem ser extremamente úteis para processos de previsão, mas eles têm três limitações sérias.

A primeira limitação é que o modelo é obtido para dados que foram coletados em determinadas condições, se fatos posteriores alterarem tais condições o modelo de regressão pode perder sua utilidade. Isso é especialmente importante para dados socioeconômicos ou financeiros, nos quais as mudanças de legislação, governos e eventuais avanços tecnológicos podem

[12] Ln significa logaritmo natural, um logaritmo cuja base é o número de Euler, igual a 2,718281828459045... e **e** significa o número de Euler

[13] Há outros métodos de estimação dos parâmetros dos modelos, como o de máxima verossimilhança.

simplesmente tornar os dados anteriores obsoletos.

A segunda limitação é a amplitude dos dados, lembrando do conceito de intervalo do Capítulo 2: o modelo de regressão é calculado, e somente é válido, para os valores da variável independente X situados *entre* o mínimo e o máximo dos valores coletados. Para o caso do Exemplo 3.5, o modelo obtido no Exemplo 3.7 vale apenas para médias no ensino médio *ENTRE* 80 e 98. Para alunos com média abaixo de 80, ou acima de 98, o modelo teoricamente não é válido: ele permite fazer cálculos de interpolação (por exemplo, para alguém com média no ensino médio igual a 93, inexistente no conjunto de dados, mas entre 80 e 98), mas não há grande confiança em cálculos de extrapolação para valores de X fora do intervalo.

A terceira limitação reside na quantidade dos dados. O objetivo é obter um modelo de regressão robusto, ou seja, um modelo que seja pouco afetado por valores discrepantes. Observe atentamente as figuras a seguir.

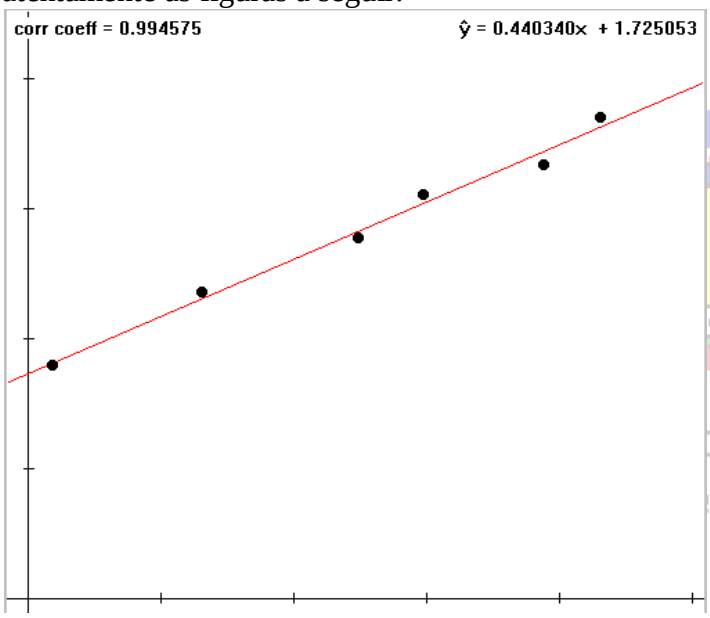


Figura 11 - Diagrama de dispersão - poucos dados - 1º caso

Há apenas seis pontos neste diagrama, e por sua disposição é possível perceber que há forte correlação linear entre as variáveis. O coeficiente de correlação linear de Pearson foi calculado, está no canto superior da figura, e é igual a 0,9945, quase igual a 1, indicando fortíssima correlação linear positiva.

A reta traçada por entre os pontos quase passa por todos eles, e trata-se de uma reta crescente (coeficiente angular igual a 0,440, no canto superior direito da figura). Mas, a quantidade de dados é muito pequena, e se ocorresse um valor discrepante? Veja o que acontece na Figura 12.

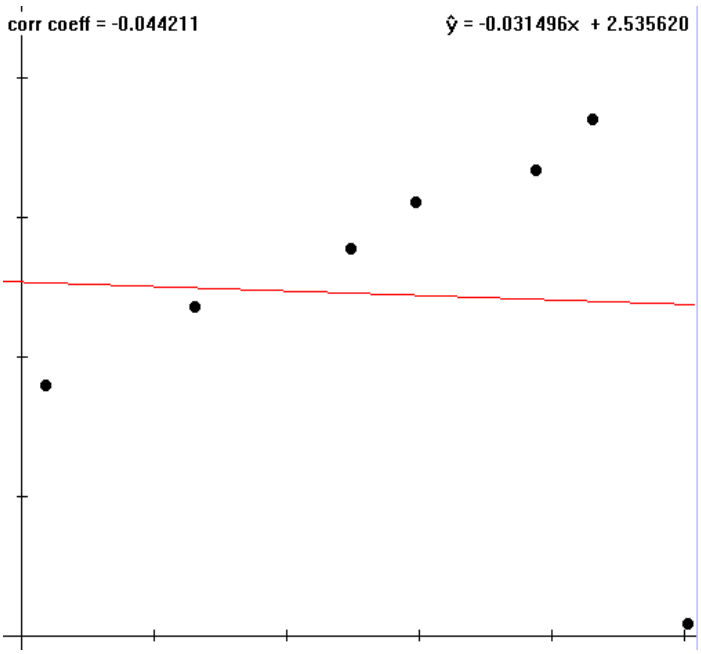
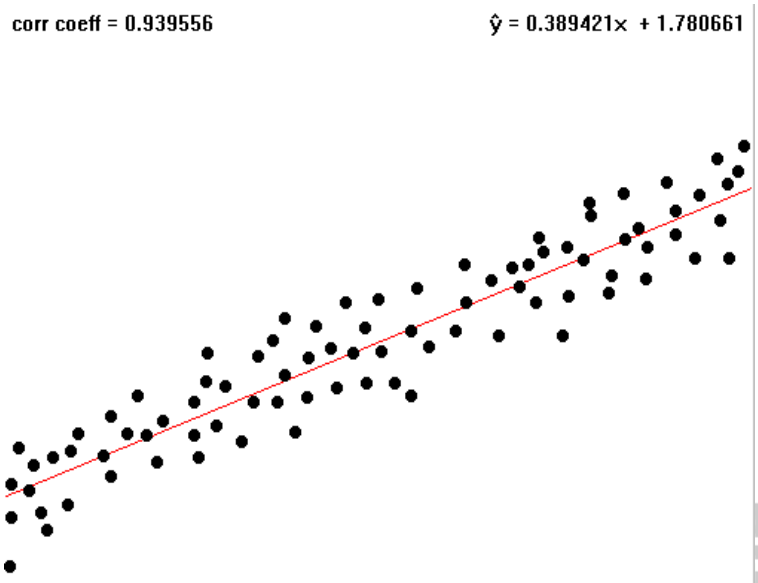


Figura 12 - Diagrama de dispersão - poucos dados - 2o caso

Foi acrescentado apenas um ponto ao conjunto mostrado na Figura 11. Mas este ponto é discrepante, no canto inferior direito da figura, e seu efeito foi devastador, devido à pequena quantidade de dados.

O coeficiente de correlação linear caiu para -0,044, indicando correlação linear quase nula, e a reta que era crescente passou a ser decrescente (coeficiente angular igual a -0,031). Decisões tomadas a partir deste conjunto poderiam ser tremendamente prejudicadas, simplesmente devido à pequena quantidade de dados.

Imagine agora uma situação em que fosse possível coletar uma grande quantidade de dados, para as mesmas duas variáveis, e um diagrama de dispersão fosse construído, tal como o da Figura 13.



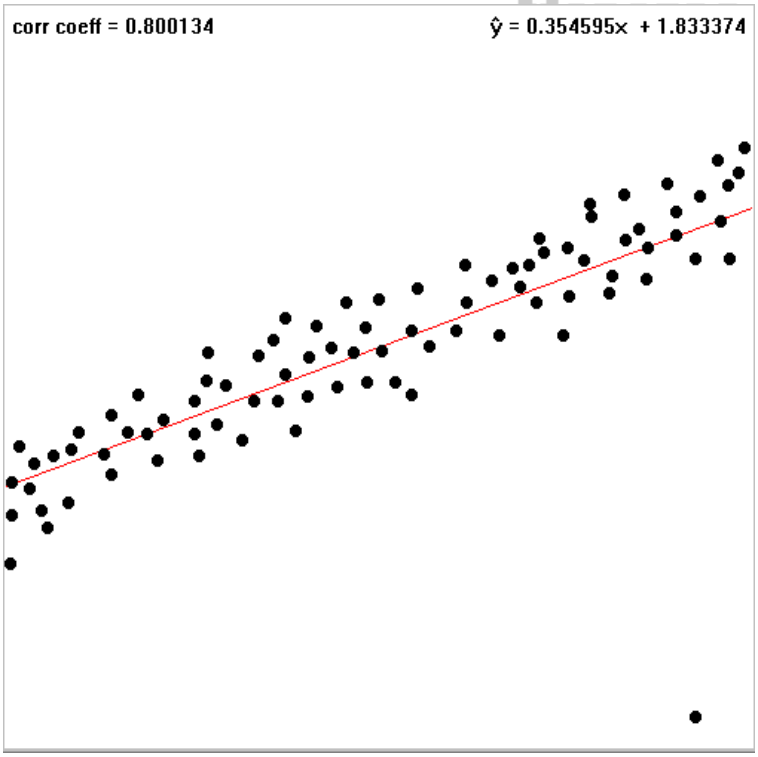
Pela disposição dos dados é fácil perceber que há correlação linear positiva entre as variáveis. Há uma "nuvem" de pontos que indica que à medida que aumentam os valores de X aumentam os de Y.

O coeficiente de correlação linear de Pearson vale 0,9395, indicando forte correlação linear positiva.

A reta ajustada aos dados é crescente, com o coeficiente angular valendo 0,3894.

Devido à grande quantidade de dados mesmo que ocorram alguns valores discrepantes seu efeito não será tão marcante quanto foi no caso mostrado na Figura 12. Veja a Figura 14.

Figura 13 - Diagrama de dispersão com muitos dados - 1º caso



Apesar do valor discrepante (no canto inferior direito da Figura 14), não houve grande mudança na equação da reta e no coeficiente de correlação linear de Pearson.

O coeficiente de correlação linear de Pearson caiu de 0,9395 para 0,8001, ainda indicando forte correlação linear positiva, um visível contraste com o que ocorreu na Figura 12.

Já o coeficiente angular da reta caiu menos ainda, de 0,3894 para 0,3545, indicando robustez no modelo.

Figura 14 - Diagrama de dispersão com muitos dados - 2º caso

Sempre que possível devemos coletar a maior quantidade possível de dados, seja regressão simples ou múltipla, para que o modelo obtido seja robusto e não sofra grandes alterações devido aos valores discrepantes.

3.2.5 - Coeficiente de Determinação

Alguns novos conceitos precisam ser introduzidos:

$\overline{Y}$ é a média aritmética dos valores **observados** de **Y** (n é o número total de observações).

$\hat{Y}_i$ é um valor genérico **predito** de **Y** através do modelo de regressão (qualquer modelo).

$\sum_{i=1}^{n}\left(Y_i-\overline{Y}\right)^2$: medida da variabilidade **total** dos dados em torno da média de **Y**.

$\sum_{i=1}^{n}\left(\hat{Y}_i-\overline{Y}\right)^2$: medida da variabilidade dos dados em torno da média de **Y** "**explicada**" pela regressão.

$\sum_{i=1}^{n}\left(Y_i-\hat{Y}_i\right)^2$: medida da variabilidade dos dados em torno da média de **Y** "**não explicada**" pela regressão, chamada também de variação **residual**.

E: $\sum_{i=1}^{n}\left(Y_i-\overline{Y}\right)^2=\sum_{i=1}^{n}\left(\hat{Y}_i-\overline{Y}\right)^2+\sum_{i=1}^{n}\left(Y_i-\hat{Y}_i\right)^2$ (variação explicada + variação residual = variação total).

Neste ponto é interessante introduzir o coeficiente de determinação, $\mathbf{r^2}$. Este coeficiente descreve a proporção da variabilidade média de **Y** que é explicada pela variação de **X** através do modelo de regressão (QUALQUER modelo). Sua fórmula geral é:

$$r^2=\frac{\sum_{i=1}^{n}\left(\hat{Y}_i-\overline{Y}\right)^2}{\sum_{i=1}^{n}\left(Y_i-\overline{Y}\right)^2}=\frac{variância\,explicada}{\text{var}\,iância\,total}$$

Para o caso linear o r^2 será simplesmente o quadrado do coeficiente de correlação linear de Pearson (**r**), e como ele será um valor adimensional, mas pode variar apenas de 0 a +1. O r^2 é uma boa medida da aderência do modelo de regressão aos dados, quanto mais próximo de +1 maior a parcela da variabilidade média total de **Y** que é explicada pela variação de **X** através do modelo.

Para r^2 superiores a 0,5 (mais de 50% da variabilidade média total de Y é explicada pela variação de X através do modelo de regressão). Para o caso *linear* isso significa que o **módulo** do coeficiente r deve ser maior do que 0,7 para que a regressão linear seja uma boa opção.

Exemplo 3.8 - Calcule e interprete o resultado do coeficiente de determinação para o modelo linear ajustado no Exemplo 3.7.

Como se trata de um modelo linear, podemos obter o coeficiente de determinação elevando o coeficiente de correlação linear de Pearson (calculado no Exemplo 3.5) ao quadrado.

$$r^2 = 0{,}9^2 = 0{,}81$$

Em média 81% da variabilidade de Y pode ser "explicada" pela variabilidade de X através do modelo linear $\hat{Y}=-13{,}52+0{,}18\times X$.

O valor do r^2 *é substancialmente maior do que 0,5, indicando que o modelo linear apropriado para os dados (corroborando as conclusões dos Exemplos 3.5 e 3.7).*

Embora útil, o coeficiente de determinação não é suficiente para avaliar se um modelo de regressão é apresenta bom ajuste aos dados. Precisamos fazer uma análise dos resíduos do modelo.

3.2.6 - Análise de resíduos

Idealmente a adequação de um modelo de regressão é realizada através da análise dos seus resíduos. Os resíduos são as diferenças entre os valores ***observados*** da variável independente e os valores ***preditos*** da variável independente através do modelo de regressão. Para tornar a análise mais confiável, sem que as grandezas dos resíduos venham a prejudicá-la recomenda-se ***padronizar*** os resíduos: calcula-se o desvio padrão dos resíduos e divide-se cada um deles pelo desvio padrão.

Para fazer a análise de resíduos precisamos construir pelo menos dois diagramas de dispersão:
- um que relacione os resíduos padronizados com os próprios valores preditos da variável independente;
- outro que relacione os resíduos padronizados com os valores da variável independente[14].

Se o modelo de regressão é adequado os resíduos padronizados não podem apresentar quaisquer padrões, eles devem distribuir-se de forma aleatória nos dois diagramas, atendendo os seguintes critérios:
- a quantidade de resíduos padronizados positivos deve ser aproximadamente igual à quantidade de negativos.
- a grandeza dos resíduos padronizados positivos deve ser aproximadamente igual a dos negativos, para todos os valores preditos da variável dependente, e para todos os valores da variável independente.
- não pode haver padrões não aleatórios (tendências crescentes ou decrescentes, curvas, etc.) em nenhum dos diagramas; em outras palavras é preciso que os pontos sejam dispostos em "**nuvem**".

Somente se ***todas*** estas condições forem satisfeitas é que podemos considerar o modelo de regressão apropriado. Se houver dois ou mais modelos apropriados escolhemos o mais simples, ou aquele que apresentar o mais alto coeficiente de determinação. Os diagramas deveriam ser como a Figura 15.

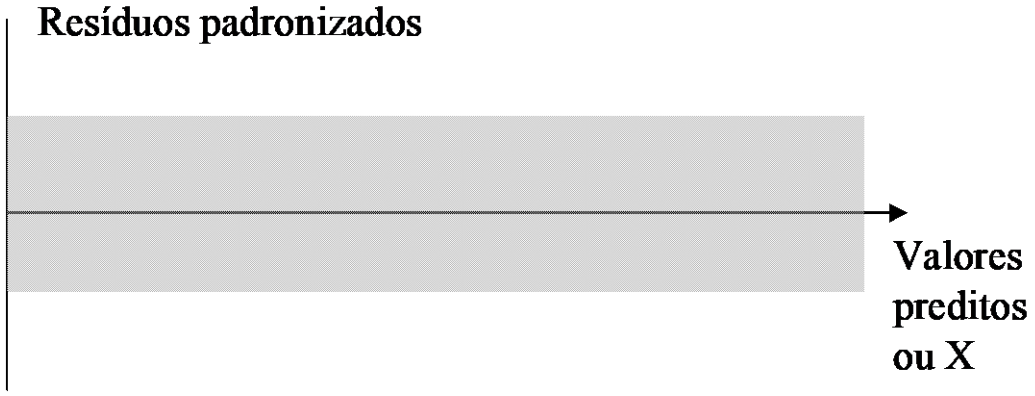


Figura 15 - Formato esperado dos resíduos se modelo é apropriado

Exemplo 3.9 - Estamos avaliando o relacionamento entre as variáveis venda de refrigerantes e temperatura ambiente nos meses de verão. Na Figura 16 vemos o diagrama de dispersão das duas variáveis (temperatura é a independente e vendas é a dependente), com dois modelos ajustados através do Microsoft Excel: reta e parábola (polinômio de 2^{o} grau). Queremos saber qual dos dois modelos é mais apropriado através da análise de seus resíduos. As Figuras 17 e 18 apresentam os diagramas de dispersão dos resíduos padronizados (em função da temperatura e dos valores preditos

[14] Se houver mais de uma variável independente faz-se um diagrama de dispersão para cada uma delas.

pelo modelo de regressão) para a reta, e as Figuras 19 e 20 apresentam os respectivos diagramas para a parábola.

a) Faça a análise do diagrama de dispersão das variáveis. Na sua opinião qual dos modelos apresenta o melhor ajuste aos dados?

b) Faça a análise dos resíduos para o modelo da reta.

c) Faça a análise dos resíduos para o modelo da parábola.

d) Com base nas respostas anteriores, qual dos dois modelos parece ser o mais apropriado para descrever o relacionamento entre as variáveis?

e) Utilizando o modelo escolhido no item d, faça a previsão de vendas para os seguintes valores de temperatura: e.1 - 27^o C e.2 - 32^o C e.3 - 38^o C

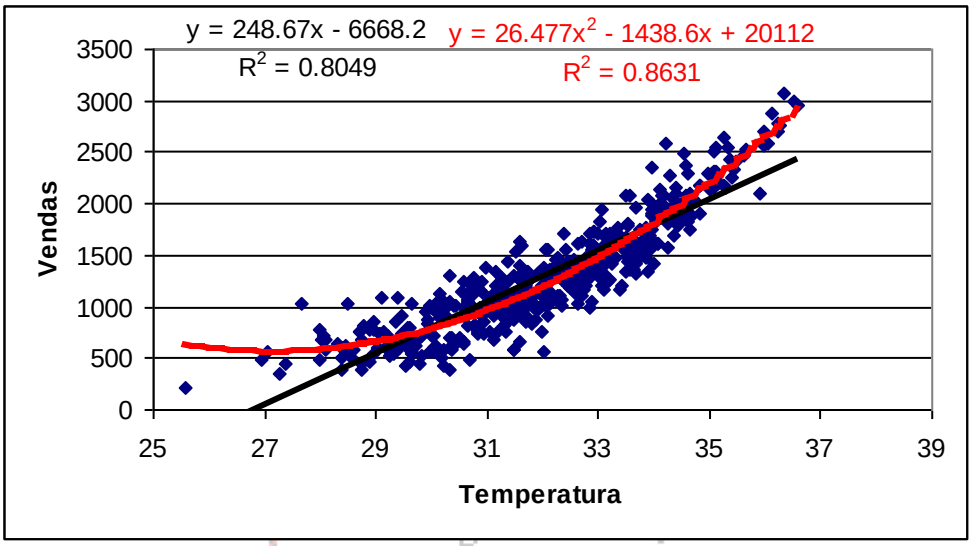


Figura 16 - Diagrama de dispersão vendas por temperatura: ajuste de reta e parábola

a) Observando o diagrama podemos ver que a parábola (polinômio de 2^o grau) aparenta ter melhor ajuste aos dados, pois ela "segue" melhor o seu comportamento do que a reta. Os resíduos do modelo de parábola provavelmente serão menores do que os da reta, o que pode ser constatado também pelo seu coeficiente de determinação (0,8631), que é maior do que o da reta (0,8049). Ambos os modelos, porém, conseguem "explicar" grande parte da variação média das vendas, pois seus coeficientes de determinação são substancialmente maiores do que 0,5.

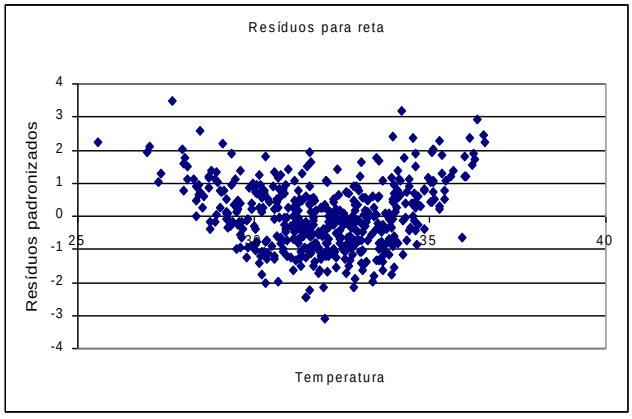


Figura 17 - Resíduos da reta por temperatura

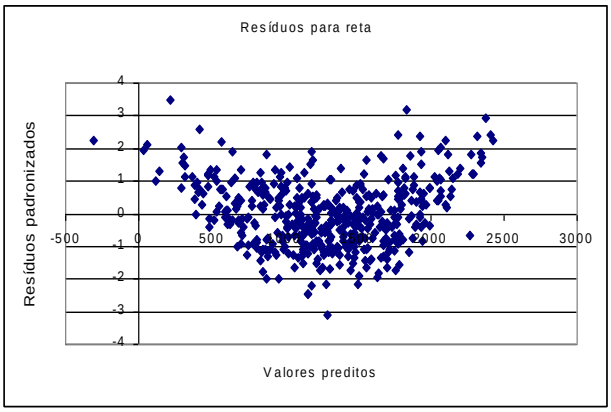


Figura 18 - Resíduos da reta por valores preditos

b) Devemos levar em conta os três aspectos mencionados anteriormente.

- Número de resíduos positivos e negativos. Aparentemente a quantidade de resíduos padronizados positivos e negativos é semelhante (deveríamos contá-los por meio de algum procedimento computacional), a linha do zero parece "dividir" o número de pontos em duas partes iguais em ambos os diagramas.

- Grandeza dos resíduos positivos e negativos. A maioria esmagadora dos pontos positivos

concentra-se abaixo de 2 desvios padrões (linha do 2), e maioria dos negativos também (acima da linha -2), em ambos os diagramas.
- Existência de padrões. Há claramente padrão em ambos os diagramas. Para valores menores de temperatura e valores preditos os resíduos são positivos e maiores. À medida que a temperatura e os valores preditos vão aumentando os valores dos resíduos vão diminuindo, tornando-se negativos, até que passam a subir novamente. Em outras palavras, o comportamento dos resíduos do modelo da reta NÃO É ALEATÓRIO.

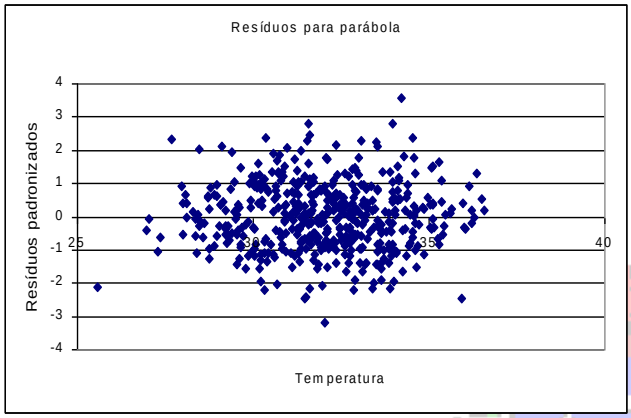


Figura 19 - Resíduos da parábola por temperatura

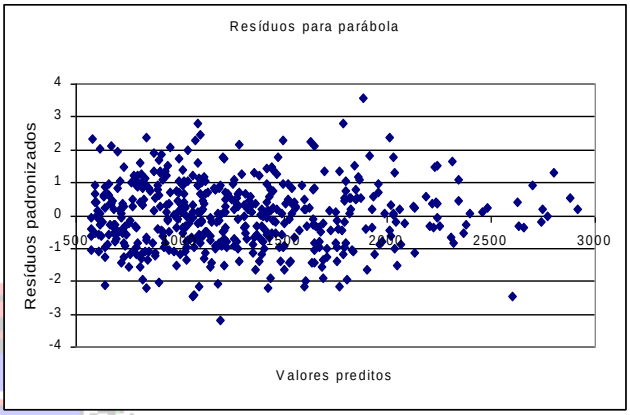


Figura 20 - Resíduos da parábola por valores preditos

c) Para o caso da parábola vamos avaliar novamente os três aspectos.
- Número de resíduos positivos e negativos. A quantidade de resíduos positivos e negativos é aparentemente bastante semelhante em ambos os diagramas (a linha do zero divide os pontos em duas "metades" similares).
- Grandeza dos resíduos positivos e negativos. Em ambos os diagramas os resíduos positivos e negativos têm grandezas semelhantes, distantes no máximo a 2 desvios padrões do zero, para a maioria dos pontos.
- Existência de padrões. Em ambos os diagramas NÃO são identificados padrões, os pontos parecem distribuir-se de forma aleatória, formando uma "nuvem".

d) Com base na análise de resíduos o modelo da parábola (polinômio de 2º grau) é o mais apropriado para descrever o relacionamento entre vendas de refrigerante e temperatura ambiente, porque os seus resíduos distribuem-se aleatoriamente, tanto em função dos valores da variável independente quanto dos valores preditos pelo próprio modelo.

e) O modelo de parábola estimado pelo Microsoft Excel é (ver Figura 16*, sendo Y = Vendas e X = Temperatura):*
$Vendas = 26{,}477 \times Temperatura^2 - 1438{,}6 \times Temperatura + 20112$
Para fazer as previsões basta substituir os valores da temperatura na equação acima.
e.1 - 27º C: $Vendas = 26{,}477 \times (27)^2 - 1438{,}6 \times 27 + 20112 = 571{,}533$
e.2 - 32º C: $Vendas = 26{,}477 \times (32)^2 - 1438{,}6 \times 32 + 20112 = 1189{,}248$
e.3 - 38º C: $Vendas = 26{,}477 \times (38)^2 - 1438{,}6 \times 38 + 20112 = 3677{,}988$
Mas, como não há temperatura de 38º C registrada nos dados originais (conferir na Figura 16*), a previsão de vendas para esta temperatura não é confiável.*

REGRA IMPORTANTE:

E se a análise de resíduos identificar que ***todos*** os modelos são apropriados? Neste caso devemos selecionar aquele que apresentar o maior coeficiente de determinação. Se, porém, os modelos tiverem coeficientes de determinação próximos (diferenças inferiores a 5%) devemos ser parcimoniosos, e escolher o modelo mais simples (regra da PARCIMÔNIA).