INE 7001 - Procedimentos de Análise Bidimensional de variáveis QUANTITATIVAS utilizando o Microsoft Excel.
Professor Marcelo Menezes Reis

O objetivo deste texto é apresentar os principais procedimentos de Análise Bidimensional de variáveis quantitativas, tal como apresentados em sala, mas utilizando a planilha eletrônica Excel. Os dados estão na planilha "Temperatura e vendas", do arquivo Bidimensional.xls, disponível nas páginas das disciplinas: contém as informações sobre 250 pares de observações temperatura (em graus Celsius) e quantidade vendida de refrigerantes.

Os procedimentos foram preparados utilizando a versão 2003 do Excel. Há algumas diferenças em relação às versões mais modernas (2007, 2010), mas a essência permanece a mesma.

1. Construção de diagrama de dispersão para as variáveis.

No presente caso, em que há apenas 2 variáveis, é possível construir um diagrama de dispersão, relacionando temperatura e vendas. O objetivo é avaliar a força, a direção e a forma de uma eventual correlação entre elas: com isso será possível avaliar qual modelo de regressão aplicar para prever os valores de uma variável em função dos da outra. Os dados de interesse estão mostrados na figura 1:

	A	B
1	Temperatura	Vendas
2	31.19	1321
3	31.28	1492
4	29.85	1495
5	32.41	1386
6	32.17	1672
7	31.87	1498
8	34.89	2702
9	30.84	1413
10	36.21	3252
11	31.36	1502
12	33.83	1937
13	33.45	2136
14	29.82	1480
15	33.27	2014

Figura 1 - Temperatura e vendas

Na coluna A encontram-se os valores de Temperatura, e na coluna B os das Vendas. É preciso identificar corretamente qual variável é a independente e qual é a dependente: caso contrário o diagrama estará completamente errado, o modelo eventualmente ajustado também, e as decisões tomadas com base neles pouca validade terão. É razoável imaginar que a Temperatura possa influenciar as Vendas de refrigerante: maiores valores de Temperatura poderiam causar maiores valores de Vendas. Sendo assim, Temperatura será a variável independente, sendo então representada no eixo X, e Vendas a variável dependente, ocupando o eixo Y.
Passamos agora a construção do diagrama de dispersão propriamente dito, clicando sobre o ícone "Assistente Gráfico", na barra de ferramentas do Excel, resultando na figura 2. Selecionando o gráfico Dispersão (XY), obtemos a figura 3.

Figura 2 - Assistente gráfico - 1a etapa

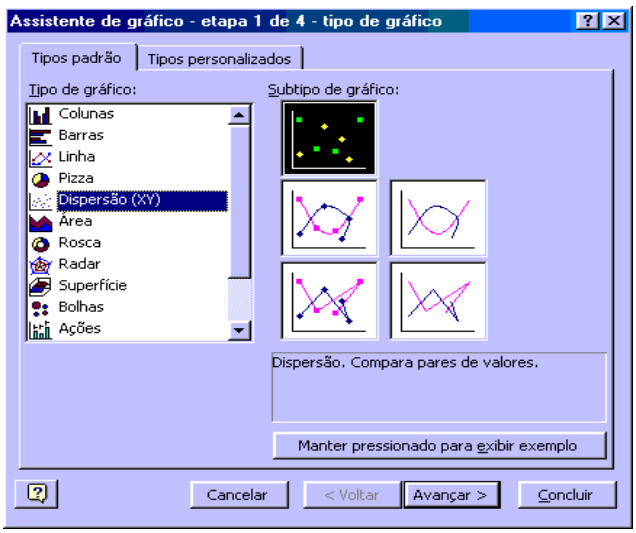


Figura 3 - Assistente gráfico - Diagrama de dispersão

Para os nossos interesses o subtipo mais interessante é o padrão, marcado em preto na figura 3. Pressionando "Avançar" chegaremos a uma tela semelhante à figura 4.

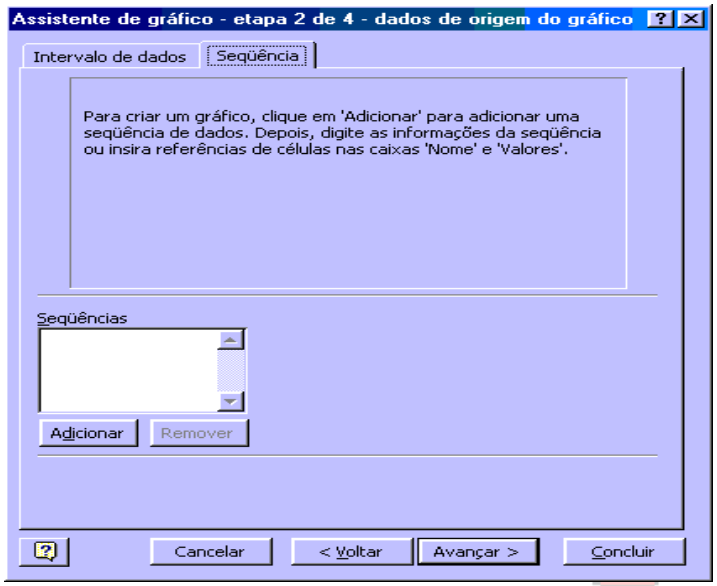


Figura 4 - Assistente gráfico - 2a etapa

Em alguns casos o Excel automaticamente adiciona as seqüências de dados necessárias para criar o gráfico. Muitas vezes estas seqüências incluem dados que não nos interessam. Se isso ocorrer, pressione "Remover" até que todas as seqüências sejam retiradas, resultando na tela mostrada na figura 4.
Agora podemos adicionar as seqüências de dados de interesse, pressionando "Adicionar", o que resultará na figura 5.

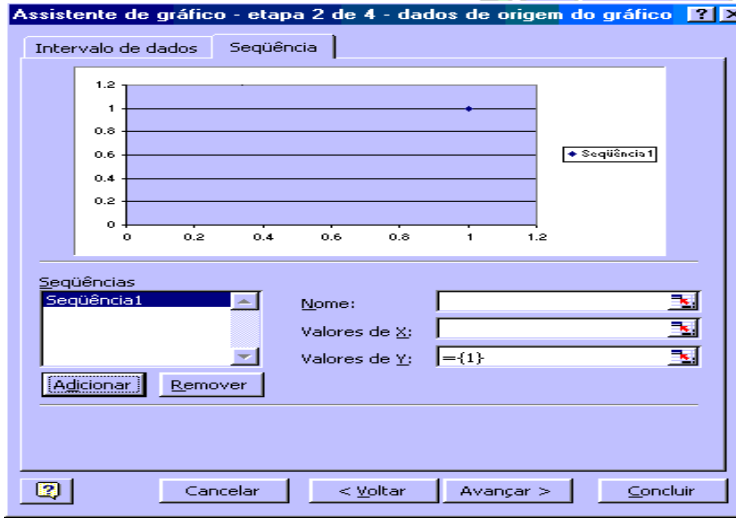


Figura 5 - Assistente gráfico: adição de seqüências

Precisamos adicionar os valores de X e de Y (não há necessidade de adicionar valores em "Nome"). Podemos fazer isso de duas formas: ou digitando as referências das células (em "Valores de X" teríamos A2:A251; em "Valores de Y" teríamos B2:B251), ou marcando as células na planilha (pressionando a seta vermelha na extrema direita de cada janela, e marcando as células de interesse na planilha).
Após a adição dos dados, o resultado será uma tela semelhante à da figura 6.

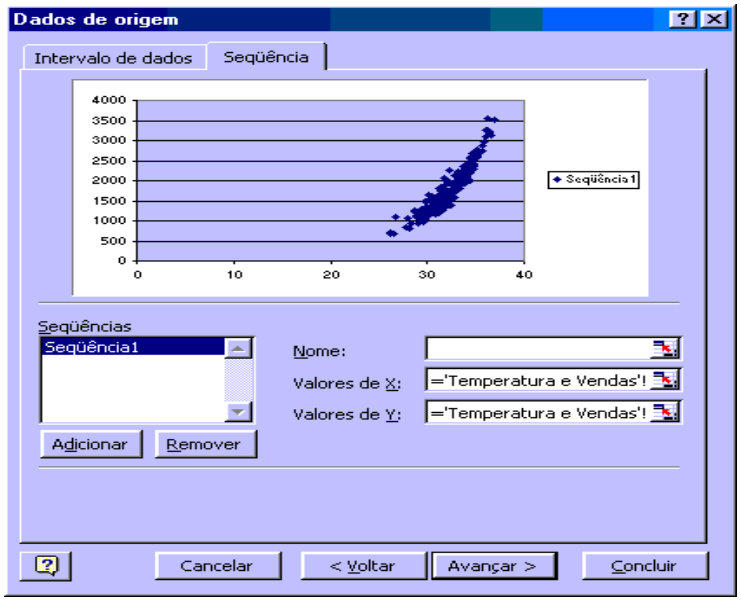


Figura 6 - Assistente gráfico: dados inseridos

Observe que já é possível ter uma idéia do diagrama de dispersão: os dados parecem distribuir-se de forma curva, com os valores de X começando acima de 20, e os valores de Y variando de 500 até quase 4000. Possivelmente teremos que modificar a escala do eixo X, para que a visualização do gráfico seja mais apropriada: da forma como está o gráfico os dados estão muito agrupados, o que pode dificultar a análise do diagrama de dispersão.
Pressionando "Avançar" chegaremos à tela mostrada na figura 7.

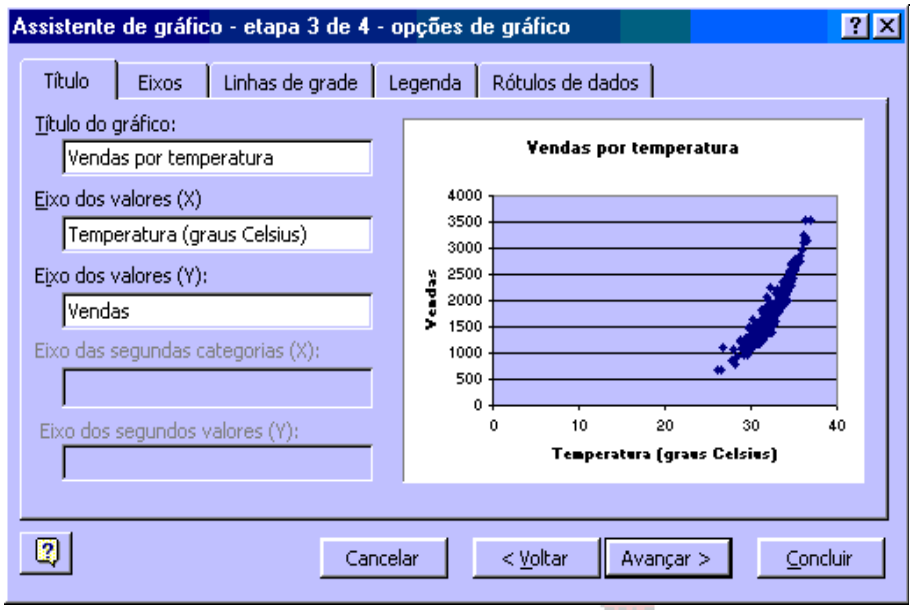


Figura 7 - Assistente gráfico - 3a etapa

É necessário pôr um título no gráfico, e identificar as variáveis em cada eixo, incluindo suas unidades.
Título: Vendas por temperatura.
Eixo X: temperatura (em graus Celsius).
Eixo Y: Vendas.
Retiramos a legenda, pois não há necessidade neste gráfico.
Ao pressionar "Avançar" chegamos na tela mostrada na figura 8.

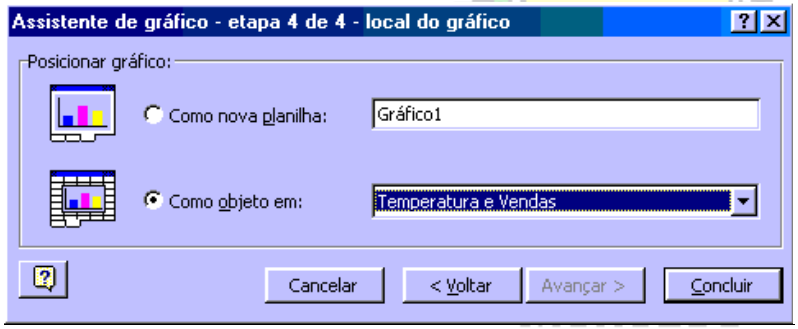


Figura 8 - Assistente gráfico - 4a etapa

Escolhe-se onde queremos que o diagrama seja posicionado. Selecionando "Como objeto em:" o gráfico será colocado na planilha onde estão os seus dados, o que pode ser mais interessante. O diagrama resultante está na figura 9.

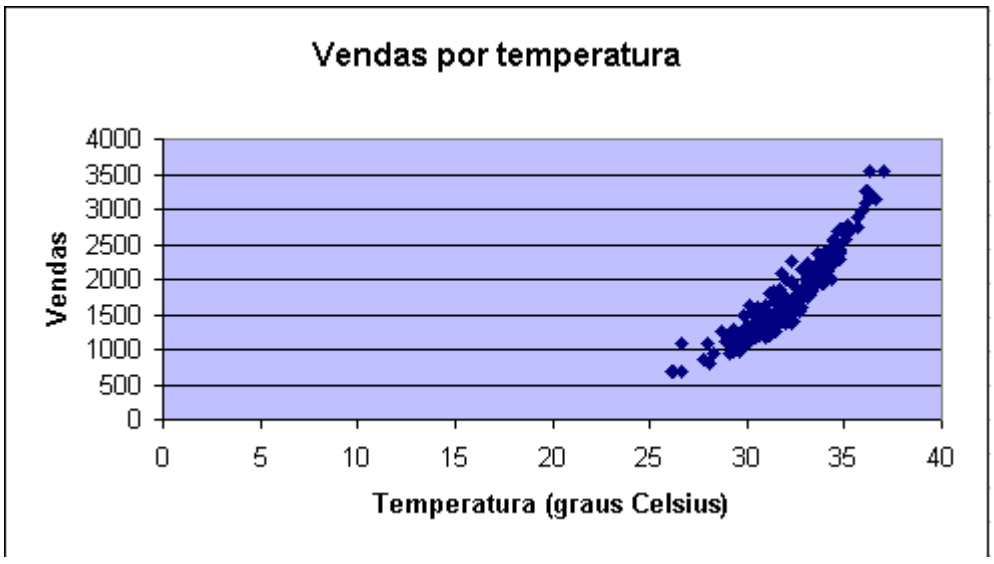
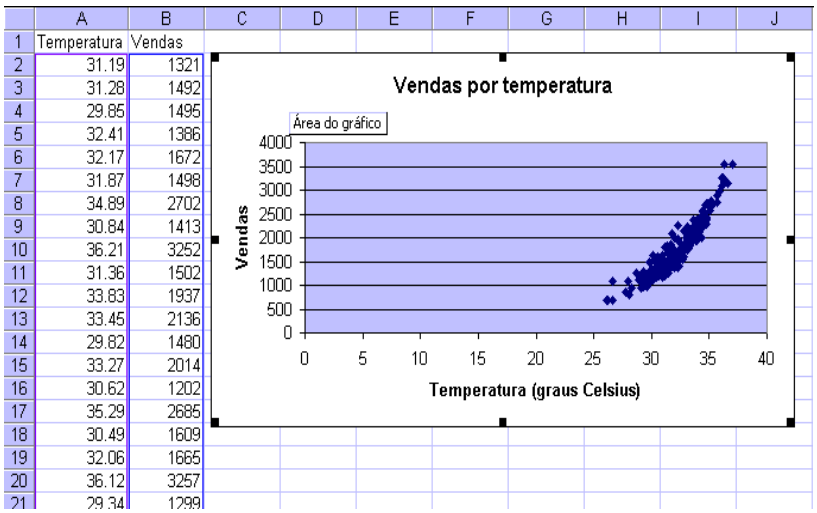


Figura 9 - Diagrama de dispersão: Vendas por temperatura

Se colocarmos o mouse sobre o gráfico (na parte branca) e pressionarmos o botão esquerdo, teremos uma situação semelhante à mostrada na figura 10.



	A	B
1	Temperatura	Vendas
2	31.19	1321
3	31.28	1492
4	29.85	1495
5	32.41	1386
6	32.17	1672
7	31.87	1498
8	34.89	2702
9	30.84	1413
10	36.21	3252
11	31.36	1502
12	33.83	1937
13	33.45	2136
14	29.82	1480
15	33.27	2014
16	30.62	1202
17	35.29	2685
18	30.49	1609
19	32.06	1665
20	36.12	3257
21	29.34	1299

Figura 10 - Diagrama de dispersão: gráfico e dados

Observe que ao selecionar o gráfico as células que contém os dados que o geraram tem suas bordas coloridas, o que pode ser útil para avaliar se não houve erros ou falta de alguns valores. O gráfico das figuras 9 e 10 apresenta alguns problemas: a escala do eixo X deixou os dados muito próximos, o que pode dificultar a análise do diagrama; o fundo cinza do gráfico pode resultar em gasto desnecessário de tinta se decidirmos imprimi-lo depois.

Temos que modificar a escala do eixo X, e o fundo cinza. Começaremos por este último, precisamos selecionar a área de plotagem do gráfico: ao colocarmos o cursor sobre o gráfico, sobre a parte cinza, e pressionando o botão esquerdo do mouse, vamos obter a tela mostrada na figura 11.

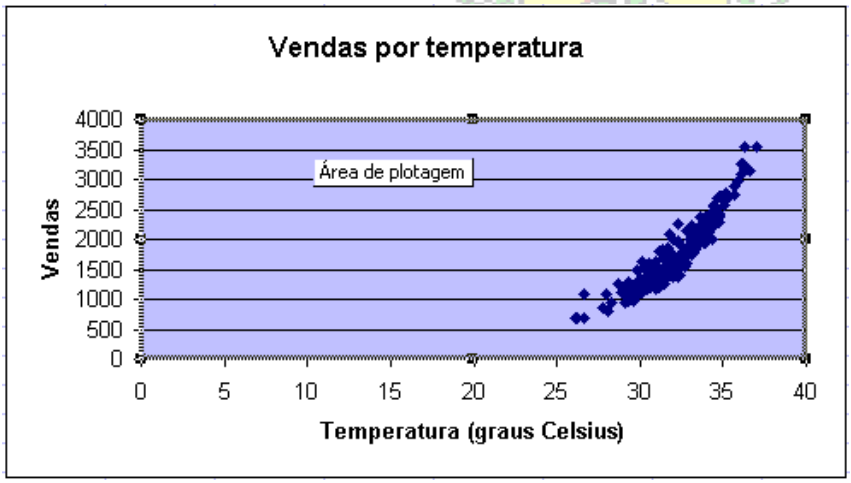


Figura 11 - Seleção da área de plotagem

Colocando o cursor sobre a área de plotagem, já selecionada, e pressionando o botão direito do mouse teremos a tela mostrada na figura 12, com as várias opções possíveis.

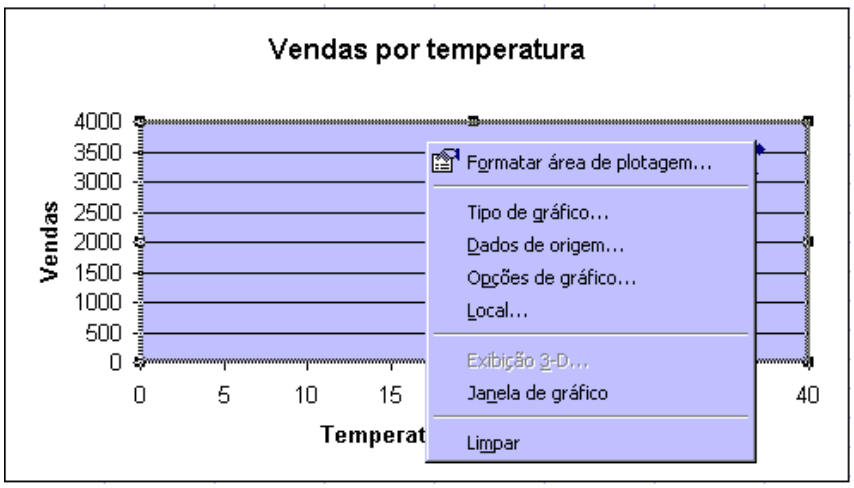


Figura 12 - Opções para a área de plotagem

Estamos interessados na primeira opção: "Formatar área de plotagem". Escolhendo esta opção o Excel apresentará a tela mostrada na figura 13.

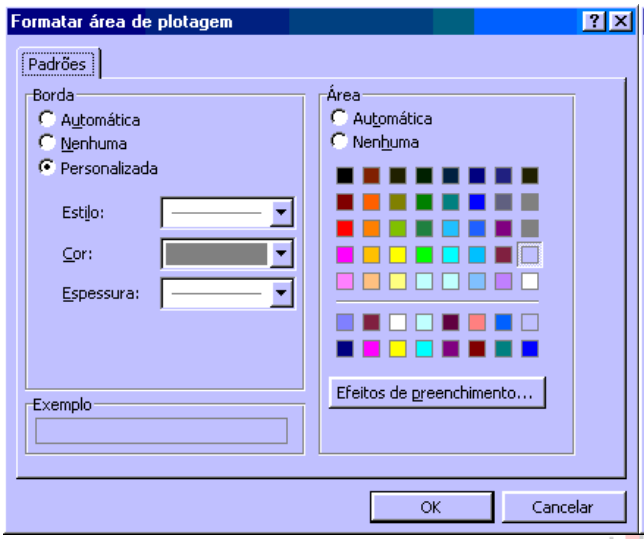


Figura 13 - Formatação padrão da área de plotagem

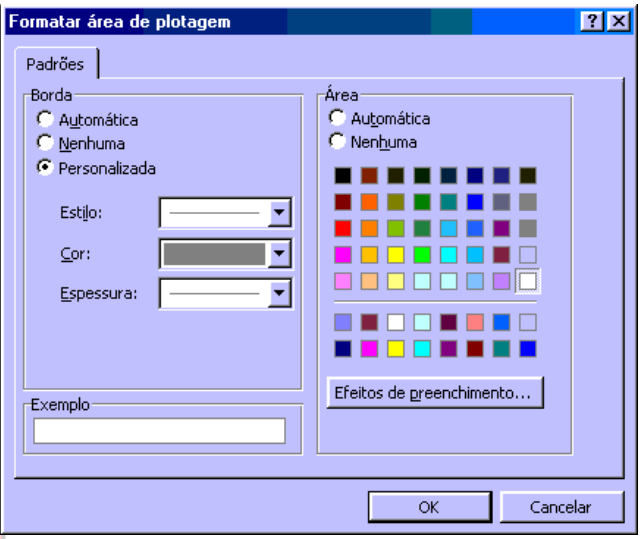


Figura 14 - Área de plotagem com fundo branco

Na figura 13 vemos a formatação padrão da área de plotagem, com fundo cinza: observe no campo "Área" que a cor cinza está selecionada, fazendo com que o campo "Exemplo" também tenha cor cinza. Na figura 14 selecionamos a cor branca, fazendo com que o campo "Exemplo" passe a ser branco também. Pressionando "OK" o gráfico passará a ser como o da figura 15.

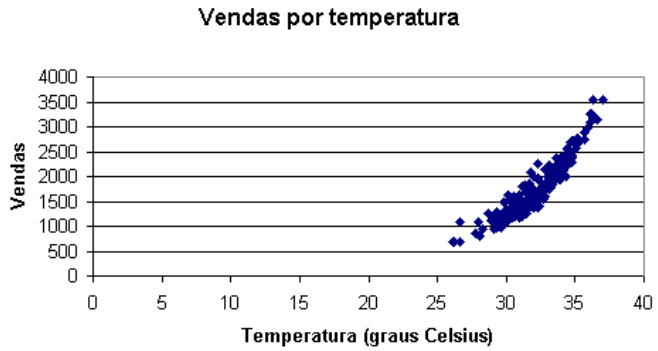


Figura 15 - Diagrama de dispersão com fundo branco

Resolvemos o problema do fundo, agora precisamos modificar a escala. Para tanto é preciso colocar o cursor exatamente sobre o eixo X, e pressionando o botão esquerdo do mouse teremos uma situação como a exposta na figura 16. Posteriormente, mantendo o cursor sobre o eixo e pressionando o botão direito do mouse vamos ter acesso às opções relativas ao eixo X, como mostrado na figura 17.

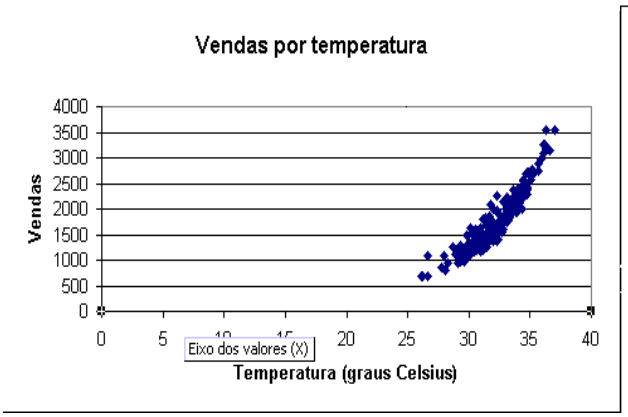


Figura 16 - Seleção do eixo X

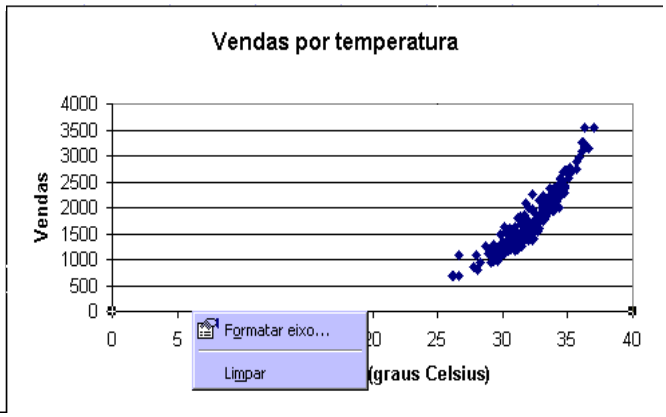


Figura 17 - Opções para o eixo X

Pressionando "Formatar eixo" vamos ter acesso a uma série de opções de modificação do eixo X, mostradas na figura 18.

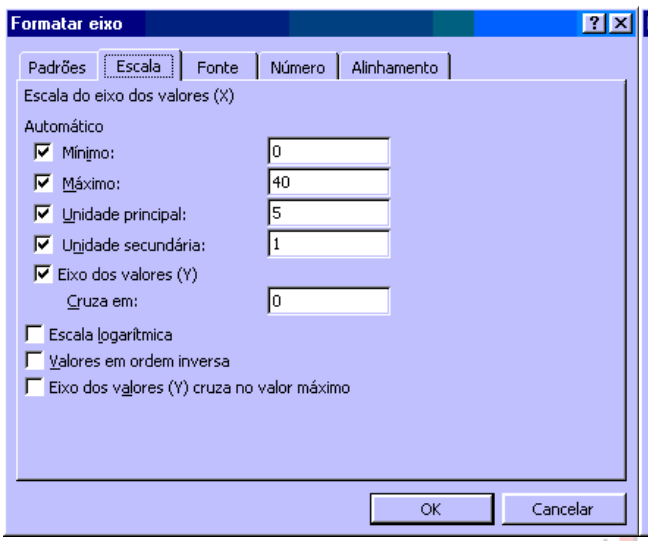


Figura 18 - Opções de formatação de eixo: escala

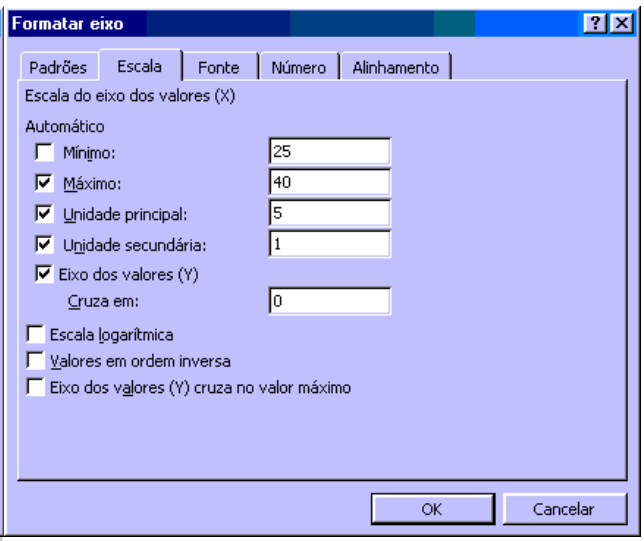


Figura 19 - Formatação de eixo: escala modificada

Escolhendo a opção "Escala" chegamos à figura 18. O comportamento padrão do Excel é construir a escala do gráfico com os valores mínimo e máximo encontrados nos dados. Mas algumas vezes, como no nosso problema, isso pode ser modificado, levando a um gráfico em que os dados estão muito concentrados. Como TODOS os valores de temperatura estão acima de 25 graus Celsius, vamos mudar o "Mínimo" da escala para 25, o que pode ser visto na figura 19. Pressionando "OK" vamos chegar ao gráfico mostrado na figura 20.

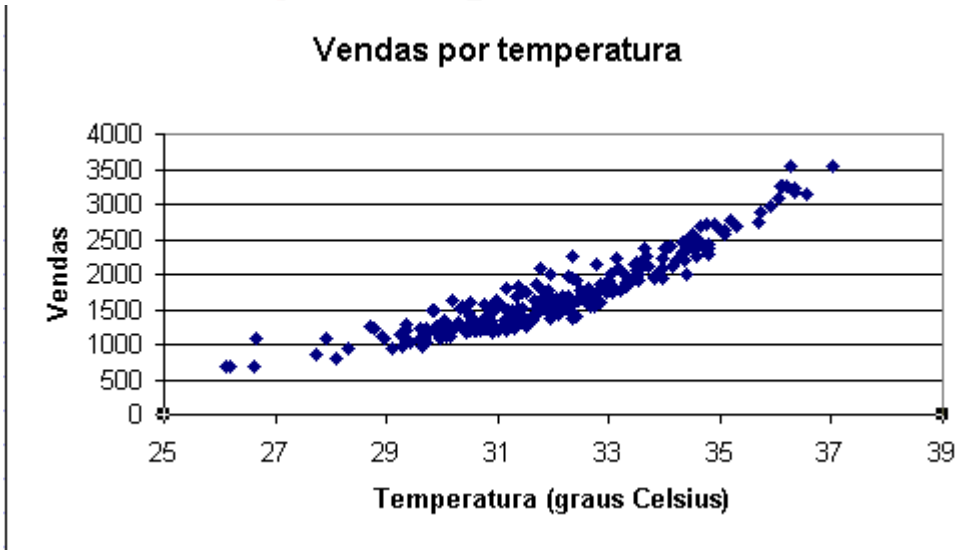


Figura 20 - Diagrama de dispersão vendas por temperatura - Final

Agora podemos fazer uma análise do diagrama de dispersão:

- as variáveis parecem estar fortemente correlacionadas, porque os pontos encontram-se bastante próximos.
- a correlação entre elas parece ser positiva, pois se observa que a nuvem de pontos tem um comportamento crescente, ou seja, maiores valores de temperatura, maiores valores de vendas (e é razoável imaginar que realmente um aumento na temperatura cause um aumento nas vendas).
- quanto à forma do relacionamento, isto é, que tipo de curva poderíamos ajustar aos dados para realização de previsões, talvez seja interessante pensar em um polinômio de segundo grau, ou uma exponencial; a utilização de uma reta talvez não seja uma boa idéia.

2. Ajuste de uma tendência a um diagrama de dispersão.

Imagine que quiséssemos ajustar uma reta ao diagrama de dispersão mostrado na figura 20, não obstante a análise feita. Como proceder? O Excel permite ajustar uma variedade de curvas aos dados mostrados em um diagrama de dispersão, e ainda calcula os coeficientes das equações das curvas, pelo método dos mínimos quadrados (ou seja, obtém os coeficientes minimizam a soma dos quadrados dos desvios entre os valores observados e os previstos por cada curva).

Para fazer o ajuste de qualquer curva, que no Excel significa adicionar uma linha de tendência, o primeiro passo é colocar o cursor sobre os pontos do gráfico e pressionar o botão esquerdo do mouse. Alguns pontos mudarão de cor, tal como mostrado na figura 21.

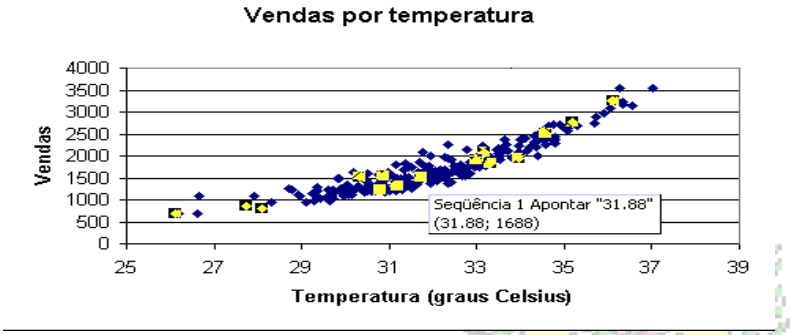


Figura 21 - Seleção de pontos no gráfico

Em seguida, mantendo o cursor sobre os pontos, precisamos pressionar o botão direito do mouse, e surgirão as opções possíveis para os dados, entre elas "Adicionar linha de tendência", tal como mostrado na figura 22.

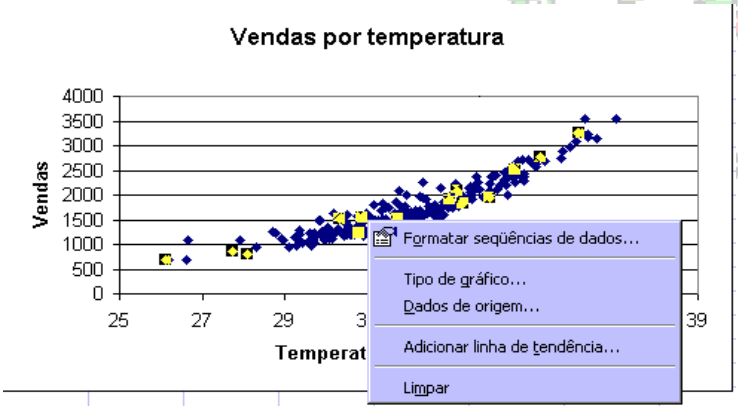


Figura 22 - Opções de modificação dos dados

Se pressionarmos "Adicionar linha de tendência" na figura 22 chegaremos à tela mostrada na figura 23. O tipo padrão de linha é a linear (reta), mas podemos selecionar outras. No nosso problema vamos manter a curva linear, mas queremos que o Excel exiba a equação e o valor de R-quadrado (coeficiente de determinação) no gráfico. Então, em "Opções" (figura 24) selecionamos ambos. Pressionando "OK" o gráfico ficará como o da figura 25.

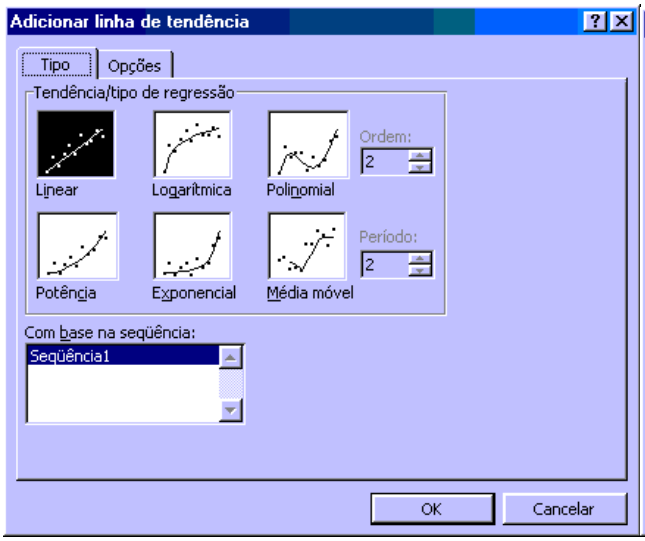


Figura 23 - Tipos de curva

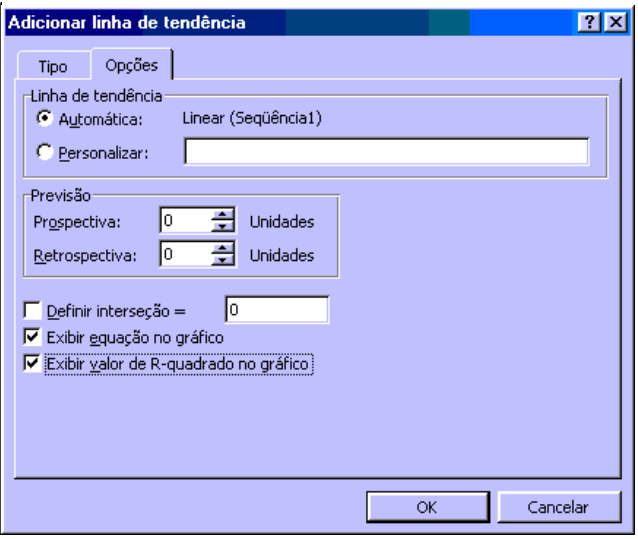
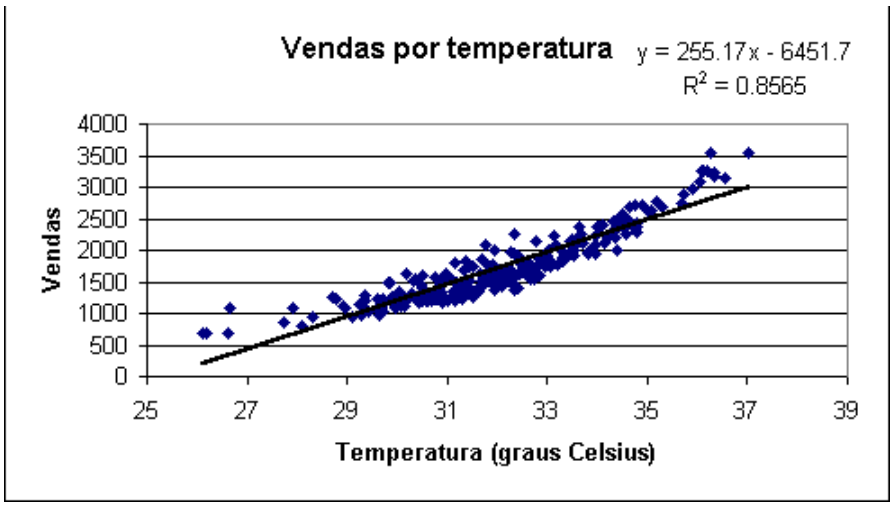


Figura 24 - Opções para os tipos de curva



Observe no canto superior direito da figura a equação da reta, com um coeficiente angular positivo (reta crescente), e o coeficiente de determinação, que vale 0,8565. Este valor significa que cerca de 85,65% da variabilidade média das vendas pode ser explicada pela variabilidade média da temperatura, através do modelo de regressão.

Figura 25 - Diagrama de dispersão com reta

Embora o valor de R^2 sugira que a reta é um bom modelo de regressão, devemos observar com cuidado o gráfico, e lembrar a análise feita na figura 20. Realmente a reta passa "entre" a maioria dos pontos, mas talvez outra curva apresente um melhor ajuste aos dados (polinômio de segundo grau ou exponencial, conforme sugerido anteriormente). Para realmente saber se o modelo ajustado é bom precisamos analisar seus resíduos.

3. Análise de resíduos

Uma vez tendo construído o diagrama de dispersão para as duas variáveis, e adicionado a linha de tendência a ele, pode ser interessante realizar a análise dos resíduos do modelo. Se o modelo for apropriado os resíduos deverão ter um comportamento aleatório, sem nenhum padrão identificável, mostrando que a variação residual, que não pode ser explicada pelo modelo é realmente casual, e ele poderá ser utilizado para realizar previsões e seus resultados serão úteis na tomada de decisão. Se, porém, algum padrão for detectado nos resíduos a variância residual não é aleatória, o que significa que o modelo não está conseguindo "explicar" de maneira consistente o relacionamento entre as variáveis, e, portanto, as previsões feitas pelo modelo são questionáveis. Isso pode acontecer mesmo que o R^2 assuma um valor elevado. Sendo assim a análise de resíduos é indispensável para avaliar a adequação de qualquer modelo de regressão, sendo especialmente importante nos casos de regressão múltipla, onde muitas vezes não é possível plotar um gráfico dos dados.

Pensando nos dados de Vendas e Temperatura, estudados nos itens 1 e 2, que culminaram no gráfico mostrado na figura 25, queremos analisar os resíduos do modelo linear (reta). O primeiro passo é calcular os valores de vendas previstos pelo modelo linear: na célula C2 da planilha inserimos a fórmula com a equação da reta obtida pelo Excel, tal como na figura 26.

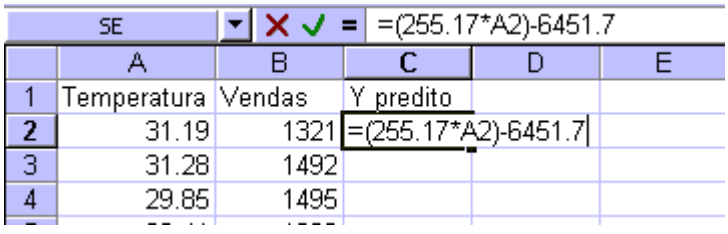
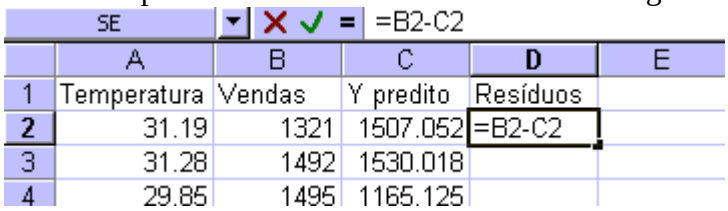


Figura 26 - Fórmula de previsão de vendas (reta)

Observe que a fórmula é construída em função da temperatura (cujo primeiro valor está na célula A2). Após digitar a fórmula e pressionar "Enter" (ou "Return", dependendo do computador), podemos colocar o cursor sobre a célula C2, selecionando-a.

Para estender os cálculos a todos os valores de temperatura basta "arrastar" a fórmula até a última linha do arquivo. As previsões de vendas através do modelo linear estarão então completas.

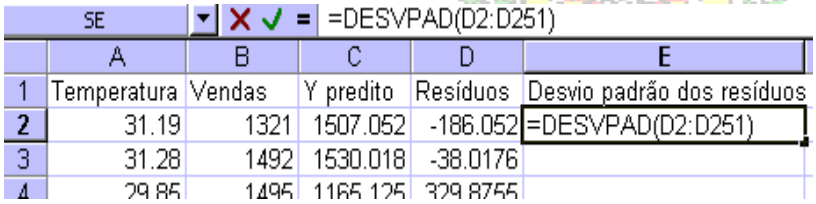
Para calcular os resíduos devemos obter a diferença entre os valores observados de Vendas e os valores previstos através do modelo linear. A figura 27 mostra isso.



Novamente, basta construir a fórmula para o primeiro valor e "arrastá-la" até a última linha para obter todos os resíduos do modelo.

Figura 27 - Cálculo dos resíduos

A obtenção dos resíduos é muito importante, mas dependendo da unidade das variáveis os resíduos poderão ser consideravelmente grandes em valores absolutos, embora em termos relativos sejam pequenos, ou o contrário. Podemos ter resíduos pequenos em termos absolutos, mas substancialmente grandes em relativos. Para que a análise seja feita objetivamente é preciso *padronizar* os resíduos: subtraí-los de sua média esperada (que deve ser igual a zero se o modelo for bom) e dividir pelo seu desvio padrão. O cálculo do desvio padrão dos resíduos está mostrado na figura 28.



Inserimos a fórmula do desvio padrão amostral, com os dados das células D2 a D251, que contêm os resíduos calculados anteriormente. O resultado está mostrado na figura 29.

Figura 28 - Cálculo do desvio padrão dos resíduos

Para obter os resíduos padronizados basta dividir cada resíduo pelo desvio padrão. Para que não haja problemas ao "arrastar" a fórmula é preciso dar uma referência absoluta ao denominador da fórmula: acrescentar $ antes da letra que designa a coluna e antes do número que designa linha, tal como na figura 29.

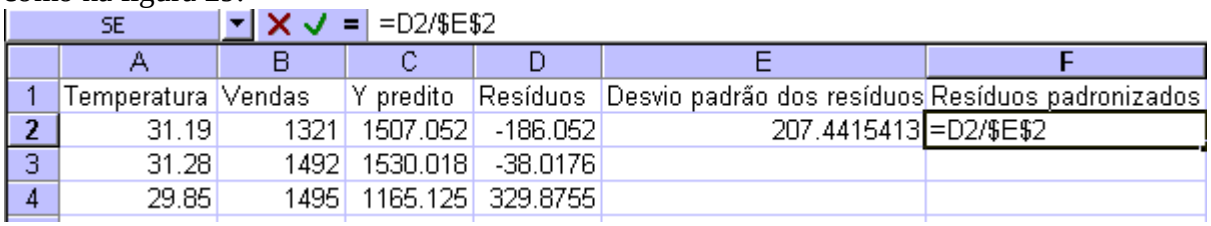


Figura 29 - Cálculo dos desvios padronizados

Para obter todos os resíduos basta "arrastar" a fórmula até a última linha do arquivo.

Uma vez obtidos os resíduos padronizados podemos fazer a sua análise propriamente dita. Precisamos construir dois diagramas de dispersão dos resíduos: resíduos padronizados em função de X (Temperatura), e resíduos padronizados em função dos valores preditos. O procedimento é semelhante ao visto no item 1, mudando apenas os valores de X e de Y, e escrevendo os títulos adequados, o que é mostrado nas figuras 30 e 31.

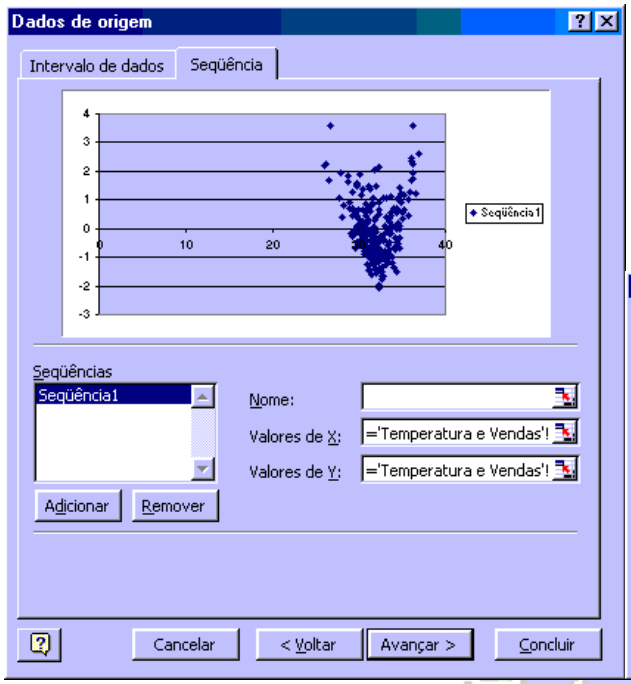


Figura 30 - Dados para o análise de resíduos

Observe a escala do diagrama. Novamente precisamos modificá-la, bem como o fundo cinza. Devemos fazer o mesmo procedimento também para o diagrama dos resíduos padronizados pelos valores preditos. Os diagramas resultantes estão nas figuras 32 e 33.

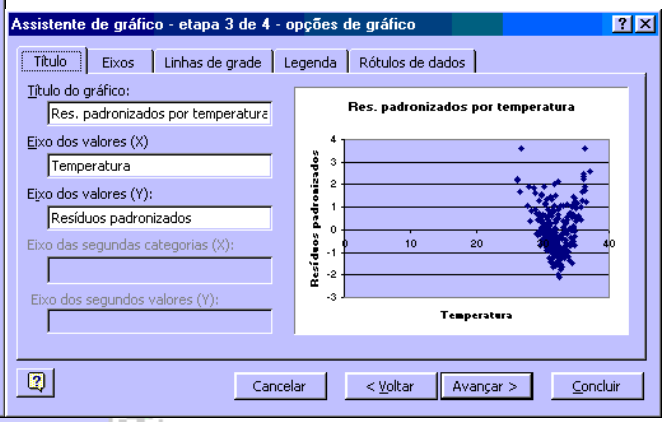


Figura 31 - Títulos do diagrama de dispersão

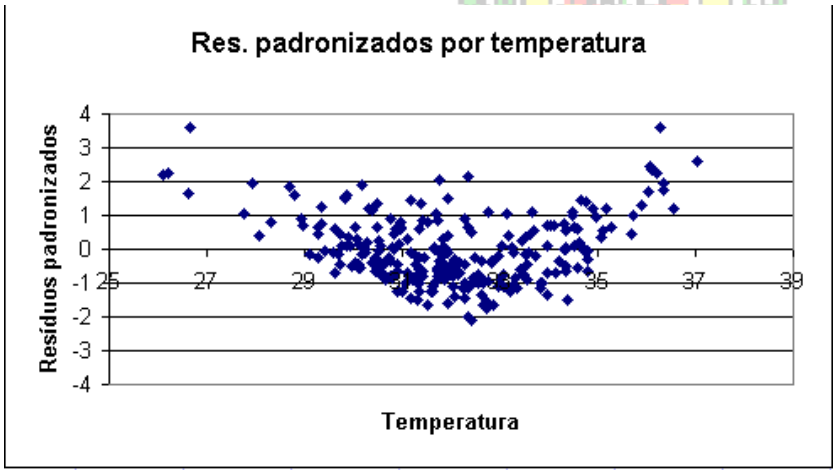


Figura 32 - Resíduos padronizados por temperatura - Modelo linear

Fazendo a análise dos resíduos mostrados na figura 32.
Observe a escala vertical do gráfico: devemos sempre torná-la simétrica ao zero, para auxiliar na análise:
1) Número de resíduos positivos é semelhante ao dos negativos.
2) As distâncias dos resíduos positivos a zero são maiores do que as dos negativos.
3) Há um padrão nos resíduos, parece uma parábola.

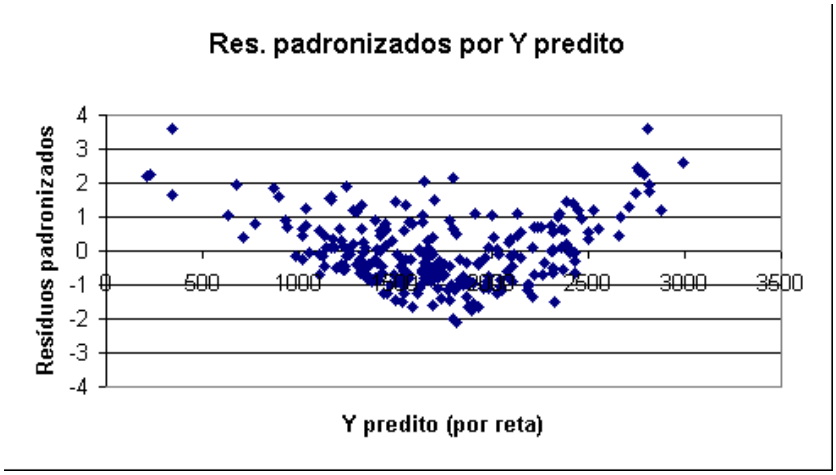


Figura 33 - Resíduos padronizados por valores previstos - Modelo linear

Fazendo a análise dos resíduos mostrados na figura 33.
1) Número de resíduos positivos é semelhante ao dos negativos.
2) As distâncias dos resíduos positivos a zero são maiores do que as dos negativos.
3) Há um padrão nos resíduos, parece uma parábola.
Juntando a análise dos dois diagramas chegamos à conclusão que o modelo linear NÃO é apropriado para o problema, pois seus resíduos não se comportam de forma aleatória.

Sugerimos a utilização de outro modelo.

Repetindo o procedimento das Figuras 21 a 23 podemos escolher o modelo Polinômio do 2º grau. O resultado pode ser visto na Figura 34, superposto ao resultado da Figura 25.

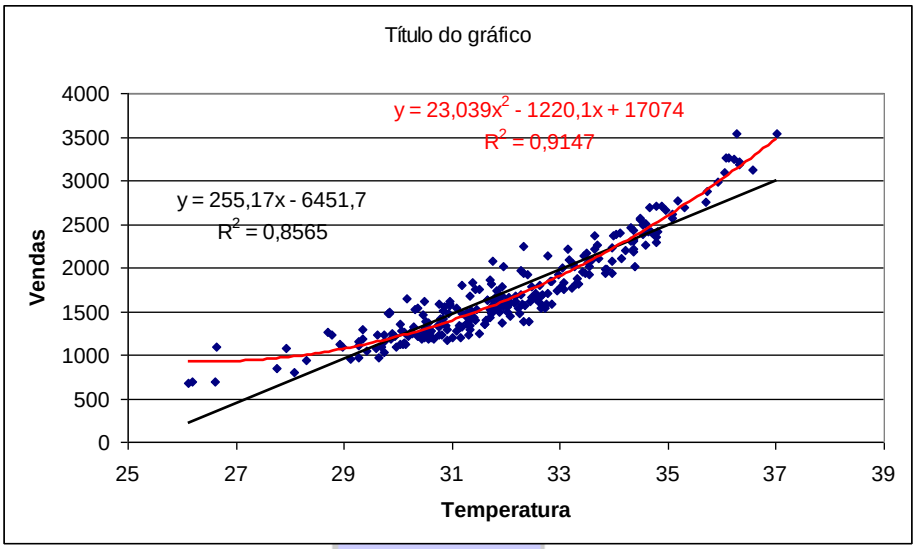


Figura 34 - Diagrama de dispersão com reta e polinômio do 2º grau

Percebe-se que o coeficiente de determinação do polinômio de 2º grau é maior do que o da reta. E, também, o ajuste da curva do polinômio de 2º grau aos pontos é bem melhor. Provavelmente os resíduos serão melhores do que os da reta. Outros modelos poderiam ser ajustados, resultando na Figura 35.

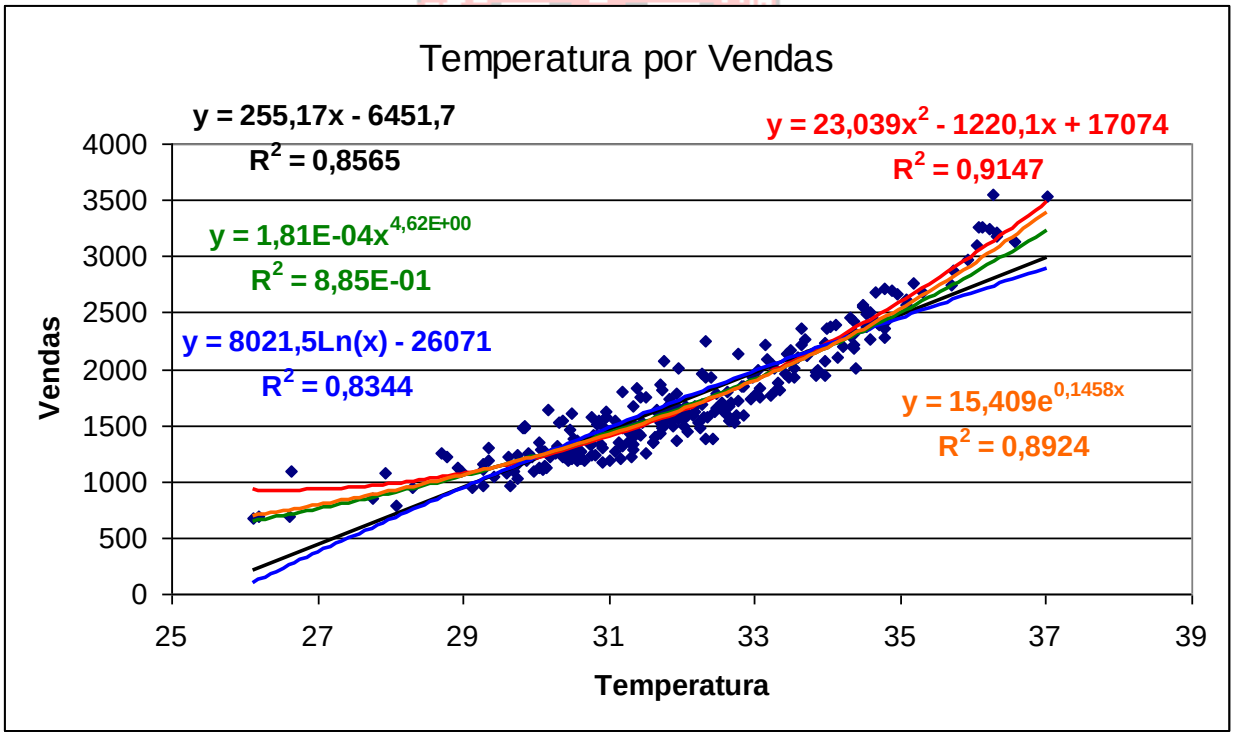


Figura 35 - Diagrama de dispersão com cinco modelos de regressão

Todos os cinco modelos aplicáveis estão no gráfico da Figura 35: reta, polinômio de 2º grau, logarítmico, exponencial e potência. Mas, observe o formato dos coeficientes no modelo potência: está científico, $1{,}81E\text{-}04x^{4{,}62E+00}$. Isso significa $0{,}000181x^{4{,}62}$, que é o formato que devemos usar nas previsões. Às vezes o Excel automaticamente apresenta as equações de um modelo em formato científico, e com um número insuficiente de casas decimais, o que pode prejudicar nossas previsões. Para mudar o formato e as casas decimais veja o procedimento a seguir.

Selecione a equação do modelo potência na Figura 35:

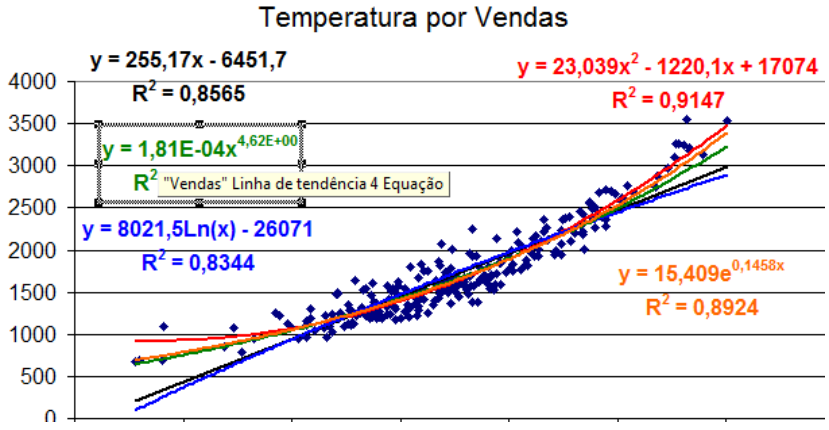


Figura 36 - Seleção de uma equação

Clicando duas vezes sobre a equação surge a tela da Figura 37.

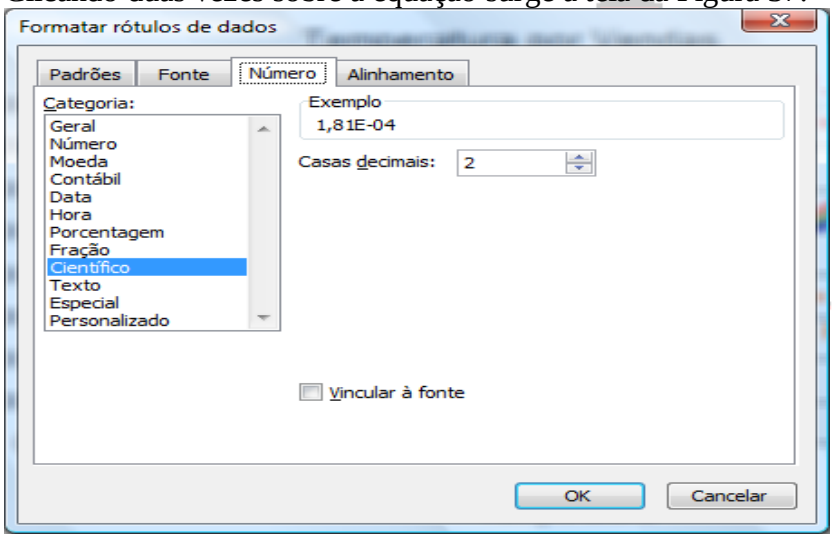


Figura 37 – Formatação de rótulo de dados: Número

Às vezes o Excel apresenta os dados em formato científico, mas na categoria "Geral". Se quisermos que os números sejam apresentados da forma usual devemos escolher "Número" e quantas casas decimais forem necessárias: no nosso caso, como o Excel usou E-04, deve-se escolher no mínimo 4, mas o ideal é um pouco mais para ganhar precisão nas previsões, 6, por exemplo. O resultado pode ser visto na Figura 38.

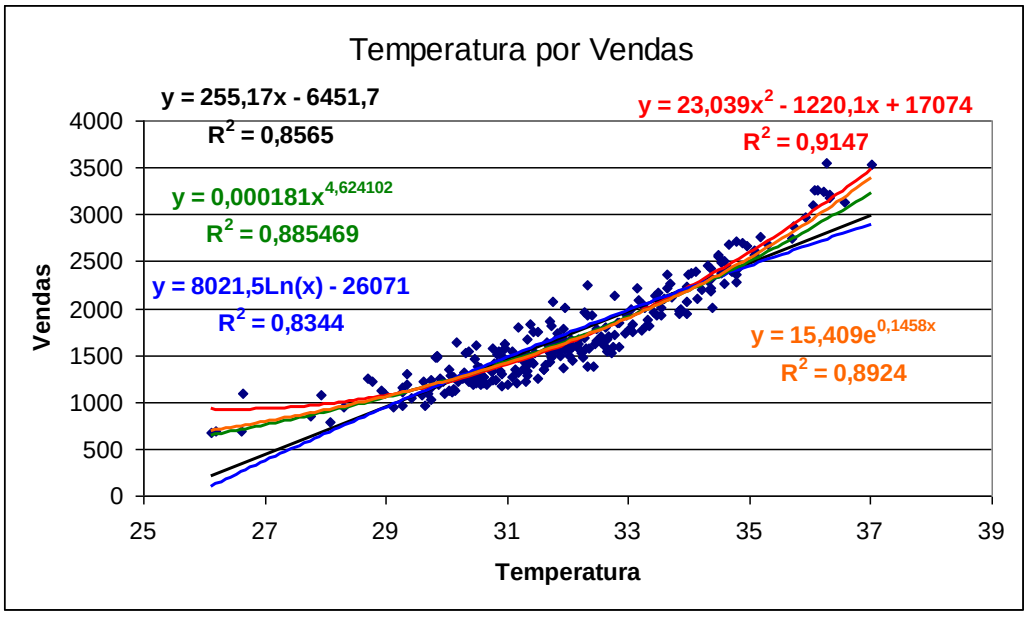


Figura 38 - Diagrama de dispersão com cinco modelos de regressão - modificado

Na Figura 26 fizemos a previsão usando o modelo de Reta, agora apresentaremos as previsões pelos outros modelos disponíveis:

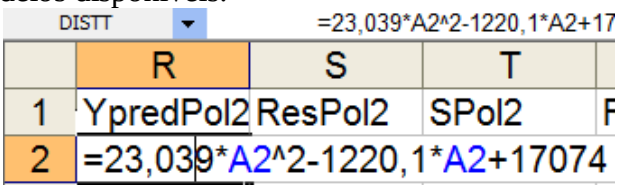


Figura 39 - Modelo polinômio de 2o grau (para equação da Figura 38)

Na Figura 39 é possível observar que no lugar de X colocamos a primeira célula do intervalo que contém os valores de temperatura (célula A2). Observe que o ^ é o símbolo de potenciação no Excel (e no Calc também). Basta arrastar até a célula R251 para completar a previsão pelo modelo polinômio de 2º grau. O cálculo dos resíduos, desvio padrão dos resíduos e resíduos padronizados é análogo ao caso da reta (para este e para os próximos modelos).

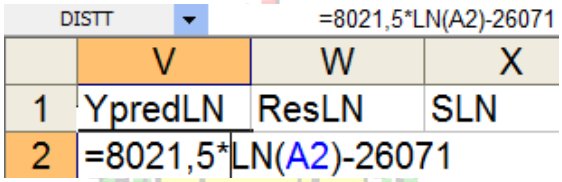


Figura 40 - Modelo logarítmico (para equação da Figura 38)

Na Figura 40 é possível observar que no lugar de X colocamos a primeira célula do intervalo que contém os valores de temperatura (célula A2). Observe que LN() é uma função do Excel (e do Calc também) que permite calcular o logaritmo neperiano (com base igual a e, a constante de Neper, igual a 2, 71828...). Basta arrastar até a célula V251 para completar a previsão pelo modelo logarítmico.

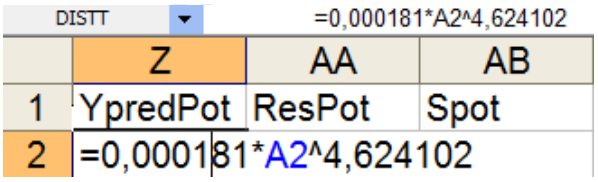


Figura 41 - Modelo potência (para equação da Figura 38)

Na Figura 41 é possível observar que no lugar de X colocamos a primeira célula do intervalo que contém os valores de temperatura (célula A2). Observe que X (no caso o conteúdo da célula A2) é elevado (^) a 4,624102, que é expoente do modelo potência (ver Figura 38). Basta arrastar até a célula Z251 para completar a previsão pelo modelo potência.

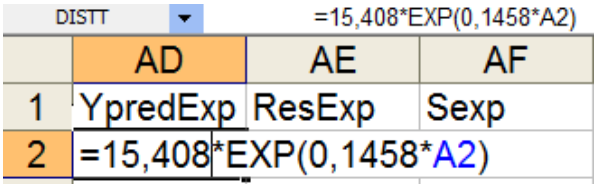


Figura 42 - Modelo exponencial (para equação da Figura 38)

Na Figura 42 é possível observar que no lugar de X colocamos a primeira célula do intervalo que contém os valores de temperatura (célula A2). Observe que EXP() é uma função do Excel (e do Calc também) que permite calcular o valor da constante de Neper (e = 2, 71828...) elevada ao produto de 0,1458 pelo conteúdo da célula A2). Basta arrastar até a célula AD251 para completar a previsão pelo modelo exponencial.

Vejam os resíduos padronizados do modelo polinômio do 2º grau:

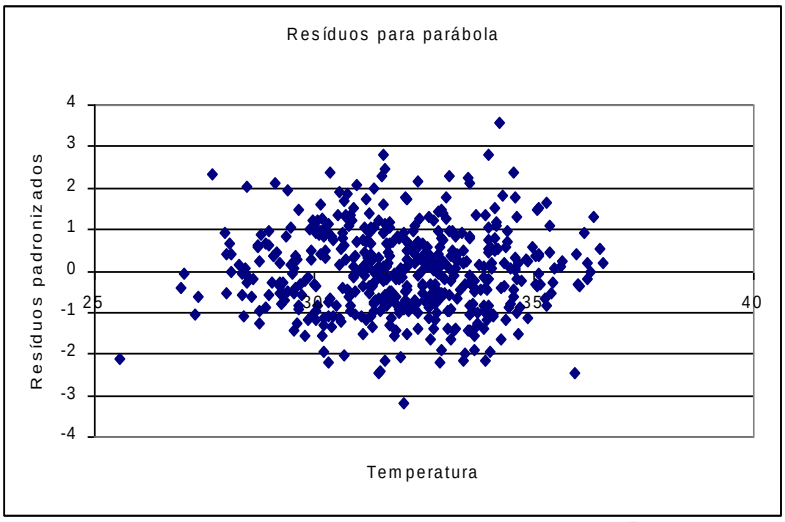


Figura 43 - Resíduos do polinômio de 2º grau por temperatura

Fazendo a análise dos resíduos mostrados na Figura 43.
1) Número de resíduos positivos é semelhante ao dos negativos.
2) As distâncias dos resíduos positivos e negativos a zero são semelhantes.
3) Os resíduos distribuem-se aleatoriamente, sem padrão.

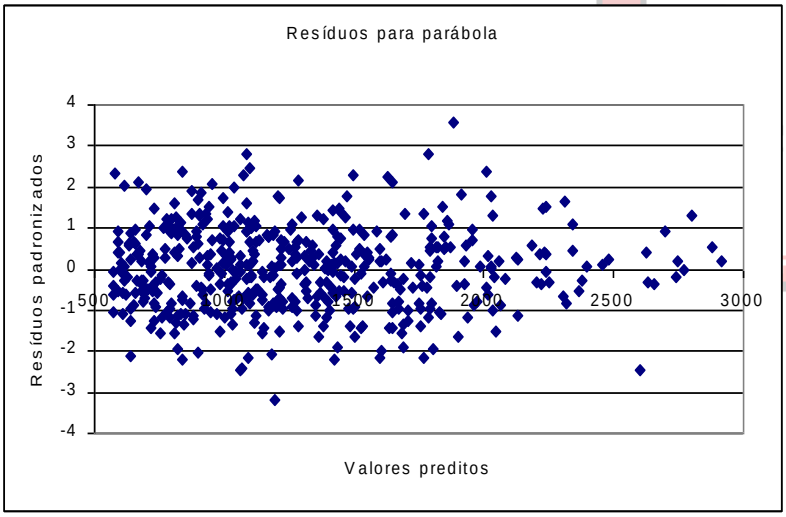


Figura 44 - Resíduos do polinômio do 2º grau por valores preditos

Fazendo a análise dos resíduos mostrados na Figura 44.
1) Número de resíduos positivos é semelhante ao dos negativos.
2) As distâncias dos resíduos positivos e negativos a zero são semelhantes.
3) Os resíduos distribuem-se aleatoriamente, sem padrão.
Juntando a análise dos dois diagramas chegamos à conclusão que o modelo de polinômio de 2º grau é apropriado para o problema, pois seus resíduos se comportam de forma aleatória.