Métodos Estocásticos da Engenharia II

Capítulo 7 - Análise de Correlação e Regressão Simples

Prof. Magno Silvério Campos

2024/2

Bibliografia

Essas notas de aulas foram baseadas nas seguintes obras:

1. BORNIA, A. C.; BARBETTA, P. A.; REIS, M. M. *Estatística para Cursos de Engenharia e Informática*. 2. ed. São Paulo: Atlas, 2009.
2. CANCHO, V.G. *Notas de Aulas sobre Noções de Estatística e Probabilidade*. São Paulo: USP, 2010.
3. HINES, W.W.; *et al. Probabilidade e Estatística na Engenharia*. 4. ed. Rio de Janeiro: LTC, 2006.
4. MONTGOMERY, D.C.; RUNGER, G.C. *Estatística Aplicada e Probabilidade para Engenheiros*. 6. ed. Rio de Janeiro: LTC, 2016.

Aconselha-se pesquisá-las para se obter um maior aprofundamento e um melhor aproveitamento nos estudos.

Conteúdo Programático

1. Seção 1 - Modelo de Regressão Linear Simples (MRLS)
   - Introdução
   - Modelo de regressão linear simples
   - Estimativas dos parâmetros do MRLS e suas propriedades
   - Testes de hipóteses na regressão linear simples
   - Intervalos de confiança
   - Previsão de novas observações
   - Estudo da adequação do modelo de regressão
2. Seção 2 - Análise de Correlação

Introdução

Muitos problemas em engenharia e ciências envolvem explorar as relações entre duas ou mais variáveis. A análise de regressão é uma técnica estatística para modelar e investigar a relação entre duas ou mais variáveis.

Exemplo

Em um processo químico, suponha que o rendimento do produto esteja relacionado à temperatura de operação do processo. A análise de regressão pode ser usada para construir um modelo para prever o rendimento em um dado nível de temperatura. Esse modelo pode também ser usado para otimização de processos, tal como encontrar o nível de temperatura que maximiza o rendimento, ou para finalidades de controlar o processo.

Sendo assim, a análise de regressão ocupa-se do estudo da dependência de uma variável, a variável dependente (ou variável resposta), em relação a uma ou mais variáveis, as variáveis explicativas (ou variáveis independentes, ou regressoras), com o objetivo de estimar e/ou prever o valor médio da variável dependente em termo dos valores conhecidos ou fixos das variáveis explicativas.

O estudo da análise de regressão será iniciado considerando o exemplo a seguir.

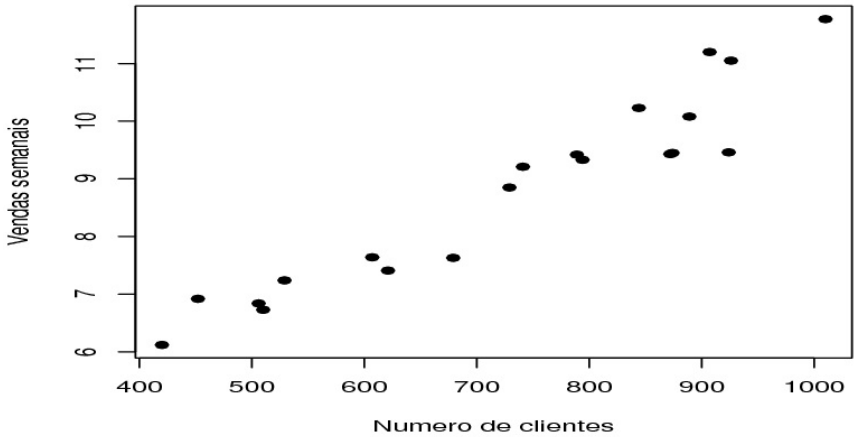
Exemplo - [Cancho(2010)]

Um administrador de uma rede de supermercados deseja desenvolver um modelo com a finalidade de estimar as vendas médias semanais (em milhares de dólares) de cada supermercado da rede. Para isso, selecionou-se uma amostra aleatória de 20 supermercados entre todos os que formam a rede. Ao se desenvolver o modelo, foi considerada, entre outras variáveis explicativas (ou independentes), a variável " número de clientes por semana". Os dados são apresentados na tabela 1:

Supermercado	N^o de clientes (X)	Vendas semanais (Y)
1	907	11,20
2	926	11,05
3	506	6,84
4	741	9,21
5	789	9,42
6	889	10,08
7	874	9,45
8	510	6,73
9	529	7,24
10	420	6,12
11	679	7,63
12	872	9,43
13	924	9,46
14	607	7,64
15	452	6,92
16	729	8,95
17	794	9,33
18	844	10,23
19	1010	11,77
20	621	7,41

Tabela: Dados para uma amostra de 20 supermercados.

Na figura abaixo, é apresentado o **diagrama de dispersão** das vendas semanais *versus* o número de clientes.



Fonte: [Cancho(2010)]

A análise desse diagrama indica que uma curva não passa exatamente por todos os pontos, mas existe uma forte evidência que os pontos estão dispersos de maneira aleatória em torno de uma linha reta. Portanto, é razoável supor que a média da variável aleatória Y, está relacionada com X pela seguinte relação

$$E(Y|X = x) = \mu_{Y|x} = \beta_0 + \beta_1 x$$

onde β_o e β_1, são respectivamente, o intercepto e a inclinação da reta e recebem o nome de coeficientes de regressão.

Mesmo que a média de Y seja uma função linear de X, o valor observado de y não cai de maneira exata sobre a reta.

A maneira apropriada para generalizar este fato como um modelo probabilístico linear, é supor que o valor esperado de Y seja uma função linear, mas, para um valor fixo de X o valor real de Y será determinado pelo valor médio da função linear ($\mu_{Y|x}$) mais um termo que representa um erro aleatório, assim:

$$Y = \mu_{Y|x} + \varepsilon = \beta_0 + \beta_1 x + \varepsilon, \tag{1}$$

onde ε é o erro aleatório. É importante observar que ε leva em conta a falha desse modelo em se ajustar exatamente aos dados. Isso pode ser devido ao efeito de outras variáveis que afetam as vendas semanais. O modelo (1) recebe o nome de **modelo de regressão linear simples**, pois tem somente uma variável explicativa ou variável regressora ou variável independente .

Modelo de Regressão Linear Simples

Como é mostrado na equação (1), os erros considerados no MRLS incidem diretamente sobre os valores observados de Y. A teoria da regressão se assenta nas seguintes suposições:

1. Os erros têm média zero e a mesma variância desconhecida, σ^2.
2. Os erros são não correlacionados, ou seja, o valor de um erro não depende de qualquer outro erro.
3. A variável explicativa X é controlada pelo experimentador e é medida sem erro, ou seja, não é uma variável aleatória.
4. Os erros tem distribuição normal.

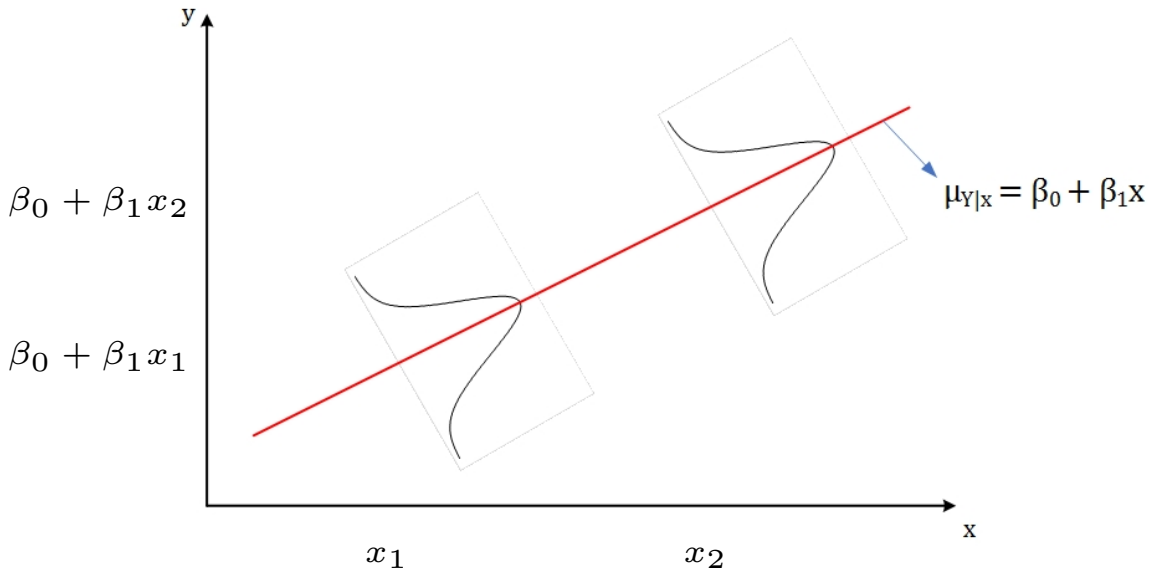
A partir das suposições 1, 2 e 4 acima, podemos escrever:

$$\varepsilon \sim NID(0, \sigma^2)$$

Se as suposições de 1 a 4 se verificarem, então:

- $E(Y|X=x) = E(\beta_0 + \beta_1 x + \varepsilon) = \beta_0 + \beta_1 x + E(\varepsilon) = \beta_0 + \beta_1 x$
- $V(Y|X=x) = V(\beta_0 + \beta_1 x + \varepsilon) = V(\beta_0 + \beta_1 x) + V(\varepsilon) = 0 + \sigma^2 = \sigma^2$
- $Y \sim N(\beta_0 + \beta_1 x; \sigma^2)$

Logo, o modelo verdadeiro de regressão, $\mu_{Y|x} = \beta_0 + \beta_1 x$, é uma linha de valores médios. Além disso, há uma distribuição de valores de Y em cada x e a variância dessa distribuição é a mesma em cada x.



Observe em

$$E(Y|X = x) = \mu_{Y|x} = \beta_0 + \beta_1 x.$$

que:

- para um acréscimo de uma unidade em X há um acréscimo de β_1 unidades na média de Y;
- se os valores de X incluem $X = 0$, então o intercepto β_0 é a média de Y quando $X = 0$. Em caso contrário, β_0 não tem interpretação prática.

Estimativas dos parâmetros do MRLS

As estimativas de mínimos quadrados da interseção e da inclinação do modelo de regressão linear simples são

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}. \tag{2}$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i y_i - \frac{\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n}}{\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}}. \tag{3}$$

onde $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$ e $\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n}$.

Considerando que:

$$
\begin{aligned}
S_{xx} &= \sum_{i=1}^{n}(x_i-\bar{x})^2 = \sum_{i=1}^{n} x_i^2 - n\bar{x}^2, \\
S_{yy} &= \sum_{i=1}^{n}(y_i-\bar{y})^2 = \sum_{i=1}^{n} y_i^2 - n\bar{y}^2, \\
S_{xy} &= \sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y}) = \sum_{i=1}^{n}(x_i-\bar{x})y_i = \sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}.
\end{aligned}
$$

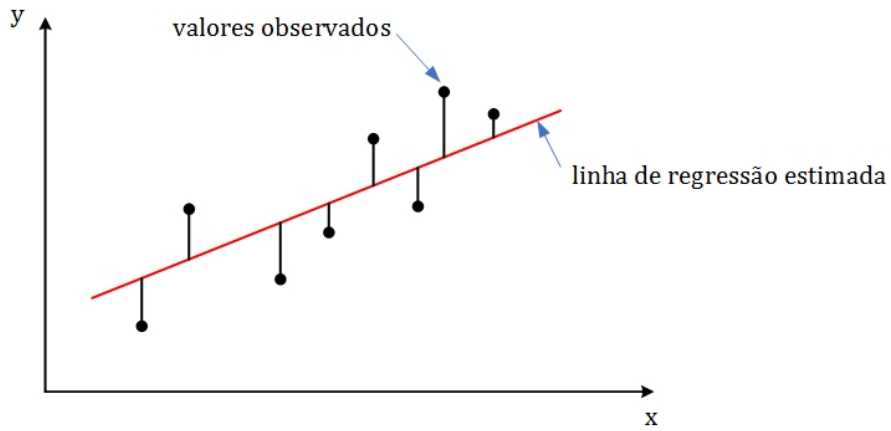
Podemos escrever:

$$
\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x}. \tag{4}
$$

$$
\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}}. \tag{5}
$$

Portanto, a linha de regressão estimada ou ajustada é $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$.

A figura a seguir mostra os desvios dos dados em relação ao modelo estimado de regressão:



Note que para cada par de observações $(x_i,\ y_i)$, satisfaz-se a relação:

$$y_i = \underbrace{\hat{\beta}_0 + \hat{\beta}_1 x_i}_{\hat{y}_i} + e_i, \quad i = 1, \ldots, n$$

onde $e_i = y_i - \hat{y}_i$ é chamado de resíduo e, descreve o erro no ajuste do modelo na i-ésima observação.

Exemplo

Considere os dados do exemplo, apresentado no início desta seção, no qual o gerente da rede de supermercados estava interessado em estimar as vendas médias semanais de cada supermercado, dado o número de clientes por supermercado.

Para se determinar o modelo de regressão estimado, foram calculadas as seguintes quantidades:

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

$$\sum_{i=1}^{n} y_i^2 = (11,20)^2 + (11,05)^2 + \cdots + (7,41)^2 = 1.602,0971$$

$$\sum_{i=1}^{n} x_i y_i = (907)(11,20) + (11,05)(926) \cdots + (7,41)(621) = 134.127,90$$

$$S_{xx} = \sum_{i=1}^{n} x_i^2 - n(\bar{x})^2 = 11.306.209 - 20(731,15)^2 = 614.603$$

$$S_{xy} = \sum_{i=1}^{n} x_i y_i - n(\bar{x})(\bar{y}) = 134.127,90 - 20(8,8055)(731,15) = 5.36$$

$$S_{yy} = \sum_{i=1}^{n} y_i^2 - n(\bar{y})^2 = 1.609,0971 - 20(8,8055) = 51,3605.$$

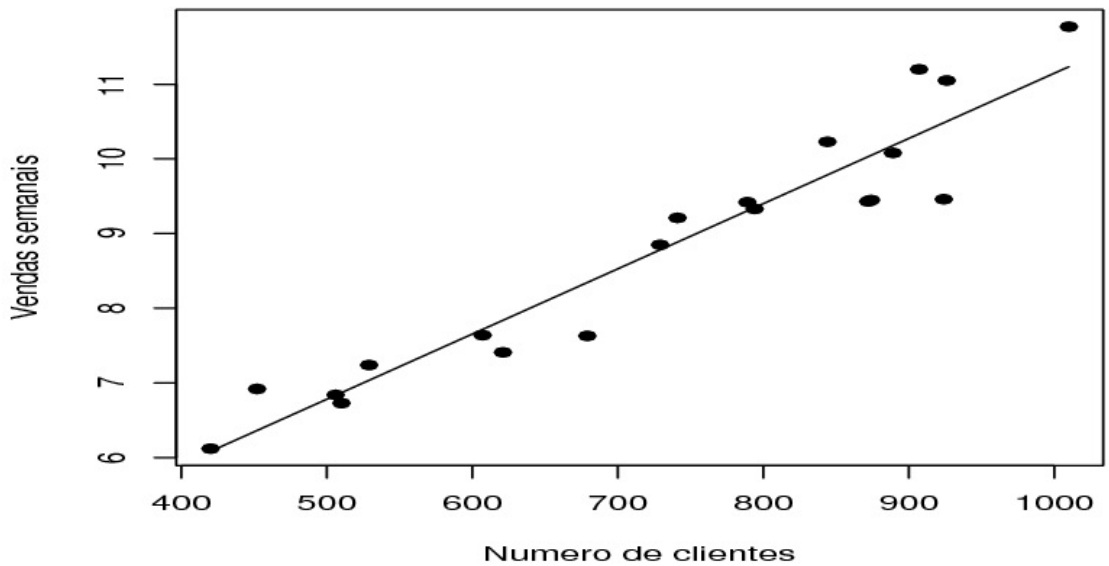
Os EMQ dos parâmetros do MRLS são:

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

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

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

$$\hat{y} = 2,423 + 0,00873x. \tag{6}$$



Fonte: [Cancho(2010)]

Comentários

A estimativa do coeficiente de regressão $\hat{\beta}_1$ foi 0,00873. Isto significa que, para cada incremento de uma unidade de X, estimamos que o valor da média de Y aumenta em 0,00873 unidades. Isto é, para cada incremento de um cliente, o modelo prevê uma estimação de um aumento nas vendas de 0,00873 mil dólares (ou 8,73 dólares). Portanto, para um incremento de 100 clientes, esperamos que as vendas semanais aumentem, em média \$ 873 dólares.

A estimativa do intercepto $\hat{\beta}_0$ foi de 2,423 mil dólares. Essa estimativa representa o valor médio Y, quando $X = 0$. Como é improvável que o número de clientes seja zero, esse valor pode ser visto como a proporção média das vendas semanais que variam em relação a fatores diferentes ao número de clientes.

Se o modelo de regressão ajustado aos dados (6) for aceitável, pode ser usado para prever os valores futuros da venda semanal.

Por exemplo, suponha que tem-se interesse em prever as vendas semanais para um supermercado com 600 clientes. No modelo de regressão ajustado em (6), é feito $X = 600$ e tem-se:

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

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, ou como uma estimação de uma futura venda de um supermercado quando o número de clientes for $X = 600$.

Abusos da Regressão

1. É possível desenvolver relações estatísticas entre as variáveis que não estejam completamente relacionadas em um sentido prático. Por exemplo, relacionar a **tensão cisalhante em pontos de solda** com o **número de espaços vazios em um estacionamento de veículos**!
2. Relações de regressão são válidas somente para valores do regressor dentro (ou próximo) da faixa dos dados originais. A extrapolação não necessariamente valida os modelos de regressão.

Propriedades dos EMQ de β_0 e β_1

Considerando que as suposições do modelo de regressão sejam válidas é possível demonstrar as seguintes propriedades:

$$E(\hat{\beta}_1) = \beta_1 \tag{7}$$

$$Var(\hat{\beta}_1) = \frac{\sigma^2}{S_{xx}}. \tag{8}$$

$$E(\hat{\beta}_0) = \beta_0 \tag{9}$$

$$Var(\hat{\beta}_0) = \sigma^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right]. \tag{10}$$

$$\hat{\beta}_j \sim N(\beta_j,\ Var(\beta_j)),\ \ j = 0, 1. \tag{11}$$

Estimação de σ^2

Para realizarmos inferências com relação aos parâmetros do MRLS β_0 e β_1, é necessário estimar o parâmetro σ^2 que aparece nas expressões de $Var(\hat{\beta}_0)$ e $Var(\hat{\beta}_1)$. O parâmetro σ^2, que é a variância do termo aleatório ε no MRLS, reflete a variação aleatória ao redor da verdadeira linha de regressão.

A soma de quadrados residuais ou soma de quadrados dos erros, denotado por SQR é:

$$SQR = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = S_{yy} - \hat{\beta}_1 S_{xy}.$$

Um estimador de σ^2 é dado por

$$\hat{\sigma^2} = \frac{SQR}{n-2} \tag{12}$$

Exemplo

Com os dados do exemplo desta seção, é feita a estimação da variância σ^2. Nesse caso, $S_{yy} = 51,3605$, $S_{xy} = 5.365,08$ e $\hat{\beta}_1 = 0,00873$. Portanto, da equação (12),

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

Definição

No modelo de regressão linear simples, o **erro padrão estimado** da inclinação é dado por:

$$EP(\hat{\beta}_1) = \sqrt{\frac{\hat{\sigma}^2}{S_{xx}}}$$

e o erro padrão do intercepto é dado por:

$$EP(\hat{\beta}_0) = \sqrt{\hat{\sigma}^2 \left[\frac{1}{n} + \frac{\bar{X}^2}{S_{xx}}\right]}$$

onde $\hat{\sigma}^2$ é calculada com a equação (12).

Testes de Hipóteses na Regressão Linear Simples

Introdução

Uma parte importante ao avaliar a adequação de um MRLS é o teste de hipóteses sobre os parâmetros do modelo e a construção de certos intervalos de confiança.

** Observação **

Para realizar testes é necessário que a suposição de que os erros sejam independentes e identicamente distribuídos segundo uma curva normal com média zero e variância σ^2 ($\varepsilon_i \sim NID(0, \sigma^2)$) seja válida.

Teste de hipóteses sobre β_1

Suponha que se deseje testar a hipótese de que a inclinação é igual a uma constante representada por $\beta_{1,0}$. As hipóteses apropriadas são:

$$\begin{aligned} H_0 : \beta_1 &= \beta_{1,0} \\ H_1 : \beta_1 &\neq \beta_{1,0} \end{aligned} \tag{13}$$

onde é considerada uma alternativa bilateral.

Estatística de teste

$$T = \frac{\hat{\beta}_1 - \beta_{1,0}}{\sqrt{\hat{\sigma}^2 / S_{xx}}} \sim t_{n-2} \tag{14}$$

se H_0 é verdadeira.

Critério de Rejeição

Rejeita-se H_0 se

$$|T_{obs}| > t_{\alpha/2,\, n-2}.$$

Um procedimento similar pode ser utilizado para testar hipóteses sobre o intercepto. Para testar

$$\begin{aligned} H_0 : \beta_0 &= \beta_{0,0} \\ H_1 : \beta_0 &\neq \beta_{0,0} \end{aligned} \tag{15}$$

usamos a estatística

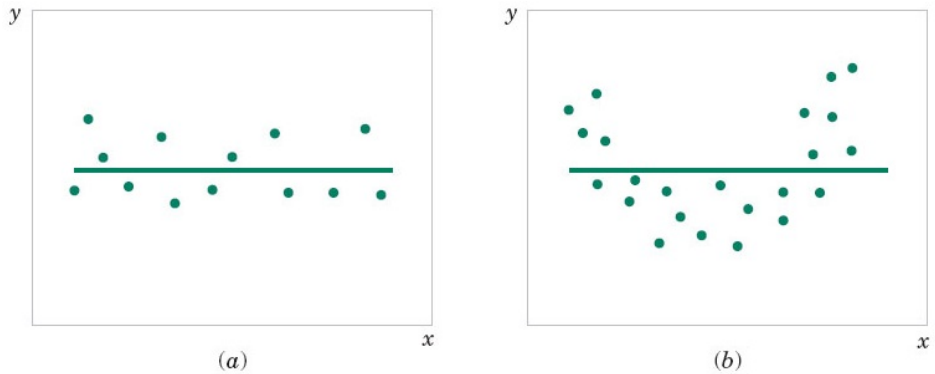
$$T = \frac{\hat{\beta}_0 - \beta_{0,0}}{\sqrt{\hat{\sigma}^2[\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}]}} \tag{16}$$

que tem distribuição t-Student com $n-2$ graus de liberdade. Rejeitamos a hipóteses nula se $|T_{obs}| > t_{\alpha/2,\, n-2}$.

Um caso particular muito importante das hipóteses é:

$$\begin{aligned} H_0 : \beta_1 &= 0 \\ H_1 : \beta_1 &\neq 0 \end{aligned} \tag{17}$$

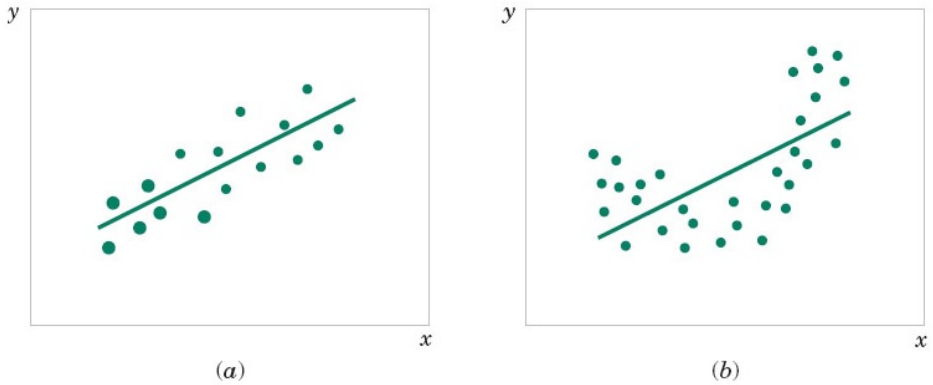
Esse teste está relacionado com a significância do modelo de regressão. Deixar de rejeitar $H_0 : \beta_1 = 0$ é equivalente a concluir que não há nenhuma relação linear entre X e Y, conforme figura abaixo:



Fonte: adaptado de [Montgomery e Runger(2016)]

Note que esse resultado pode implicar que X é pouco importante para explicar a variação Y e o melhor estimador de Y para qualquer X é $\hat{Y} = \bar{Y}$ (figura a), ou que a verdadeira relação entre X e Y não é linear

Como alternativa, se $H_0 : \beta_1 = 0$ é rejeitado, implica que X tem importância ao explicar a variabilidade de Y (veja a figura abaixo).



Fonte: adaptado de [Montgomery e Runger(2016)]

Contudo, a rejeição de $H_0 : \beta_1 = 0$ pode significar que o modelo linear é adequado (figura a), ou que, mesmo havendo um efeito linear de X, melhores resultados podem ser obtidos com a adição de termos polinomiais de ordem maior em X (figura b).

Exemplo

Realize o teste de significância para o MRLS para os dados do exemplo desta seção, usando $\alpha = 0,05$. As hipóteses são

$$\begin{aligned} H_0 : \beta_1 &= 0 \\ H_1 : \beta_1 &\neq 0. \end{aligned}$$

Foram calculados anteriormente:

$$\hat{\beta}_1 = 0,00873, \quad n = 20 \quad S_{xx} = 614,603, \quad \hat{\sigma}^2 = 0,2513,$$

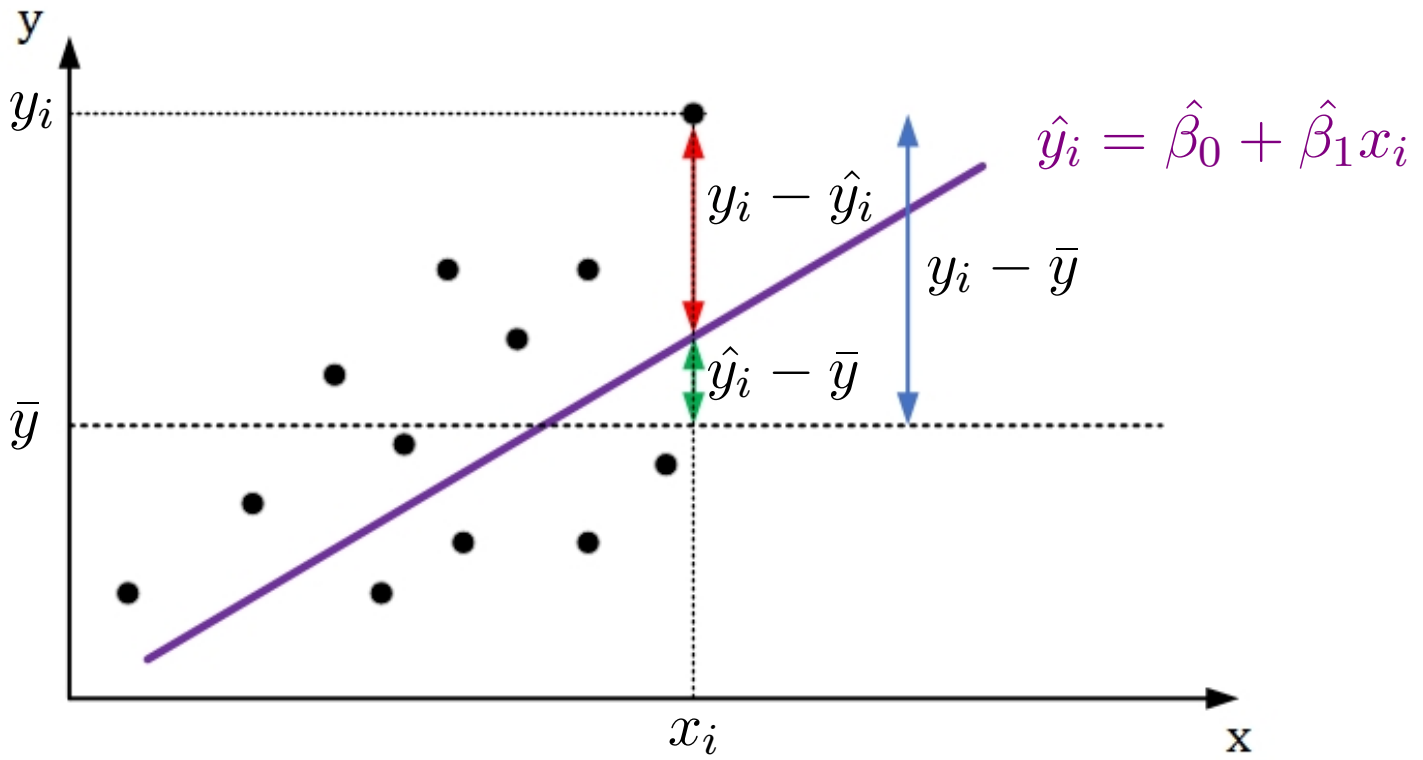
De modo que a estatística de teste, dada em (28), é:

$$T_{obs} = \frac{\hat{\beta}_1}{\sqrt{\hat{\sigma}^2/S_{xx}}} = \frac{0,00873}{\sqrt{0,2513/614.603}} = 13,65.$$

Como $T_{obs} = 13,65 > t_{0,025;18} = 2,101$, rejeita-se a hipótese $H_0 : \beta_1 = 0$. Portanto, conclui-se ao nível de significância de 5%, que existe uma relação linear significativa entre o número de clientes e as vendas semanais.

Abordagem de Análise de Variância para Testar Significância de Regressão

Para testar a significância do modelo de regressão ($H_0 : \beta_1 = 0$,) pode-se utilizar o método conhecido como **análise de variância**. O método consiste em decompor a variabilidade da variável resposta em componentes mais manejáveis.



Considere a seguinte *identidade de análise de variância*:

$$\underbrace{\sum_{i=1}^{n}(y_i-\bar{y})^2}_{\text{Variação total}} = \underbrace{\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2}_{\text{Variação explicada pela regressão}} + \underbrace{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}_{\text{Variação não explicada}} \tag{18}$$

Reescreveremos este resultado como:

$$SQT = SQ_{reg} + SQR,$$

sendo:

- SQT: a soma de quadrados total $= S_{yy}$;
- SQ_{reg}: soma dos quadrados da regressão $= \hat{\beta}_1 S_{xy}$;
- SQR: soma dos quadrados dos erros $= SQT - SQ_{reg}$.

Pode-se mostrar que a soma de quadrado total, SQT, tem $n-1$ graus de liberdade e, $SQreg$ e SQR têm respectivamente 1 e $n-2$ graus de liberdade.

Teste de Hipóteses

$$\begin{aligned} H_0 : \beta_1 &= 0 \\ H_1 : \beta_1 &\neq 0 \end{aligned} \tag{19}$$

Estatística de Teste

$$F_o = \frac{SQreg/1}{SQR/(n-2)} = \frac{QMreg}{QMR} \sim F_{1;n-2}. \tag{20}$$

Critério de Rejeição

Rejeita-se H_0 se $F_o > F_{\alpha,\,1,\,n-2}$.

As quantidades $QMreg = SQreg/1$ e $QMR = SQR/(n-2)$ são denominadas respectivamente, **quadrado médio devido à regressão** e **quadrado médio devido aos residuais**. O procedimento do teste é usualmente representado em uma tabela de análise de variância, como mostrada na tabela abaixo:

Tabela: Análise de variância para o teste de $H_0 : \beta_1 = 0$

Fonte de variação	Soma de Quadrados	Graus de Liberdade	Quadrado Médio	F
Regressão	$SQreg = \hat{\beta}_1 S_{xy}$	1	$QMreg$	$\frac{QMreg}{QMR}$
Residual	$SQR = SQT - SQreg$	$n-2$	QMR	
Total	$SQT = S_{yy}$	$n-1$		

Exemplo

A seguir é apresentado o procedimento de análise de variância para testar se de fato existe relação linear entre o número de clientes (X) e as vendas semanais (Y), no modelo proposto para os dados do exemplo desta seção. (Use $\alpha = 0,05$)

Relembre que $S_{yy} = 51,3605$, $\hat{\beta}_1 = 0,00873$, $S_{xy} = 5.365,08$ e $n = 20$. A soma de quadrados da regressão é

$$SQreg = \hat{\beta}_1 S_{xy} = (0,00873)(5.365,08) = 46,8371$$

enquanto a soma de quadrados dos residuais é:

$$SQR = SQT - \hat{\beta}_1 S_{xy} = 51,3605 - 46,8371 = 4,5234.$$

Na tabela abaixo, é apresentado um resumo da análise de variância para testar $H_0 : \beta_1 = 0$. Nesse caso, a estatística de teste é $F_{0bs} = QMreg/QMR = 46,837148/0,2512 = 186,4536$. Como $F_{obs} = 186,4536 > F_{0,05,1,18} = 4,41$ rejeita-se H_0, ao nível de significância de 5%.

Tabela: Análise de variância para o teste de $H_0 : \beta_1 = 0$ do exemplo

Fonte de variação	Soma de Quadrados	Graus de Liberdade	Quadrado Médio	F
Regressão	46,8371	1	46,8371	186,4536
Residual	4,5234	18	0,2513	
Total	51,3605	19		

Observação 1

Note, que o procedimento de análise de variância para testar a significância da regressão é equivalente o teste T dado no início desta seção. Portanto, qualquer desses procedimentos conduz às mesmas conclusões.

Observação 2

O teste T é um pouco mais flexível , pois permite testar hipóteses unilaterais, enquanto que o teste F é restrito ao teste bilateral.

Intervalos de confiança

Intervalos de confiança para β_1 e β_0

Além das estimativas pontuais para a inclinação e o intercepto da linha de regressão, é possível obter estimações por intervalos de confiança para esses parâmetros. O comprimento desses intervalos é uma medida da qualidade total da linha de regressão. Se para o MRLS é válida a suposição de que os $\varepsilon_i \sim NID(0, \sigma^2)$, então

$$IC(\beta_1; 1-\alpha) = \left(\hat{\beta}_1 \pm t_{\frac{\alpha}{2},\, n-2} \sqrt{\frac{QMR}{S_{xx}}} \right) \tag{21}$$

$$IC(\beta_0; 1-\alpha) = \left(\hat{\beta}_0 \pm t_{\frac{\alpha}{2},\, n-2} \sqrt{QMR[\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}]} \right) \tag{22}$$

Exemplo

A seguir é obtido um intervalo de 95% de confiança para a inclinação do MRLS com os dados do exemplo dessa seção.

Relembre que $n = 20$, $\hat{\beta}_1 = 0,00873$, $S_{xx} = 614,603$ e $QMR = 0,2513$. Para $1-\alpha = 0,95$, tem-se $t_{0,025,\,18} = 2,101$. Então da equação (21), vem:

$$
\begin{aligned}
IC(\beta_1; 0,95) &= \left(\hat{\beta}_1 - t_{0,025,18}\sqrt{\frac{QMR}{S_{xx}}} \; ; \; \hat{\beta}_1 + t_{0,025,18}\sqrt{\frac{QMR}{S_{xx}}} \right) \\
&= \left(0,00873 - 2,101\sqrt{\frac{0,2513}{614.603}} \; ; \; 0,00873 + 2,101\sqrt{\frac{0,251}{614.60}} \right. \\
&= (0,00873 - 0,00134; 0,00873 + 0,00134)
\end{aligned}
$$

Ou seja,

$$IC(\beta_1; 0,95) = (0,00739; 0,01007).$$

Intervalo de confiança para a resposta média

Também é possível construir intervalos de confiança para a resposta média correspondente a um valor especificado da variável explicativa, que representaremos por x_0. Ou seja, o interesse consiste em estimar um intervalo de confiança para $E(Y|X = x_0) = \mu_{Y|x_0} = \beta_0 + \beta_1 x_0$. Um estimador pontual de $\mu_{Y|x_0}$ pode ser obtido a partir do modelo de regressão ajustado

$$\hat{\mu}_{Y|x_o} = \hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x_0.$$

Considerando que a suposição de que os $\varepsilon_i \sim NID(0, \sigma^2)$ é válida, vem:

$$IC(\hat{\mu}_{Y|x}; 1-\alpha) = \left(\hat{\mu}_{Y|x_o} \pm t_{\frac{\alpha}{2},\, n-2} \sqrt{QMR[\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}} \right) \qquad (23)$$

Observe que o comprimento de intervalo de confiança para $\hat{\mu}_{Y|x}$ é mínimo quando $x_0 = \bar{x}$. E aumenta à medida que $|x_0 - \bar{x}|$ aumenta.

Exemplo

Para o problema dos supermercados, suponha que tem-se interesse em construir um intervalo de 95% de confiança para a venda média semanal para todos os supermercados com 607 clientes.
No modelo ajustado $\hat{\mu}_{Y|x_0} = 2,423 + 0,00873x_0$. Para $x_0 = 607$, obtém-se $\hat{\mu}_{Y|x_0} = 7,72211$. Também,

$$\bar{x} = 731,15, \quad QMR = 0,2513, \quad S_{xx} = 614.603, \quad n = 20 \text{ e } 1-\alpha = 0,95 \Rightarrow$$

Substituindo esses valores na equação (23), obtém-se o seguinte intervalo de confiança:

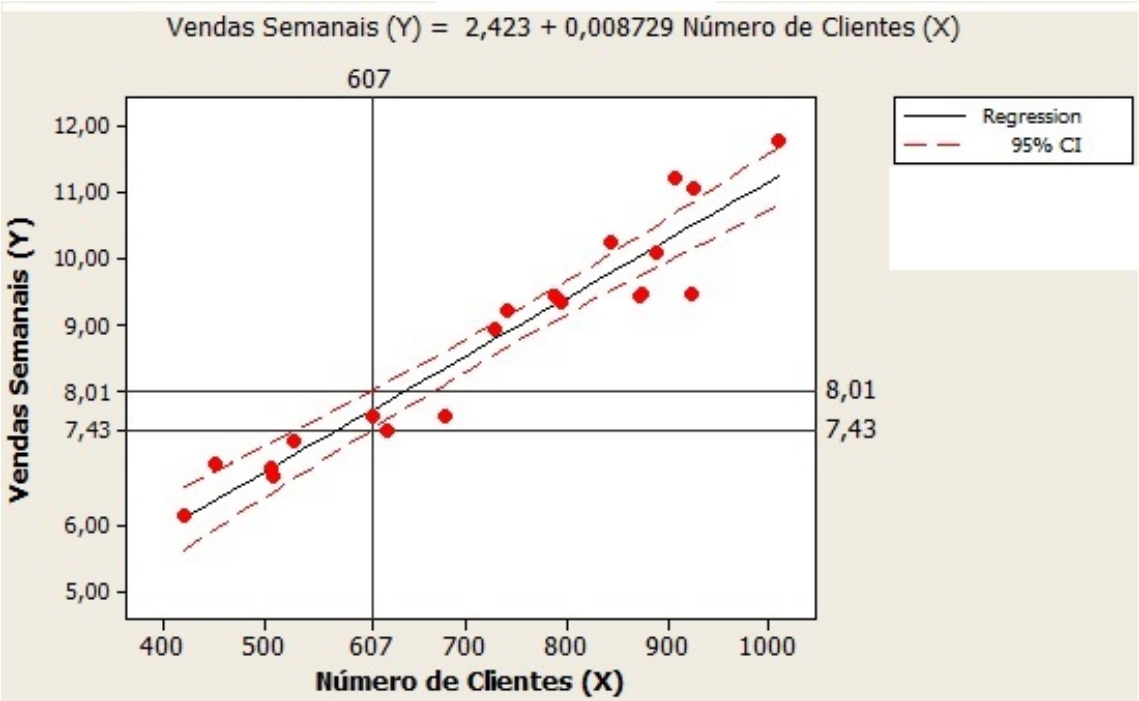
$$IC(\mu_{Y|x_0=607}; 0,95) = \left(7,72211 \pm 2,101\sqrt{0,2513[\frac{1}{20} + \frac{(607 - 731,}{614.603}\right.$$

$$= (7,72211 - 0,288; 7,72211 + 0,288)$$

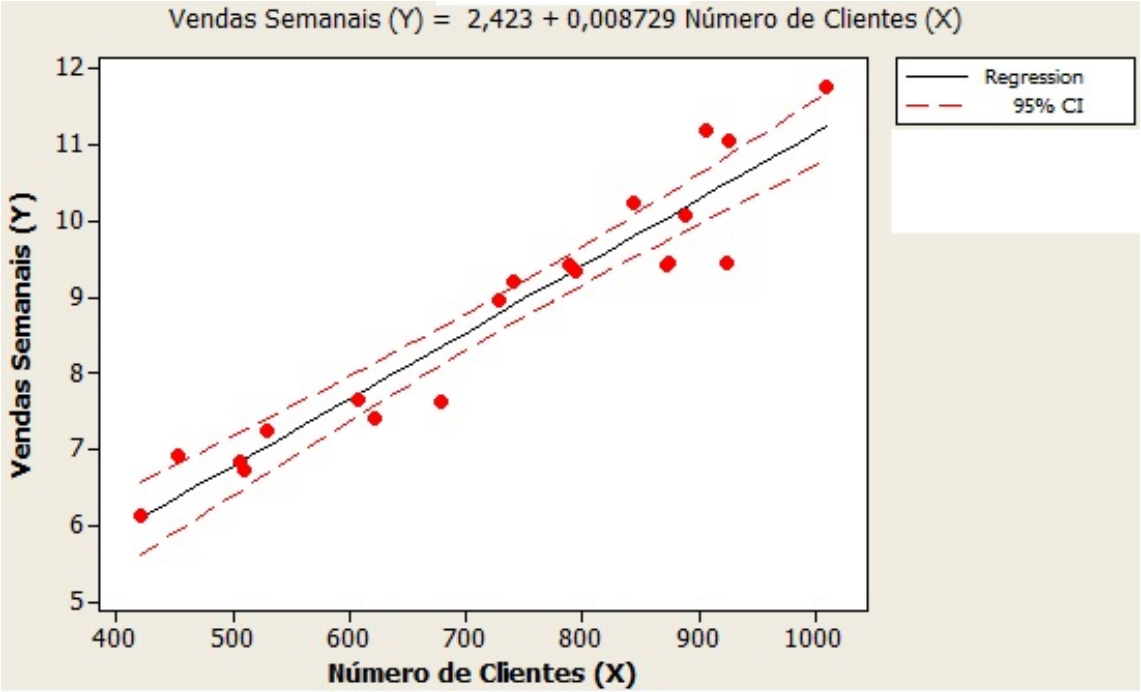
$$= (7,43411; 8,01011).$$

Observação

Ao repetir os cálculos anteriores para valores diferentes de x_0, obtém-se os limites de confiança para cada $\mu_{Y|x_0}$. Na figura abaixo, é mostrado o diagrama de dispersão com o modelo de regressão ajustado e os correspondentes limites de confiança de 95% (bandas de confiança).



Observe que o comprimento do intervalo de confiança para $\mu_{Y|x_0}$ aumenta a medida que $|x_0 - \bar{x}|$ aumenta.



Previsão de novas observações

Uma aplicação muito importante de um modelo de regressão é a previsão de novas ou futuras observações de Y, (Y_0) correspondente a um dado valor da variável explicativa X, x_0, então

$$\hat{Y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0 \tag{24}$$

é o melhor estimador pontual de Y_0.

Intervalo de confiança para a previsão

$$IC(Y_0; 1-\alpha) = \left(\hat{Y} \pm t_{\frac{\alpha}{2},\, n-2} \sqrt{QMR[1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}]} \right) \tag{25}$$

Observe que o comprimento do intervalo de confiança para a nova observação é mínimo quando $x_0 = \bar{x}$ e aumenta a medida que $|x_0 - \bar{x}|$ aumenta.

Exemplo

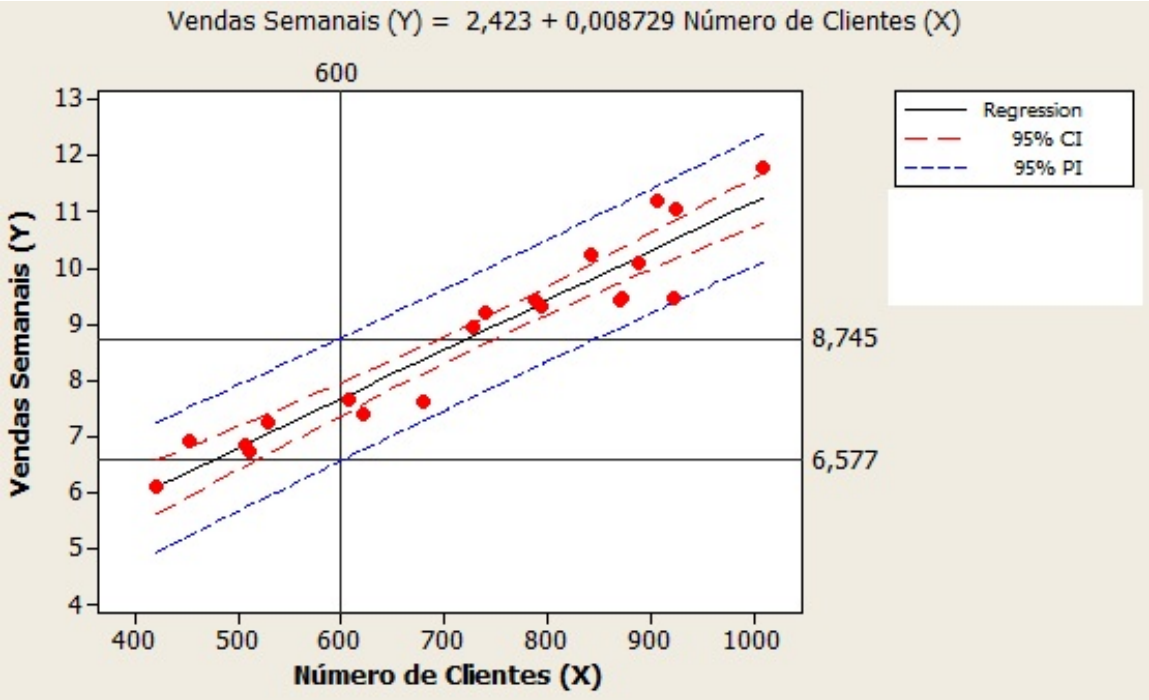
Para ilustrar a construção de um intervalo de previsão, considere os dados do exemplo desta seção e suponha agora, que se tem interesse em encontrar um intervalo de previsão de 95% das vendas semanais de um supermercado com 600 clientes.

Considerando a equação (25) e os dados do exemplo , $\hat{Y} = 7,661$ e o intervalo de previsão é:

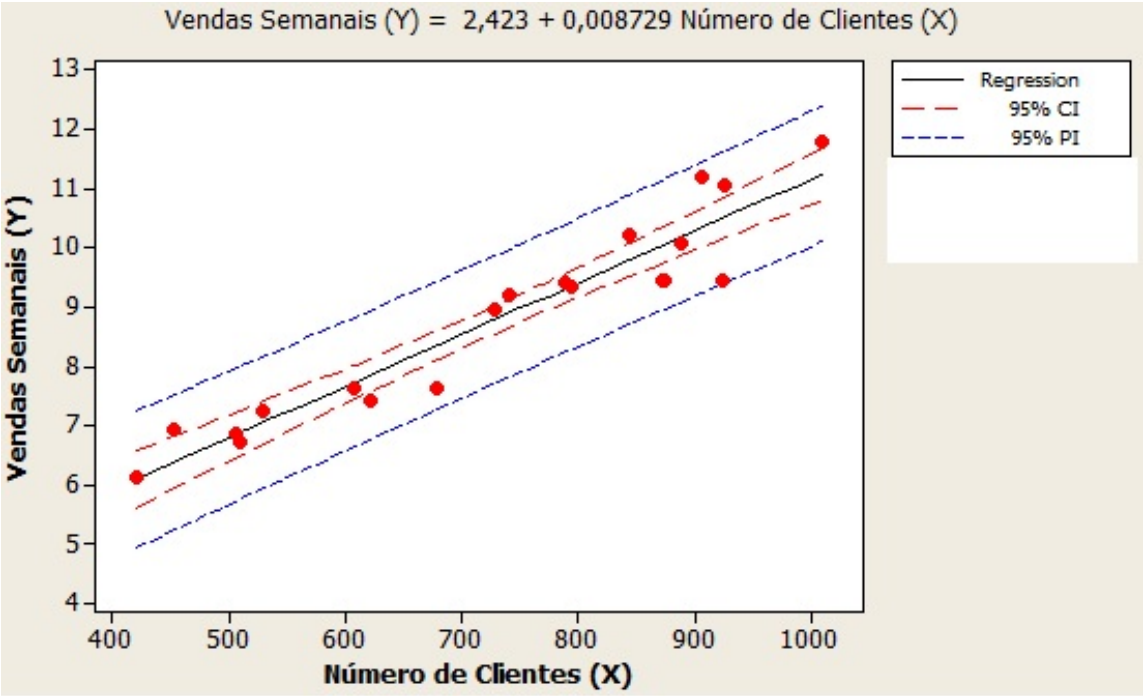
$$\begin{aligned} IC(Y_0; 0,95) &= \left(7,661 - 2,101\sqrt{0,2513[1 + \frac{1}{20} + \frac{(600 - 731,15)^2}{614.603}]};\right. \\ &\qquad \left. 7,661 + 2,101\sqrt{0,2513\left[1 + \frac{1}{20} + \frac{(600 - 731,15)^2}{614.603}\right]}\right) \\ &= (7,661 - 1,084; 7,661 + 1,084) \\ &= (6,577; 8,745). \end{aligned}$$

Observação

Ao repetir os cálculos anteriores para diferentes valores de x_0, podemos obter os intervalos de previsão de 95%, que estão representados na figura abaixo.



Observe que esse gráfico também apresenta os limites de confiança do 95% para $\mu_{Y|x_0}$, calculados anteriormente. Isto ilustra que os limites de previsão sempre são mais amplos que os limites de confiança da $\mu_{Y|x_0}$.



Estudo da adequação do modelo de regressão

O ajuste de um modelo de regressão requer várias suposições:

- A estimação dos parâmetros do modelo requer a suposição de que os erros sejam variáveis aleatórias não correlacionadas com média zero e variância constante;
- A construção de intervalos de confiança e testes de hipóteses requer que os erros sejam normalmente distribuídos;
- Além disso, é assumindo que a ordem do modelo é correta; isto é, se ajustamos um modelo de regressão linear simples, considera-se que o fenômeno realmente se comporta dessa forma.

O pesquisador deve sempre questionar a validade dessas suposições e realizar análises para verificar a adequação do modelo adotado. Nesta subseção serão discutidos métodos úteis para o estudo da adequação do modelo de regressão.

(1) - Análise Residual

A análise de resíduos é útil para verificar a suposição de que os erros são não correlacionados e têm uma distribuição que é aproximadamente normal com média zero e variância constante, assim como para determinar se é necessária a adição de termos adicionais ao modelo.

Os resíduos de um modelo de regressão são definidos como

$$e_i = y_i - \hat{y}_i, \quad i = 1, \ldots, n \qquad \rightarrow \text{resíduo regular.}$$

onde y_i é uma observação real de Y e $\hat{y}_i$ é o valor correspondente estimado através do modelo de regressão.

Um procedimento muito útil consiste em padronizar os resíduos assim:

1. $$d_i = \frac{e_i}{\sqrt{\hat{\sigma}^2}}, \quad i = 1, \ldots, n \qquad \rightarrow \text{resíduo padronizado.}$$

2. $$z_i = \frac{e_i}{\sqrt{\hat{\sigma}^2(1 - h_{ii})}}, \quad i = 1, \ldots, n \qquad \rightarrow \text{resíduo padronizado.}$$

3. $$z_i^* = \frac{e_i}{\sqrt{\hat{\sigma}^2_{(i)}(1 - h_{ii})}}, \quad i = 1, \ldots, n \qquad \rightarrow \text{resíduo estudentizado.}$$

Onde:

- $h_{ii} = (\frac{1}{n} + \frac{(x_i - \bar{x})^2}{S_{xx}})$
- $\sigma^2_{(i)}$ é a variância estimada sem utilizarmos a i-ésima observação.

Observação 1

Se os erros tem distribuição normal, então aproximadamente 95% dos resíduos padronizados devem pertencer ao intervalo $(-2, +2)$.

Os resíduos fora desse intervalo podem indicar a presença de um valor atípico (*"outlier"*). Isto é, uma observação que não é comum do restante da massa de dados

Na literatura, foram propostas várias regras para descartar valores atípicos. Porém , muitas vezes, os *"outliers"* fornecem informações importantes sobre situações pouco usuais que são de interesse para o pesquisador e não devem, em princípio, ser descartados.

Observação 2

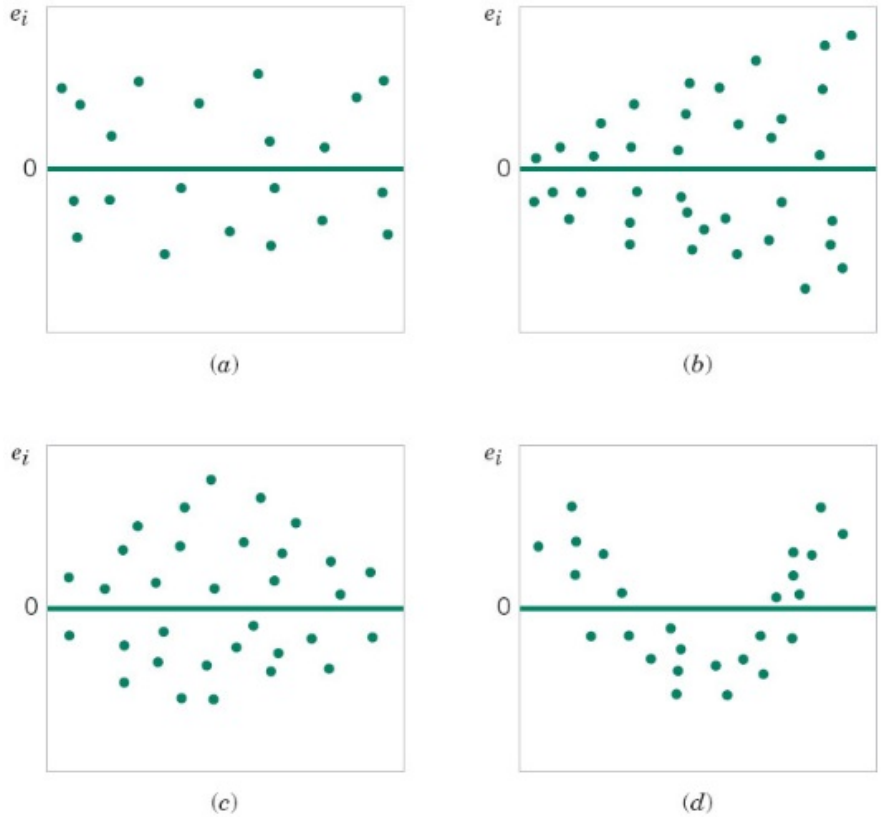
Remover observações pode afetar a estimativa da variância e também as estimativas dos parâmetros.

Um grande valor absoluto do resíduo estudentizado pode indicar que a inclusão/exclusão de uma observação no modelo aumenta/diminui a variância do erro ou que tem uma grande influência.

Geralmente, é útil fazer um gráfico dos resíduos:

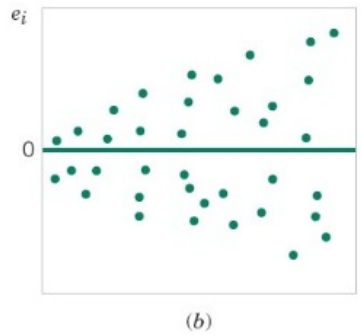
- com uma sequência no tempo (se é conhecida);
- em relação aos $\hat{y}$;
- em função da variável independente x.

Usualmente, esses gráficos tem aspecto similar a um dos quatro padrões gerais que aparecem na figura abaixo:



Fonte: adaptado de [Montgomery e Runger(2016)]

O padrão (a) dessa figura representa a situação ideal, enquanto que os padrões (b), (c) e (d) representam anomalias.

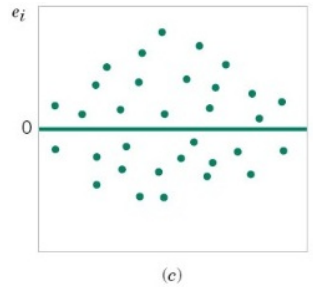


Fonte: adaptado de [Montgomery e Runger(2016)]

Se os resíduos aparecem como em (b), a variância das observações pode aumentar com o tempo ou com a magnitude de Y ou X.

Usualmente uma transformação nos dados sobre a resposta Y elimina este problema.

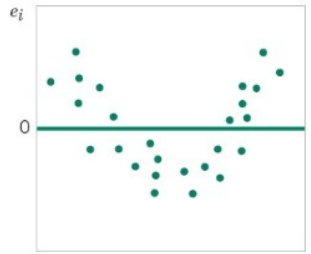
Entre as transformações mais usadas para estabilizar a variância se inclui o emprego de: $\sqrt{y}$, $\ln y$, $1/y$ ou $arcsen\sqrt{y}$.



(c)

Fonte: adaptado de [Montgomery e Runger(2016)]

Os gráficos dos resíduos com $\hat{y}$ ou com x, semelhantes a(c) também indicam uma desigualdade da variância.



(d)

Fonte: adaptado de [Montgomery e Runger(2016)]

Gráficos dos resíduos semelhantes ao de figura (d), indicam que modelo é inadequado, isto é, que é necessário adicionar ao modelo termos de ordem superior, considerar uma transformação da variável x ou da variável y (ou ambas), ou considerar outras variáveis explicativas.

A seguir é apresentada a análise residual para o modelo de regressão ajustado aos dados de exemplo desta seção.

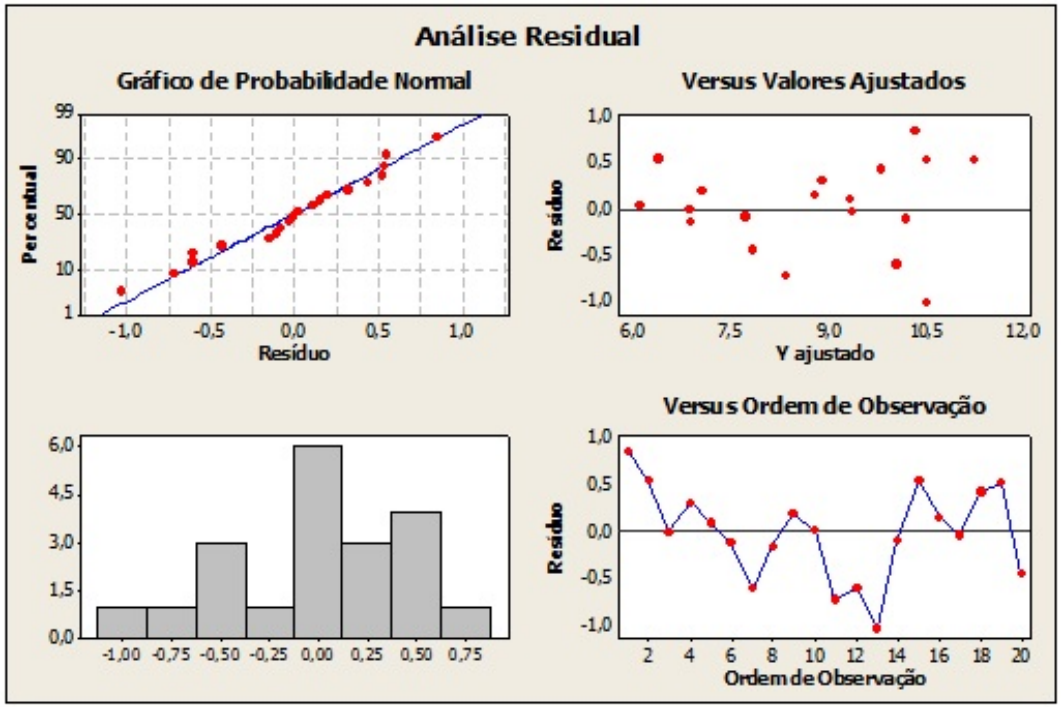
Na tabela a seguir, são apresentados os valores observados e ajustados de Y para cada valor de x que aparece no conjunto dos dados.

Supermercado	Número de Clientes	Vendas Semanais	Valor Ajustado	Resíduo	Resíduo Padronizado 1	Resíduo Padronizado 2	Resíduo Estudentizado
(i)	(x_i)	(y_i)	$(\hat{y}_i)$	$e_i = y_i - \hat{y}_i$	$z_i = \frac{e_i}{\sqrt{\hat{\sigma}^2}}$	$\frac{e_i}{\sqrt{\hat{\sigma}^2(1-h_{ii})}}$	$\frac{e_i}{\sqrt{\hat{\sigma}^2_{(i)}(1-h_{ii})}}$
1	907	11,2	10,341	0,859	1,714	1,807	1,941
2	926	11,05	10,506	0,544	1,084	1,150	1,161
3	506	6,84	6,840	0,000	0,000	0,000	0,000
4	741	9,21	8,891	0,319	0,635	0,652	0,641
5	789	9,42	9,310	0,110	0,218	0,225	0,219
6	889	10,08	10,183	-0,103	-0,206	-0,216	-0,210
7	874	9,45	10,052	-0,602	-1,202	-1,255	-1,276
8	510	6,73	6,875	-0,145	-0,289	-0,310	-0,302
9	529	7,24	7,041	0,199	0,397	0,422	0,413
10	420	6,12	6,089	0,031	0,061	0,069	0,067
11	679	7,63	8,350	-0,720	-1,437	-1,477	-1,531
12	872	9,43	10,035	-0,605	-1,207	-1,259	-1,282
13	924	9,46	10,489	-1,029	-2,053	-2,175	-2,463
14	607	7,64	7,722	-0,082	-0,163	-0,170	-0,165
15	452	6,92	6,369	0,551	1,100	1,212	1,229
16	729	8,95	8,787	0,163	0,326	0,334	0,326
17	794	9,33	9,354	-0,024	-0,048	-0,050	-0,048
18	844	10,23	9,791	0,439	0,877	0,909	0,904
19	1010	11,77	11,240	0,530	1,058	1,165	1,178
20	621	7,41	7,844	-0,434	-0,866	-0,897	-0,892

Nas figuras a seguir, são apresentados os gráficos da análise residual.

Análise 1 - Resíduos regulares

(a)



(b)

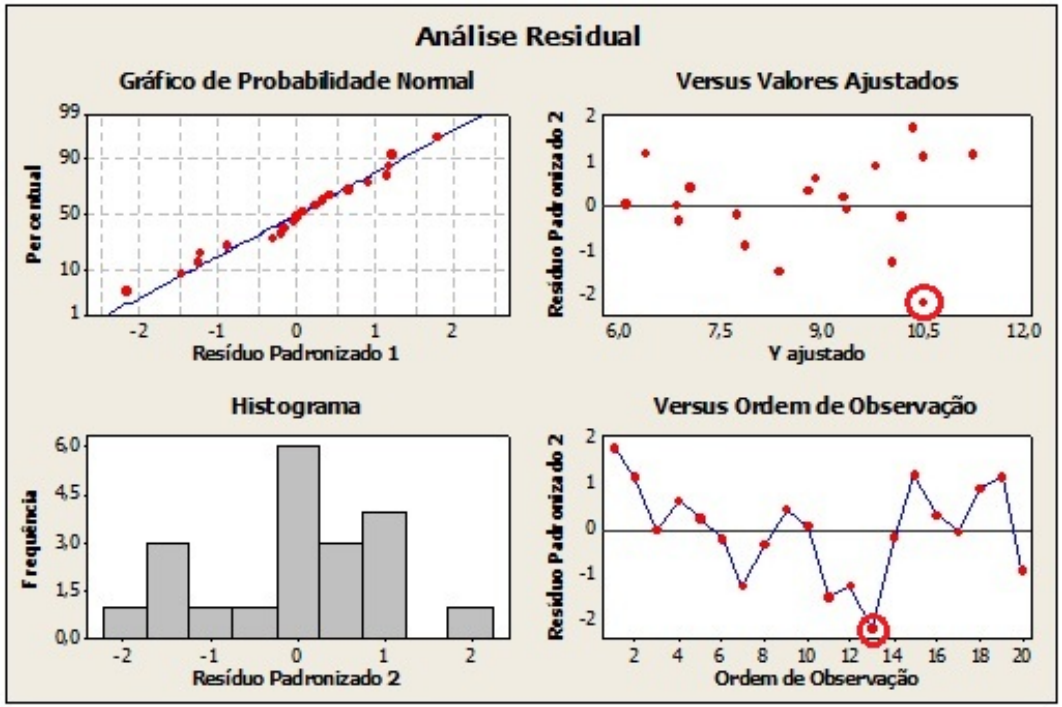
(c)

(d)

A figura (a) mostra um gráfico de probabilidade normal dos resíduos. Como esses resíduos estão localizados aproximadamente ao longo de uma linha reta, conclui-se que há uma forte indicação de que a suposição de normalidade dos erros seja adequada. Na figura (b), mostra-se o gráfico de resíduos com os valores ajustados ($\hat{y}_i$). Enquanto na figura (d), representa-se o gráfico de resíduos *versus* a ordem de observação.

Análise 2 - Resíduos padronizados do tipo 2

(a)



(b)

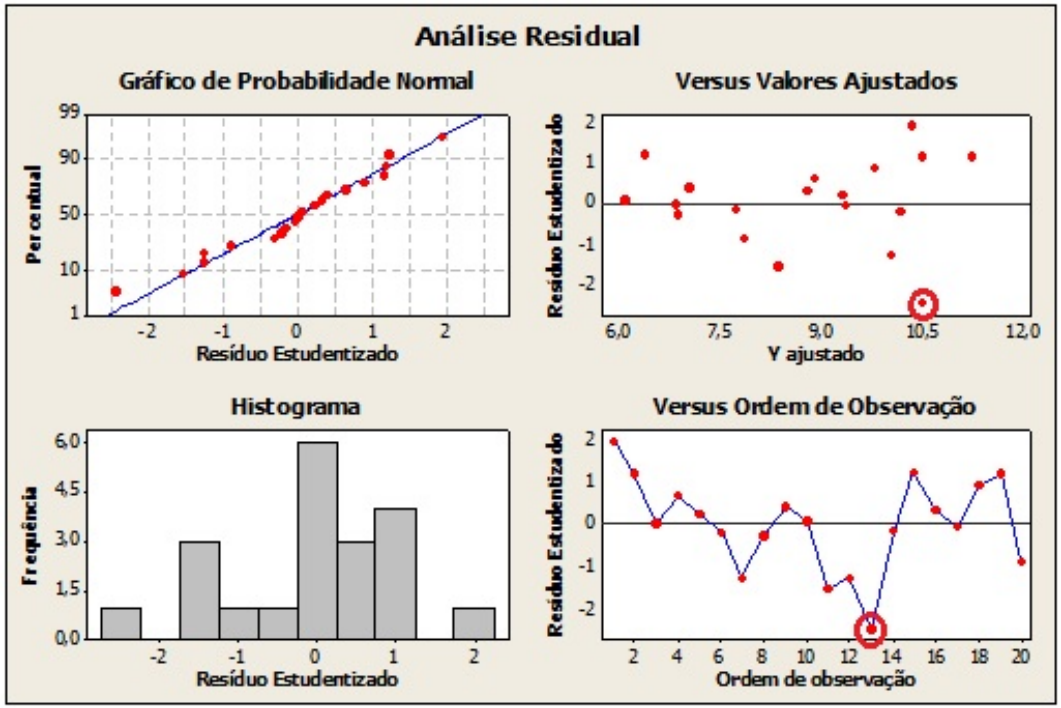
(c)

(d)

A figura (a) mostra um gráfico de probabilidade normal dos resíduos. Na figura (b), mostra-se o gráfico de resíduos padronizados com os valores ajustados ($\hat{y}_i$). Enquanto na figura (d), representa-se o gráfico de resíduos padronizados *versus* a ordem de observação. Porém, tanto as figuras (c) e (d) mostram uma observação (o supermercado 13), cujo resíduo está externo ao intervalo (-2, +2), o qual poderia ser considerado um valor atípico.

Análise 3 - Resíduos estudentizados

(a)



(b)

(c)

(d)

A figura (a) mostra um gráfico de probabilidade normal dos resíduos. Na figura (b), mostra-se o gráfico de resíduos estudentizados com os valores ajustados ($\hat{y}_i$). Enquanto na figura (d), representa-se o gráfico de resíduos estudentizados *versus* a ordem de observação. Porém, tanto as figuras (c) e (d) mostram uma observação (o supermercado 13), cujo resíduo está externo ao intervalo (-2, +2), o qual poderia ser considerado um valor atípico.

(2) - Coeficiente de determinação

Uma medida largamente usada para um modelo de regressão é a razão de soma dos quadrados.
A quantidade:

$$R^2 = \frac{SQreg}{SQT} = 1 - \frac{SQR}{SQT} \tag{26}$$

recebe o nome de **coeficiente de determinação** que é usado para ajudar a julgar a adequação do modelo de regressão.

Da identidade da análise de variância, dada em (18), temos que $0 \leq R^2 \leq 1$.

O coeficiente de determinação pode ser interpretado como a proporção da variabilidade presente nas observações da variável resposta Y, que é explicada pela variável independente X no modelo de regressão.

Exemplo

Para os dados dos supermercados do exemplo desta seção, determinar R^2.

Da equação (26) tem-se:

$$R^2 = \frac{SQreg}{SQT} = \frac{46,8371}{51,3605} = 0,912$$

Esse resultado significa que o modelo ajustado explicou 91,2% da variação na variável resposta Y (vendas semanais). Isto é, 91,2% da variabilidade de Y é explicada pela variável regressora X (número de clientes).

Observação 1

Existem várias idéias errôneas quanto a R^2:

- Em geral, R^2 não mede a magnitude da inclinação da reta de regressão. Um grande valor de R^2 não implica em um valor alto para inclinação da reta de regressão.
- Por outro lado, R^2 não mede sozinho a adequação do modelo. Por exemplo, o R^2 para a equação de regressão da figura pode ser relativamente grande, mesmo que a aproximação linear seja pobre.

Fonte: adaptado de [Montgomery e Runger(2016)]

- Finalmente, mesmo que R^2 seja grande, não implica necessariamente, que o modelo de regressão proporcione previsões precisas de observações futuras.

Observação 2

Uma alternativa para o teste de significância do modelo pode ser construída a partir do coeficiente de determinação R^2, da seguinte maneira:

$$\begin{aligned} H_0 : \beta_1 &= 0 \\ H_1 : \beta_1 &\neq 0 \end{aligned} \tag{27}$$

Estatística de teste

$$F_o = \frac{(n-2)R^2}{1-R^2} \sim F_{(1;n-2)} \tag{28}$$

se H_0 é verdadeira.

Critério de Rejeição

Rejeita-se H_0 se

$$F_o > f_{\alpha,\,1,\,n-2}.$$

Análise de Correlação

Conforme foi mencionado no início deste capítulo, a *análise de regressão* é usada quando tem-se interesse em estabelecer o tipo de relação (linear, quadrática, exponencial, etc.) que há entre uma variável dependente e uma ou mais variáveis independentes.

Mas, quando tem-se interesse estabelecer o grau (forte, moderado, fraco) dessa relação é usada a *análise de correlação.*

O coeficiente de correlação é uma quantidade adimensional que mede a força da associação linear entre duas variáveis aleatórias.

$$\rho = \frac{E[(X - \mu_x)(Y - \mu_y)]}{\sigma_x \sigma_y} \tag{29}$$

$$-1 \leq \rho \leq +1$$

Os métodos já apresentados na seção anterior podem ser empregados para análise de modelos onde X e Y são variáveis aleatórias com distribuição conjunta normal bivariada (*).

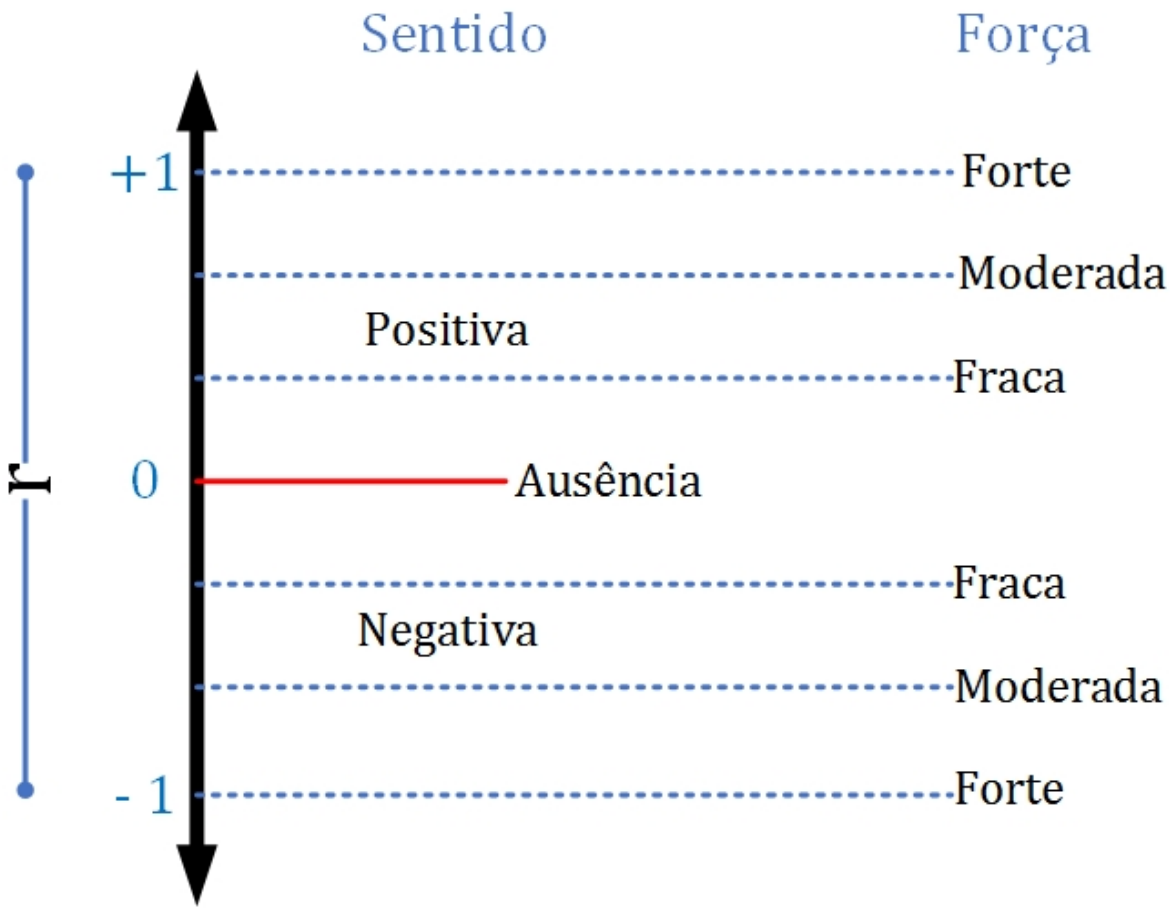
($*$) Relembrando: A distribuição conjunta de X e Y tem uma distribuição normal bivariada quando a função de densidade é dada por

$$f(x,y) = \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}} \exp\left\{\frac{1}{2(1-\rho^2)}\left[\left(\frac{x-\mu_x}{\sigma_x}\right)^2 + \left(\frac{y-\mu_y}{\sigma_y}\right)^2 - 2\rho\left(\frac{x-\mu_x}{\sigma_x}\right)\left(\frac{y-\mu_y}{\sigma_y}\right)\right]\right\} \tag{30}$$

É possível realizar inferência sobre o coeficiente de correlação ρ desse modelo. Um estimador de ρ é o coeficiente de correlação amostral, representado por r e definido por

$$r = \frac{\sum_{i=1}^{n} Y_i(X_i - \bar{X})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2 \sum_{i=1}^{n}(Y_i - \bar{Y})^2}} = \frac{S_{XY}}{\sqrt{S_{XX}S_{YY}}} \tag{31}$$

Valores possíveis de r e interpretação da correlação



É possível demonstrar que:

$$\hat{\beta}_1 = \left(\frac{S_{YY}}{S_{XX}}\right)^{1/2} r. \tag{32}$$

O coeficiente de correlação amostral r mede a força da associação linear entre X e Y, enquanto $\hat{\beta}_1$ mede a alteração esperada em Y quando X sofre uma variação unitária.

Além disso, vem:

$$r^2 = \hat{\beta}_1^2 \frac{S_{XX}}{S_{YY}} = \frac{\hat{\beta}_1 S_{XY}}{S_{YY}} = \frac{SQreg}{SQT} = R^2.$$

onde R^2 é o coeficiente de determinação definido na equação (26). Isto é, o coeficiente de determinação R^2 é igual ao quadrado do coeficiente de correlação amostral entre X e Y.

Inferência estatística sobre a correlação

Caso 1

Em análise de correlação, pode ser interessante testar se o coeficiente de correlação é igual a zero, já que, $\rho = 0$ significa ausência de relacionamento linear entre Y e X. As hipóteses a serem testadas são:

$$\begin{aligned} H_0 &: \rho = 0 \\ H_1 &: \rho \neq 0. \end{aligned} \tag{33}$$

Estatística de Teste

$$T = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \sim t_{n-2} \tag{34}$$

Critério de Rejeição

A hipótese nula deverá ser rejeitada se $|T_{obs}| > t_{\alpha/2,\, n-2}$.

* Esse teste é equivalente ao teste de hipóteses $H_0 : \beta_1 = 0$, apresentado na seção anterior.

Caso 2

O procedimento para o teste das hipóteses

$$\begin{aligned} H_0 &: \rho = \rho_0 \\ H_1 &: \rho \neq \rho_0. \end{aligned} \tag{35}$$

onde $\rho_0 \neq 0$, é um pouco mais complicado.

Estatística de Teste

Para amostras de tamanho moderado grande ($n \geq 25$), a estatística

$$Z_r = \operatorname{arctanh} r = \frac{1}{2} \ln \frac{1+r}{1-r} \tag{36}$$

tem distribuição aproximadamente normal com média

$$\mu_{Z_r} = \operatorname{arctanh} \rho = \frac{1}{2} \ln \frac{1+\rho}{1-\rho}$$

e variância

$$\sigma^2_{Z_r} = (n-3)^{-1}.$$

Portanto, para testar a hipóteses $H_0 : \rho = \rho_0$, a estatística de teste apropriada é:

$$Z = (\operatorname{arctanh} r - \operatorname{arctanh} \rho_0)\,(n-3)^{1/2}. \tag{37}$$

Critério de Rejeição

Se $H_0 : \rho = \rho_0$ é verdadeira, a estatística Z tem, aproximadamente, distribuição normal padrão. Portanto, H_0 deverá ser rejeitada se $|Z_{obs}| > z_{\alpha/2}$.

Intervalo de confiança para ρ

Além disso, é possível construir um intervalo aproximado de $100(1-\alpha)\%$ de confiança para o coeficiente de correlação ρ, que é dado por:

$$IC(\rho; 1-\alpha) = \left(\tanh\left[\operatorname{arctanh} r - \frac{z_{\alpha/2}}{\sqrt{n-3}} \right]; \tanh\left[\operatorname{arctanh} r + \frac{z_{\alpha/2}}{\sqrt{n-3}} \right] \right) \tag{38}$$

onde

$$\tanh r = \frac{e^r - e^{-r}}{e^r + e^{-r}}.$$

Exemplo - [Montgomery e Runger(2016)]

A tabela a seguir contém dados de três variáveis, que foram coletados em um estudo de observação numa indústria de semicondutores.

	Resistência à tração	Comprimento do fio	Altura do molde
Number	y	x_1	x_2
1	9.95	2	50
2	24.45	8	110
3	31.75	11	120
4	35.00	10	550
5	25.02	8	295
6	16.86	4	200
7	14.38	2	375
8	9.60	2	52
9	24.35	9	100
10	27.50	8	300
11	17.08	4	412
12	37.00	11	400
13	41.95	12	500
14	11.66	2	360
15	21.65	4	205
16	17.89	4	400
17	69.00	20	600
18	10.30	1	585
19	34.93	10	540
20	46.59	15	250
21	44.88	15	290
22	54.12	16	510
23	56.63	17	590
24	22.13	6	100
25	21.15	5	400

Nessa planta, o semicondutor final é um fio colado a uma estrutura.

As variáveis reportadas são:

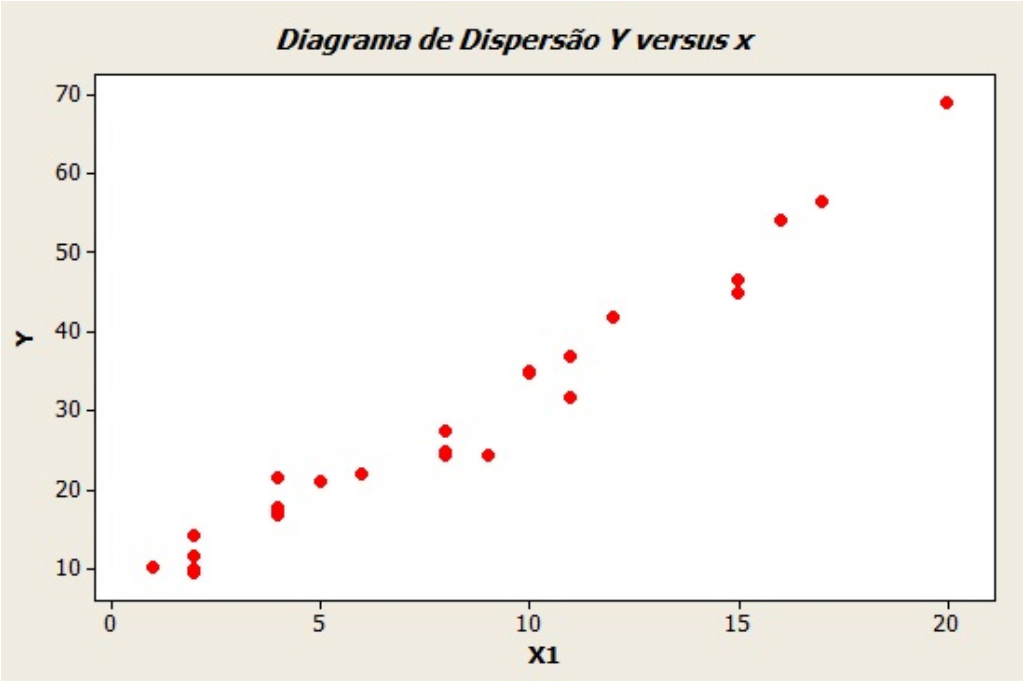
- a resistência à tração (uma medida de força requerida para romper a cola),
- o comprimento do fio, e
- a altura do molde.

Gostaríamos de encontrar um modelo relacionando a resistência à tração ao comprimento do fio e à altura do molde.

Resolução via Análise de Correlação e MRLS

Primeiramente, consideraremos que a resistência à tração e o comprimento do fio sejam distribuídos normal e conjuntamente

A figura abaixo mostra um diagrama de dispersão da resistência do fio colado *versus* o comprimento do fio.



Há evidência de uma relação linear entre as duas variáveis.

Cálculo de ρ

A partir dos dados da tabela, podemos calcular:

$$S_{xy} = 2027,7132 \quad S_{xx} = 698,56 \quad S_{yy} = SQT = 6105,9$$

Então, vem:

$$r = \frac{S_{XY}}{\sqrt{S_{XX} S_{YY}}} = 0,9818 \tag{39}$$

Suponha agora que desejemos testar a hipótese

$$\begin{aligned} H_0 &: \rho = 0 \\ H_1 &: \rho \neq 0. \end{aligned} \tag{40}$$

com $\alpha = 0,05$.

Podemos calcular a estatística T de teste da equação (34) como

$$T = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} = \frac{0,9818\sqrt{23}}{\sqrt{1-0,9640}} = 24,8$$

Uma vez que $t_{0,025;23} = 2,069$, rejeitamos H_0 e concluímos, com $\alpha = 0,05$, que $\rho \neq 0$.
Ou ainda, podemos dizer, a um nível de 5% de significância, que existe correlação linear positiva forte entre as variáveis resistência à tração e comprimento do fio

Além disso, podemos construir um intervalo aproximado de confiança de 95% para ρ a partir de (38).

$$IC(\rho; 1-\alpha) = \left(\tanh \left[\operatorname{arctanh} r - \frac{z_{\alpha/2}}{\sqrt{n-3}} \right] ; \tanh \left[\operatorname{arctanh} r + \frac{z_{\alpha/2}}{\sqrt{n-3}} \right] \right)$$

$$IC(\rho; 95\%) = \left(\tanh \left[\operatorname{arctanh} 0,9818 - \frac{1,96}{\sqrt{22}} \right] ; \tanh \left[\operatorname{arctanh} 0,9818 + \frac{1,9}{\sqrt{2}} \right] \right.$$

$$IC(\rho; 95\%) = \left(\tanh \left[2,3452 - \frac{1,96}{\sqrt{22}} \right] \leq \rho \leq \tanh \left[2,3452 + \frac{1,96}{\sqrt{22}} \right] \right)$$

$$IC(\rho; 95\%) = (0,9585 \leq \rho \leq 0,9921)$$

.

Uma vez mensurada a força da relação linear entre as variáveis resistência à tração e comprimento do fio, podemos construir um modelo de regressão linear simples para essa relação.

O modelo ajustado é dado a seguir:

The regression equation is

Y = 5,11 + 2,90 X1

Predictor	Coef	SE Coef	T	P
Constant	5,115	1,146	4,46	0,000
X1	2,9027	0,1170	24,80	0,000

S = 3,09342 R-Sq = 96,4% R-Sq(adj) = 96,2%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	5885,9	5885,9	615,08	0,000
Residual Error	23	220,1	9,6		
Total	24	6105,9			

Cancho, V., 2010. Notas de aulas sobre noções de estatística e probabilidade - São Paulo: USP.

Montgomery, D., Runger, G., 2016. Estatística Aplicada e Probabilidade para Engenheiros. Rio de Janeiro: LTC.