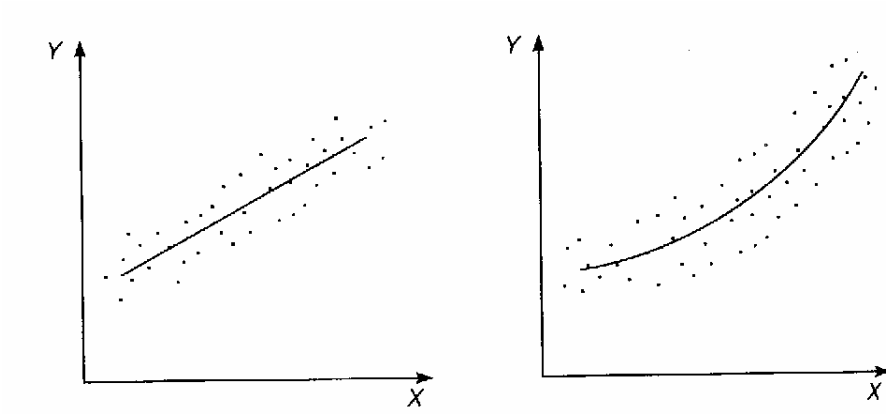
Prof. Paulo Ricardo B. Guimarães

Análise de Regressão

1. Introdução

Os modelos de regressão são largamente utilizados em diversas áreas do conhecimento, tais como: computação, administração, engenharias, biologia, agronomia, saúde, sociologia, etc. O principal objetivo desta técnica é obter uma equação que explique satisfatoriamente a relação entre uma variável resposta e uma ou mais variáveis explicativas, possibilitando fazer predição de valores da variável de interesse. Este relacionamento pode ser por uma equação linear ou uma função não linear, conforme figura abaixo:

Figura 1: formas lineares e não lineares de relação entre pares de variáveis



LINEAR NÃO LINEAR

2. Regressão linear simples

Se uma relação linear é válida para sumarizar a dependência observada entre duas variáveis quantitativas, então a equação que descreve esta relação é dada por:

$$Y = a + b\,X$$

Esta relação linear entre X e Y é determinística, ou seja, ela "afirma" que todos os pontos caem exatamente em cima da reta de regressão. No entanto este fato raramente irá ocorrer, ou seja, os valores observados não caem todos exatamente sobre esta linha reta. Existe uma diferença entre o valor observado e o valor fornecido pela equação. Esta diferença é denominada erro e é representada por ε, é uma variável aleatória que quantifica a falha do modelo em ajustar-se aos dados exatamente. Tal erro pode ser devido ao efeito, dentre outros, de variáveis não consideradas e de erros de medição. Incorporando esse erro à equação acima temos:

$$Y = a + bX + \varepsilon$$

que é denominado modelo de regressão linear simples. a e b são os parâmetros do modelo.

A variável X, denominada variável regressora, explicativa ou independente, é considerada uma variável controlada pelo pesquisador e medida com erro desprezível. Já Y, denominada variável resposta ou dependente, é considerada uma variável aleatória, isto é, existe uma distribuição de probabilidade para Y em cada valor possível de X. É muito freqüente, na prática, encontrarmos situações em que Y tenha distribuição Normal. Este é um dos principais pressupostos para aplicação desta técnica.

Exemplo 1: O preço de aluguel de automóveis de uma agência é definido pela seguinte equação: $Y = 8 + 0.15\ X$, onde Y = Taxa de aluguel (R$); X = distância percorrida (km).

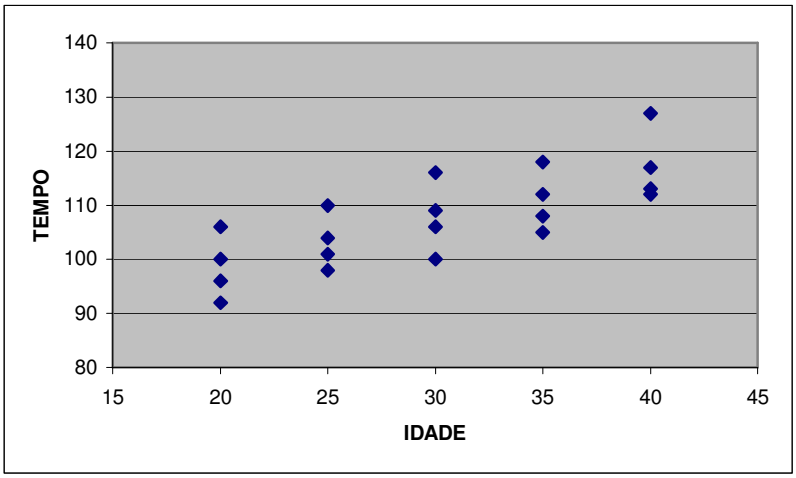
Assim, a taxa de aluguel inicia com o preço de R$ 8,00 e vai aumentando à medida que a distância percorrida aumenta. Assim, se fosse percorrida uma distância de 100 km, a taxa de aluguel seria de 8 + 0,15 x 100 = R$ 23,00. No entanto, como essa equação foi obtida baseada em dados de automóveis de diversas marcas certamente haverá uma variação no preço, por causa de diversos outros fatores. Assim, essa equação terá uma margem de erro, que é devida a esses inúmeros fatores que não foram controlados.

Exemplo 2: Um psicólogo investigando a relação entre o tempo que um indivíduo leva para reagir a um certo estímulo e sua idade obteve os seguintes resultados:

Tabela 1: Idade (em anos) e tempo de reação a um certo estímulo (em segundos)

Y - Tempo de reação (segundos)	X - Idade (em anos)
96	20
92	20
106	20
100	20
98	25
104	25
110	25
101	25
116	30
106	30
109	30
100	30
112	35
105	35
118	35
108	35
113	40
112	40
127	40
117	40

Figura 2: diagrama de dispersão entre a idade (X) e o tempo de reação (Y)



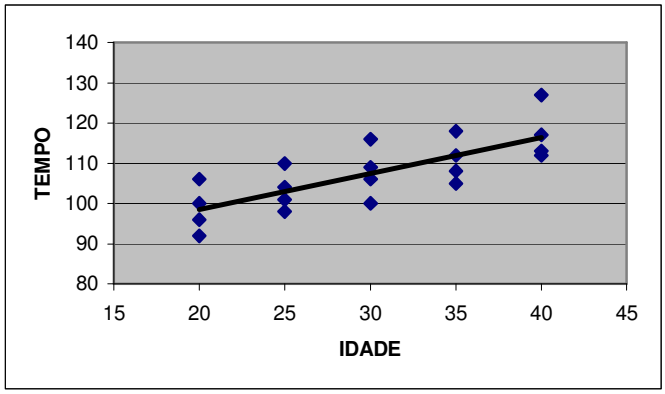
A partir da representação gráfica desses dados, mostrada na figura acima, é possível visualizar uma relação linear positiva entre a idade e o tempo de reação. O coeficiente de correlação de Pearson para esses dados resultou em r = 0,768, bem como seu respectivo teste de significância em t_{cal} = 5,09, que comparado ao valor tabelado $t_{tab,5\%}$ = 2,1 , fornece evidências de relação linear entre essas duas variáveis, ou seja, há evidências de considerável relação linear positiva entre idade e tempo de reação.

Podemos, então, usar um modelo de regressão linear simples para descrever essa relação. Para isso, é necessário estimar, com base na amostra observada, os parâmetros desconhecidos a e b deste modelo. O método de estimação denominado Mínimos Quadrados Ordinários (MQO) é freqüentemente utilizado em regressão linear para esta finalidade e será apresentado mais adiante.

Continuando a análise dos dados do exemplo, é possível obter o seguinte modelo de regressão linear simples ajustado:

$$Y = 80,5 + 0,9X$$

Figura 3: reta de regressão ajustada aos dados



Como a variação dos dados em X não inclui x = 0, não há interpretação prática do coeficiente a = 80,5. Por outro lado, b = 0,9 significa que a cada aumento de 1 ano na idade das pessoas, o tempo de reação médio (esperado) aumenta em 0,9 segundos.

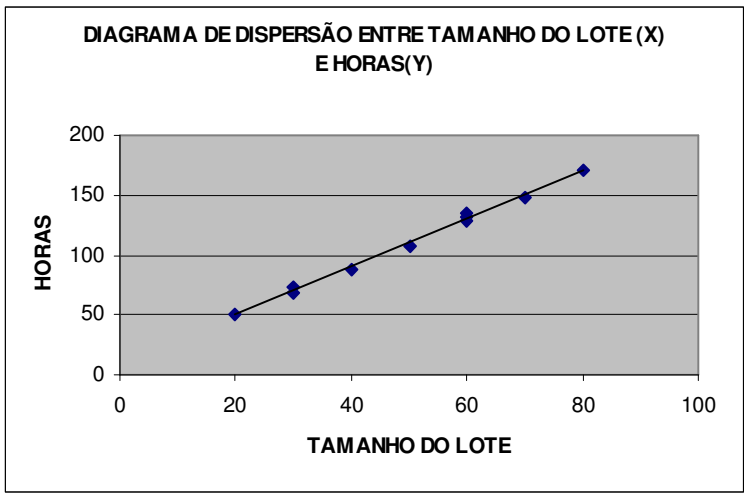
Assim, se: x = 20 anos, teremos y = 98,5 seg.
Para x = 21 anos, y = 99,4 seg.
x = 22 anos, y = 100,3 seg.
Assim, de ano para ano, o aumento no tempo de reação esperado é de 0,9 segundos.

Exemplo 3:
Uma certa peça é manufaturada por uma companhia, uma vez por mês, em lotes que variam de tamanho de acordo com as flutuações na demanda. A tabela abaixo contém dados sobre tamanho do lote e número de horas gastas na produção de 10 recentes lotes produzidos sob condições similares. Estes dados são apresentados graficamente na figura 4, tomando-se horas-homem como variável *dependente* ou variável *resposta* (Y) e o tamanho do lote como variável *independente* ou *preditora* (X).

Tabela 2 - Tamanho de lote e número de horas gastas na produção de cada lote.

Lote (i)	Horas (Y_i)	Tamanho do lote (X_i)
1	73	30
2	50	20
3	128	60
4	170	80
5	87	40
6	108	50
7	135	60
8	69	30
9	148	70
10	132	60

Figura 4 - Relação estatística entre Y e X, referente aos dados da tabela 2.



A figura sugere claramente que há uma relação linear positiva entre o tamanho do lote e o número de horas, de modo que, maiores lotes tendem a corresponder a maiores números de horas-homem consumidas. Porém, a relação não é perfeita, ou seja, há uma dispersão de pontos sugerindo que alguma variação no número de horas não é dependente do tamanho do lote. Por exemplo, dois lotes de 30 unidades (1 e 8) demandaram quantidades um pouco diferentes de horas. Na figura foi traçada uma linha (reta) de relacionamento descrevendo a relação estatística entre horas e tamanho do lote. Ela indica a tendência geral da variação em horas-homem quando há trocas no tamanho do lote.

Observa-se que grande parte dos pontos da figura não cai diretamente sobre a linha de relacionamento estatístico. A dispersão dos pontos em torno da linha de relacionamento representa a variação em horas que não é associada ao tamanho do lote, e que é usualmente considerada aleatória. Relações estatísticas são geralmente úteis, mesmo não tendo uma relação funcional exata.

3. Método dos mínimos quadrados ordinários (MQO)

Para estimar os parâmetros do modelo é necessário um método de estimação. O método estatístico utilizado e recomendado pela sua precisão, é o método dos mínimos quadrados que ajusta a melhor "equação" possível aos dados observados.

Com base nos n pares de observações (y_1 ,x_1) , (y_2,x_2) ,... , (y_n, x_n) , o método de estimação por MQO consiste em escolher a e b de modo que a soma dos quadrados dos erros, ε_i (i=1 ,..., n), seja mínima.

Para minimizar esta soma, que é expressa por:

$$SQ = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{I-1}^{n} \left(y_i - a - bX_i \right)$$

devemos, inicialmente, diferenciar a expressão com respeito a "a" e "b" e, em seguida, igualar a zero as expressões resultantes. Feito isso, e após algumas operações algébricas, os estimadores resultantes são:

$$b = \frac{\sum x_i y_i - n\bar{y}\bar{x}}{\sum x_i^2 - n\bar{x}^2}$$

$$a = \bar{y} - b\bar{x}$$

onde $\bar{y}$ é a média amostral dos y_i's e $\bar{x}$ a média amostral dos x_i's.

Logo, E(Y | x) = a + bx é o modelo de regressão linear simples ajustado, em que E(Y|x), denotado também $\hat{Y}$ por simplicidade, é o valor médio predito de Y para qualquer valor X = x que esteja na variação observada de X.

No exemplo 2, as estimativas dos parâmetros resultaram em a = 80,5 e b = 0,9. Veja como esses valores foram obtidos:

$\sum X_i = 600$ $\quad \sum Y_i = 2150$ $\quad$ n = 20 $\quad \sum X_i Y_i = 65400$

$\bar{X} = 30$ $\quad \bar{Y} = 107{,}5$ $\quad \sum X_i^2 = 19000$

$$b = \frac{\sum x_i y_i - n\bar{y}\bar{x}}{\sum x_i^2 - n\bar{x}^2} = \frac{65400 - 20.107{,}5.30}{19000 - 20.(30)^2} = \frac{900}{1000} = 0{,}9$$

$$a = \bar{y} - b\bar{x} = 107{,}5 - 0{,}9.30 = 80{,}5$$

No exemplo 3 as estimativas dos parâmetros a e b são:

$\sum X_i = 500$ $\quad \sum Y_i = 1100$ $\quad$ n = 10 $\quad \sum X_i Y_i = 61800$

$\bar{X} = 50$ $\quad \bar{Y} = 110$ $\quad \sum X_i^2 = 28400$

$$b = \frac{\sum x_i y_i - n\bar{y}\bar{x}}{\sum x_i^2 - n\bar{x}^2} = \frac{61800 - 10.110.50}{28400 - 10.(50)^2} = \frac{6800}{3400} = 2$$

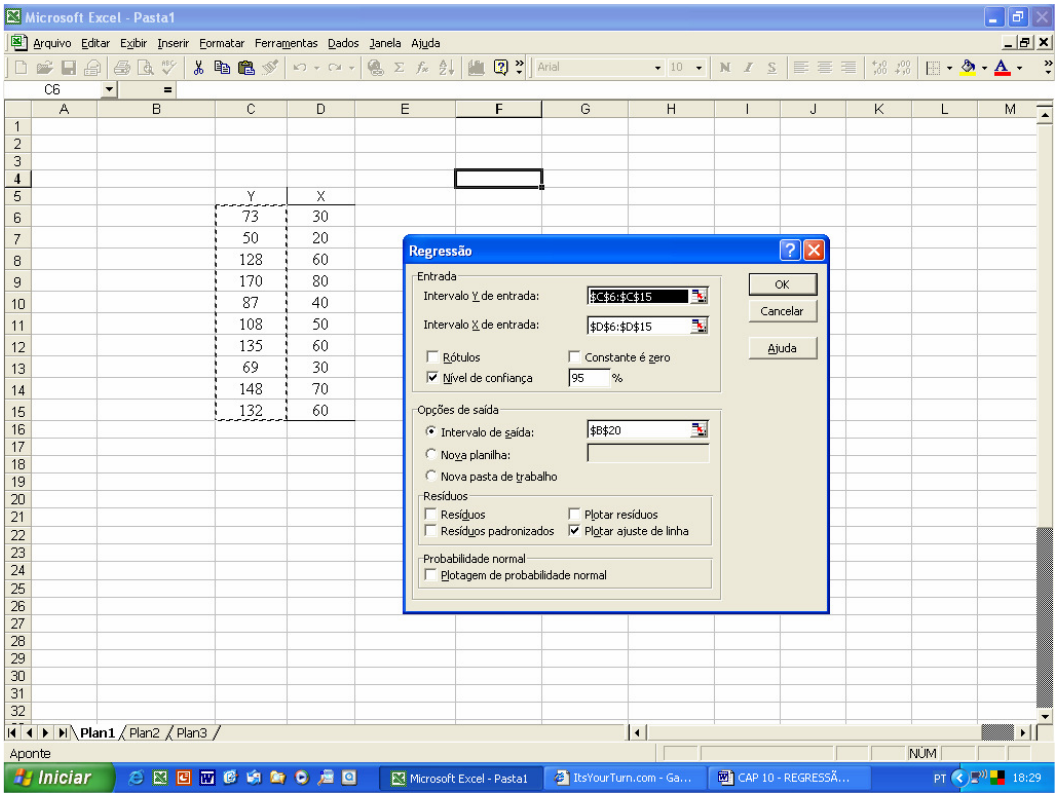
$$a = \bar{y} - \hat{\beta}_1\bar{x} = 110 - 2.50 = 10$$

Assim, a equação de regressão linear entre X e Y será dada por:

$$Y = 10 + 2X + \varepsilon$$

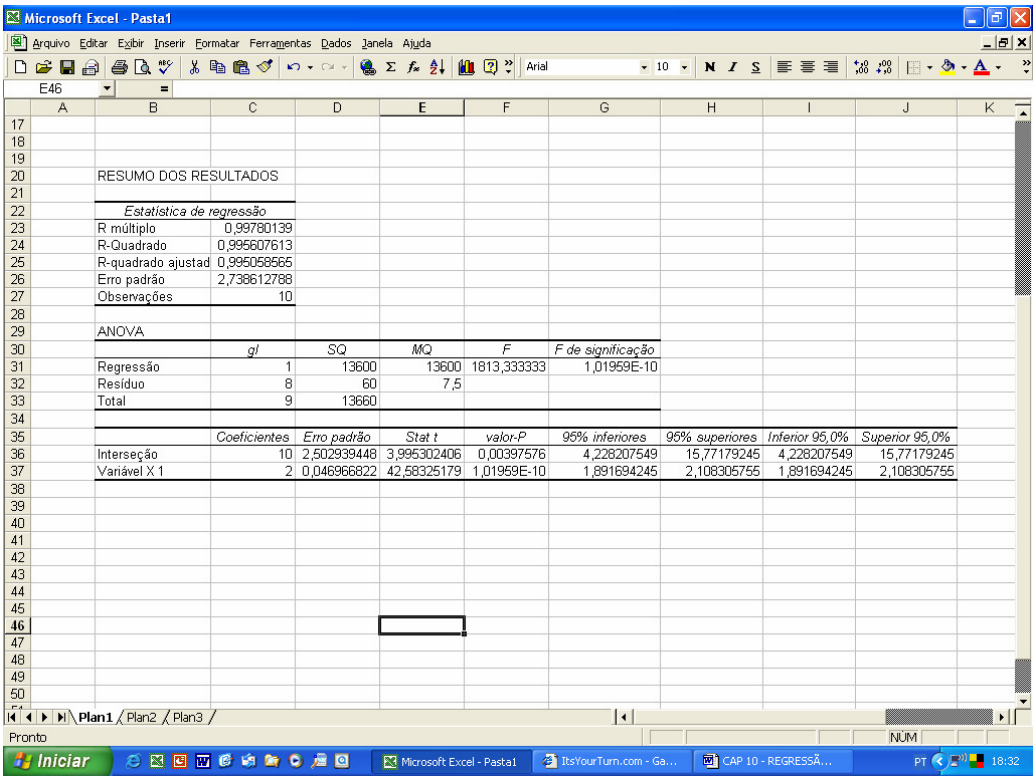
Interpretando o modelo acima, poderemos observar que, aumentando o tamanho do lote em uma unidade, o número de horas gastas na produção será aumentado de 2 horas.

Obtendo a reta de regressão com ajuda da planilha Excel, teremos que selecionar a opção REGRESSÃO no módulo de Análise de dados (em ferramentas):



Y	X
73	30
50	20
128	60
170	80
87	40
108	50
135	60
69	30
148	70
132	60

A saída fornecida pela planilha é a seguinte:



Observe que o Excel fornece, além dos coeficientes de correlação, a Anova da regressão para testar a sua significância e os coeficientes estimados com seus respectivos testes de significância.

4. Análise de Variância da Regressão

Para verificar a adequação do modelo aos dados, algumas técnicas podem ser utilizadas. A "análise de variância da Regressão" é uma das técnicas mais usadas. Assim, podemos analisar a adequação do modelo pela ANOVA da regressão a qual é geralmente apresentada como na tabela abaixo:

Fonte de Variação	g.l.	S.Q.	Q.M.	F	p-valor
Regressão	p-1	SQreg	SQreg/p-1	QMreg/QMres	
Resíduos	n-p	SQres	SQres/n-p		
Total	n-1	SQtotal	Sqtotal/n-1		

Onde:

- SQreg = soma dos quadrado devido à regressão:

$$SQreg = \sum_{i=1}^{n} (\hat{Y}_i - \bar{y})^2$$

- SQres = soma dos quadrado devido aos erros:

$$\text{SQres} = \text{SQtotal} - \text{Sqreg} = \sum_{i=1}^{n} (y_i - \widehat{Y}_i)^2$$

- SQtotal = soma dos quadrados totais:

$$\text{SQtotal} = \sum_{i=1}^{n} (y_i - \overline{y})^2$$

- p = número de variáveis do modelo
- n = numero de observações.

Caso o p-valor seja inferior ao nível de significância estabelecido então consideramos a regressão como significativa.

Uma maneira auxiliar de medir o "ganho" relativo introduzido pelo modelo é usar o coeficiente de determinação o qual é definido por R^2 que é calculado por SQreg/SQtotal.

Para os exemplos 2 e 3, a tabela da ANOVA seria construída de seguinte forma:

Exemplo 2:

$$\text{SQreg} = \sum_{i=1}^{n} (\widehat{Y}_i - \overline{y})^2 = \sum_{i=1}^{n} (80{,}5 + 0{,}9x_i - 107{,}5)^2 = 810$$

Para obter a soma de quadrados acima, deveremos substituir em x_i todos os valores de Idade da tabela 1.

$$\text{SQtotal} = \sum_{i=1}^{n} (y_i - \overline{y})^2 = \sum_{i=1}^{n} (y_i - 107{,}5)^2 = 1373$$

Para obter a soma de quadrados acima, deveremos substituir em y_i todos os valores de tempo de reação da tabela 1.

SQres = 1373 – 810 =563

Fonte de Variação	g.l.	S.Q.	Q.M.	F	p-valor
Regressão	1	810	810	25,90	<0,01
Resíduos	18	563	31,27		
Total	19	1373			

O que indica que a regressão entre X e Y é significativa. O modelo Y = 80,5 +0,9 X pode ser considerado adequado para realizar predições de Y. O coeficiente r^2 de determinação para esse modelo é de 0,59 o que representa um poder apenas razoável de explicação dos valores de tempo de reação pela Idade. Muito provavelmente outras variáveis estejam influenciando o tempo de reação.

Exemplo 3:

$$SQreg = \sum_{i=1}^{n} (\hat{Y}_i - \bar{y})^2 = \sum_{i=1}^{n} (10 + 2x_i - 110)^2 = 13600$$

Para obter a soma de quadrados acima, deveremos substituir em x_i todos os valores do tamanho do lote da tabela 2.

$$SQtotal = \sum_{i=1}^{n} (y_i - \bar{y})^2 = \sum_{i=1}^{n} (y_i - 107{,}5)^2 = 13660$$

Para obter a soma de quadrados acima, deveremos substituir em y_i todos os valores de número de horas gastas da tabela 2.

SQres = 13660 – 13600 = 60

Fonte de Variação	g.l.	S.Q.	Q.M.	F	p-valor
Regressão	1	13600	13600	1813,33	<0,01
Resíduos	8	60	7,5		
Total	9	13660			

O que indica que a regressão entre X e Y é significativa. O modelo Y = 10 + 2 X pode ser considerado de boa qualidade para realizar predições de Y. O coeficiente r^2 de determinação para esse modelo é de 0,996.

5. Erro padrão de estimação e intervalos de predição

O erro padrão da estimação é um desvio padrão condicional, na medida em que indica o desvio padrão da variável dependente Y, dado um valor específico da variável dependente X. O erro padrão baseado em dados amostrais é dado por:

$$\hat{\sigma}_u = \sqrt{\frac{\sum (y - \hat{Y})^2}{n - 2}}$$

Para fins de cálculo, é mais conveniente uma versão alternativa da fórmula:

$$\hat{\sigma}_u = \sqrt{S_y^2 \left(1 - r^2\right)}$$

onde $S_y^2 = \dfrac{\sum_{i=1}^{n} \left(y - \bar{y}\right)}{n}$

O erro padrão pode ser usado para estabelecer um intervalo de predição para a variável dependente, dado um valor específico da variável independente.

Uma vez que o erro padrão de estimação está baseado em dados de amostra, é apropriado o uso da distribuição t de Student com n-2 graus de liberdade. Assim, um intervalo de predição para a variável dependente Y, em análise de regressão simples é:

$$[\hat{Y} \pm t_{n-2;\alpha/2}.\hat{\sigma}_u]$$

Para os dados do exemplo 2 teríamos o erro padrão da estimação dado por:

Dado que $S_y^2 = 68{,}65$ e $r^2 = 0{,}5911$ então

$$\hat{\sigma}_u = \sqrt{S_y^2(1-r^2)} = \sqrt{68{,}65(1-0{,}5911^2)} = 6{,}683$$

E o intervalo de predição, com 95% de confiança, para um valor de Y=112 seria:

$$[\hat{Y} \pm t_{n-2;\alpha/2}.\hat{\sigma}_u] = [112 \pm 2{,}10 . 6{,}68] = [97{,}96 \; ; \; 126{,}03]$$

Ou seja, para uma pessoa com 35 anos, o tempo de reação predito estaria entre 97,96 e 126,03 segundos, com 95% de confiança.

Para os dados do exemplo 3 teríamos o erro padrão da estimação dado por:

Dado que $S_y^2 = 1366$ e $r^2 = 0{,}996$ então

$$\hat{\sigma}_u = \sqrt{S_y^2(1-r^2)} = \sqrt{1366(1-0{,}996^2)} = 3{,}3$$

E o intervalo de predição, com 95% de confiança, para um valor predito de Y= 110 seria:

$$[\hat{Y} \pm t_{n-2;\alpha/2}.\hat{\sigma}_u] = [110 \pm 2{,}31 . 3{,}3] = [102{,}37 \; ; \; 117{,}62]$$

Ou seja, para um lote de tamanho 50, seriam necessárias de 102,37 a 117,62 horas, com 95% de confiança.

6. Análise de Resíduos

Os desvios $e_i = y_i - \hat{y}_i$ (i = 1, ..., n) são denominados resíduos e são considerados uma amostra aleatória dos erros. Por este fato, uma análise gráfica dos resíduos é, em geral, realizada para verificar as suposições assumidas para os erros ε_i.

Para verificação dos pressupostos necessários para ajuste de um modelo de regressão é necessário realizar uma Análise de Resíduos. Os 3 tipos de resíduos mais comumente utilizados são:

- Resíduos brutos;
- Resíduos padronizados;
- Resíduos estudentizados.

7. Ampliando seus conhecimentos

Análise de Regressão Múltipla

A regressão múltipla envolve três ou mais variáveis, ou seja, uma única variável dependente, porém duas ou mais variáveis independentes (explicativas).

A finalidade das variáveis independentes adicionais é melhorar a capacidade de predição em confronto com a regressão linear simples. Mesmo quando estamos interessados no efeito de apenas uma das variáveis, é aconselhável incluir as outras capazes de afetar Y, efetuando uma análise de regressão múltipla, por 2 razões:

a) Para reduzir os resíduos. Reduzindo-se a variância residual (erro padrão da estimativa), aumenta a força dos testes de significância;
b) Para eliminar a tendenciosidade que poderia resultar se simplesmente ignorássemos uma variável que afeta Y substancialmente.

Uma estimativa é tendenciosa quando, por exemplo, numa pesquisa em que se deseja investigar a relação entre a aplicação de fertilizante e o volume de safra, atribuímos erroneamente ao fertilizante os efeitos do fertilizante mais a precipitação pluviométrica.

O ideal é obter o mais alto relacionamento explanatório com o mínimo de variáveis independentes, sobretudo em virtude do custo na obtenção de dados para muitas variáveis e também pela necessidade de observações adicionais para compensar a perda de graus de liberdade decorrente da introdução de mais variáveis independentes.

A equação da regressão múltipla tem a forma seguinte:

$$Y = a + b_1x_1 + b_2x_2 + \ldots + b_k x_k + e_i \text{ , onde:}$$

- a = intercepto do eixo y;
- b_i = coeficiente angular da i-ésima variável;
- k = número de variáveis independentes.

Enquanto uma regressão simples de duas variáveis resulta na equação de uma reta, um problema de três variáveis resulta um plano, e um problema de k variáveis resulta um hiperplano.

Também na regressão múltipla, as estimativas dos mínimos quadrados são obtidas pela escolha dos estimadores que minimizam a soma dos quadrados dos desvios entre os valores observados Y_i e os valores ajustados $\hat{Y}$.

Na regressão simples:
b = aumento em Y, decorrente de um aumento unitário em X.

Na regressão múltipla:

b_i = aumento em Y se X_i for aumentado de 1 unidade, mantendo-se constantes todas as demais variáveis X_j.

extraído de http://www.erudito.fea.usp.br/PortalFEA/

8. Atividades de Aplicação

1. Os encargos diários com o consumo de gás propano (Y) de uma empresa dependem da temperatura ambiente (X). A tabela seguinte apresenta o valor desses encargos em função da temperatura exterior:

Temperatura (^oC)	5	10	15	20	25
Encargos (dólares)	20	17	13	11	9

Seja $Y = \beta_0 + \beta_1 X + \varepsilon$ o correspondente modelo de regressão linear.

(a) Determine, usando o método dos mínimos quadrados, a respectiva reta de regressão e represente-a no diagrama de dispersão.
(b) Quantifique a qualidade do ajuste obtido e interprete.
(c) Determine um intervalo de confiança a 95% para os encargos médios com gás propano num dia em que a temperatura ambiente é de 17^oC.

2. Suponha que um analista toma uma amostra aleatória de 9 carregamentos feitos recentemente por caminhões de uma companhia. Para cada carregamento registra-se a distância percorrida em Km (X) e o respectivo tempo de entrega (Y). Obteve-se:

$$\sum x_i = 6405, \quad \sum y_i = 23.5, \quad \sum x_i^2 = 5628075, \quad \sum y_i^2 = 74.75, \quad \sum x_i y_i = 20295.$$

(a) Estime, usando o modelo de regressão linear, o tempo esperado de entrega para uma distância de 1050 Km.
(b) Comente a afirmação "o tempo de entrega é explicado em aproximadamente 94% pela distância percorrida".

3. Seja Y o número de chamadas telefônicas atendidas num determinado serviço de atendimento a clientes decorridos X minutos após as 8h30. Em determinado dia da semana observaram-se os seguintes pares de valores:

Tempo após 8h30 (min)	1	3	4	5	6
Número de chamadas atendidas	2	5	10	11	12

Seja $Y = \beta_0 + \beta_1 X + \varepsilon$ o correspondente modelo de regressão linear.

(a) Estime β_0 e β_1 usando o método dos mínimos quadrados e represente a correspondente reta de regressão no diagrama de dispersão.
(b) Determine o correspondente coeficiente de determinação, bem como o coeficiente de correlação; como interpreta os valores obtidos?
(c) Estime a variância do erro.
(d) Seja $E[Y(2)] = E[Y \mid x = 2]$. Estime $E[Y(2)]$; determine um intervalo de confiança para $E[Y(2)]$ com 95% de confiança.