

REGRESSÃO LINEAR SIMPLES. Marcos Alves dos Santos, Vilma Mayumi Tachibana. – Probabilidade e Estatística - Estatística - Departamento de Matemática, Estatística e Computação – Faculdade de Ciências e Tecnologia – Campus de Presidente Prudente.

Em muitos experimentos deseja-se investigar como uma mudança ocorrida em uma ou mais variáveis denominadas variáveis explicativas (ou variáveis independentes) afeta uma outra variável denominada variável resposta. Por exemplo, a quantidade procurada de um bem em um mercado pode ser considerada como função do seu preço. Para investigar esses e outros problemas foi estudada a regressão linear simples. A regressão linear simples estuda a relação entre duas variáveis quantitativas (ou qualitativas) de tal forma que uma variável pode ser “predita” a partir da outra variável de forma linear, ou seja, por uma linha reta.

O interesse da regressão linear simples é expressar estatisticamente esse relacionamento, para isso é utilizado o seguinte modelo:

$$Y = \beta_0 + \beta_1 X + \varepsilon,$$

em que Y é denominado variável resposta ou dependente e X é a variável regressora ou preditora, na qual ela pode ser prefixada ou aleatória. Neste trabalho foi estudado o caso em que a variável regressora é prefixada.

O modelo de regressão linear simples (MRLS) é representado pela soma de uma quantidade determinística e uma quantidade aleatória. A parte determinística do modelo de regressão linear simples é uma função linear com parâmetros β_0 e β_1 . O parâmetro β_0 é o ponto da reta que corta o eixo das ordenadas e o parâmetro β_1 é a variação média de Y a cada unidade da variável X . A quantidade aleatória do modelo, chamado de erro, é um erro aleatório cometido ao aproximar os pontos à reta de regressão.

Os parâmetros β_0 e β_1 são desconhecidos, mas podem ser estimados a partir de uma amostra aleatória. Um dos métodos mais utilizados de obtenção dos estimadores dos MRLS é o método de mínimos quadrados. A reta a ser representada pelo método de mínimos quadrados é a reta que possui a menor soma de quadrados da diferença entre a observação y_i e o valor estimado pela reta, cujos estimadores são:

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \quad \text{e} \quad \hat{\beta}_1 = \frac{\sum_{i=1}^n y_i (x_i - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{S_{xy}}{S_{xx}}$$

e o estimador de mínimos quadrados da variância é dado por:

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n - 2}$$

Montgomery et al. (2001) citam que pelo teorema de Gauss-Markov os estimadores de mínimos quadrados de β_0 e β_1 são estimadores não viciados de variância mínima, ou seja, são os melhores estimadores pontuais para β_0 e β_1 .

Os estimadores por intervalo são construídos com base em algumas suposições sobre o modelo de regressão, necessárias para inferências sobre a população de interesse. Uma das principais suposições sobre o modelo de regressão é que os erros tenham distribuição Normal com média 0 e variância σ^2 . Outras suposições são: os erros, ε s, são independentes e têm uma variância constante σ^2 (não depende da variável Y ou X), ou seja, $\varepsilon \sim N(0, \sigma^2)$ e $Cov(\varepsilon_i, \varepsilon_j) = 0$, $i \neq j$. Dessa forma, fixado $X_i = x_i$, as variáveis Y_i são independentes e $Y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$, $i = 1, \dots, n$.

Para verificar se o modelo ajustado de regressão é estatisticamente significativo realiza-se um teste de hipótese, comparando-o com um modelo simples, um modelo que não utiliza variável auxiliar,

isto é, $Y_i = \mu_y + \varepsilon_i$. O erro desse modelo alternativo, ou seja, o desvio de uma observação em relação à média é decomposto como o desvio da observação em relação ao valor ajustado pela regressão mais o desvio do valor ajustado em relação à média. Da dificuldade de comparar esses erros individualmente, Bussab e Morettin (2005) trabalham com as respectivas somas dos quadrados desses desvios e o teste pode ser resumido em uma tabela de análise de variância, denominada tabela de ANOVA.

A regressão não estaria completa sem verificar a adequação do MRLS. O coeficiente de determinação, comumente chamado de R^2 , é uma boa ferramenta para verificar o quanto o MRLS ou outro modelo de regressão qualquer explica o relacionamento dos dados. Este coeficiente é dado pela razão entre a soma de quadrados da regressão, que representa a variação em torno da reta de regressão, e a soma de quadrados total, que representa a variação total em torno da média.

O R^2 é fácil de ser calculado, mas segundo Bussab (1986), na situação em que para um número razoável de valores de x existirem duas ou mais observações para y , pode-se procurar ajustar o modelo mais simples $Y_i = \mu_{y_j} + \varepsilon_i$, em que μ_{y_j} representa a média para cada um dos níveis de x , fazendo-se o teste da falta de ajuste para testar a hipótese de adequação do modelo.

Quando o MRLS não é adequado, a solução é explorar outros modelos. Algumas situações indicam teoricamente que a reta de regressão deva passar pela origem, outras, porém, exigem modelos que expliquem o relacionamento entre duas variáveis, cujos estimadores são difíceis de serem obtidos, necessitando-se, muitas vezes recorrer a soluções numéricas. Uma primeira tentativa seria transformar o modelo em um de regressão linear simples, mas nem todos os modelos são linearizáveis.

Outro motivo que pode influenciar na inadequação do modelo é a existência de pontos influentes ou discrepantes. Nessa situação, recomenda-se usar regressão resistente que é baseada nas medianas.

Um dos maiores usos de um modelo de regressão é para obtenção do intervalo de confiança para a média \hat{y}_i com determinado nível da variável regressora e a predição para uma observação da variável resposta Y , que são construídos de maneira semelhante aos intervalo de confiança dos estimadores do MRLS.

A suposição de normalidade do erro é muito importante, por que os testes de hipótese e intervalos de confiança do MRLS foram construídos sob essa suposição. Não sendo válida, os testes de hipótese e intervalos de confiança se tornam ineficientes. Assim, a comprovação dessa suposição é feita pela análise de resíduos, principalmente pela forma gráfica.

Os dados da Tabela 1 apresentam o peso em kg de carros de uma certa categoria e a velocidade máxima em km/h que conseguem atingir durante o percurso.

Tabela 1: Dados referentes ao peso e a velocidade máxima de carros.

Peso (kg)	Velocidade máxima (km/h)			
720	257	263	266	
730	269	271		
740	275	274	278	
750	278	275	280	284
760	281	286	289	
770	293	297		
780	295	299	299	301
790	296	303	308	

Fonte: Charnet et al. (1999).

Sejam X_i : peso em kg do i -ésimo carro e Y_i : velocidade máxima (km/h) do i -ésimo carro. Da Tabela 1 foram obtidas as seguintes informações:

$$\begin{array}{lll}
 n = 24 & \sum_{i=1}^n x_i = 18150 & \sum_{i=1}^n x_i^2 = 13738300 \\
 \sum_{i=1}^n y_i = 6817 & \sum_{i=1}^n y_i^2 = 1940779 & \sum_{i=1}^n x_i y_i = 5162560
 \end{array}$$

O que implica que a reta de mínimos quadrados é $\hat{y} = -156,633 + 0,58271x$, com $\hat{\beta}_1 = 0,58271$, $\hat{\beta}_0 = -156,633$ e $\hat{\sigma}^2 = 12,24555$.

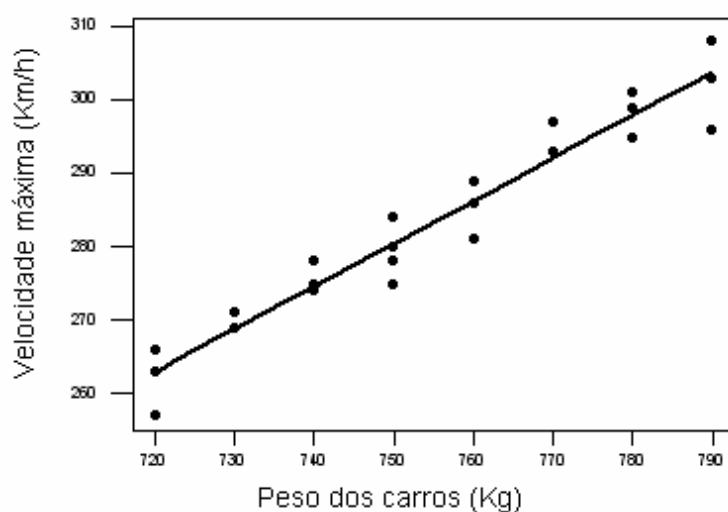


Figura 1: Ajuste do modelo de regressão.

O MRLS se ajustou bem aos dados como mostra a Figura 1. Mas isso não quer dizer que não exista outro modelo de regressão melhor que o MRLS. Para verificar estatisticamente se o modelo é adequado foi utilizado o teste da falta de ajuste, resumido na tabela de ANOVA, dada pela Tabela 2. Com p-valor de 0,4273 o modelo é adequado, cujo coeficiente de determinação foi de aproximadamente 0,94.

Tabela 2: Tabela de ANOVA.

Fonte (Fonte de variação)	GL (Graus de liberdade)	SQ (Soma de Quadrados)	QM (Quadrado médio)	F_0
Regressão	1	4197,556	4197,556	0,4273
Erro	22	269,402	12,24555	
(falta de ajuste)	6	41,652	6,942	
(erro puro)	14	227,75	16,2678	
Total	23	4466,958		

O modelo de regressão linear simples foi construído com base em suposições, caso não sejam verdadeiras, as previsões e intervalos de confiança seriam imprecisos. Pelo teste de normalidade de Anderson-Darling, os resíduos têm distribuição normal, com p-valor de 0,563 e não há indicio de heterocedasticidade. Portanto, as suposições sob o MRLS são verdadeiras.

Se um carro tem 755 Kg, então com 95% de confiança a velocidade do carro estaria entre o limite inferior de 275,908 e o limite superior de 290,719; para carros com mesmo peso e coeficiente de confiança de 95% o intervalo da velocidade seria de 281,830 a 284,797.

Pode-se concluir que a regressão linear simples é utilizada em várias áreas de atuação, como engenharia, ciências biológicas, ciências econômicas, entre outras. Isso pode ser verificado no exemplo apresentado neste resumo e no projeto.

O enfoque foi estudar o relacionamento linear entre duas variáveis, que é possível utilizando-se o modelo de regressão linear simples. Os recursos computacionais, tais como os *software* SAS e Minitab, são importantes ferramentas estatística para a análise de conjunto de dados. Esses *software* foram utilizados na resolução dos gráficos, fundamentais para a análise de resíduos.

Referências Bibliográficas

BUSSAB, W. O. *Análise de Variância e Regressão*. São Paulo: Atual, 1986. 147 p.

BUSSAB, W. O., MORETTIN, P. A. *Estatística Básica*. São Paulo: Saraiva, 2005. 526 p.

CHARNET, R., FREIRE, C. A. L., CHARNET, E. M. R., BONVINO, H. *Análise de modelos de regressão linear com aplicações*. São Paulo: Unicamp, 1999. 354p.

MONTGOMERY, D. C.; PECK, E.; VINING, G. G., *Introduction to linear regression analysis*. 3. ed. New York: Wiley, 2001. 641 p.