Universidade Estadual da Paraíba
Centro de Ciências e Tecnologia
Departamento de Estatística

Humberto Moreira de Almeida

Análise de regressão linear múltipla com estudo relacionado a horas de máquinas paradas na linha de produção de uma indústria de calçados

Campina Grande
Agosto 2014

Humberto Moreira de Almeida

Análise de regressão linear múltipla com estudo relacionado a horas de máquinas paradas na linha de produção de uma indústria de calçados

Trabalho de conclusão de curso apresentado ao curso de Bacharel em Estatística do Departamento de Estatística do centro de Ciências e Tecnologia da Universidade Estadual da Paraíba em cumprimento as exigências legais para obtenção do título de bacharel em Estatística

Orientador:

Gustavo Henrique Esteves

Campina Grande

Agosto 2014

A447a Almeida, Humberto Moreira de.
Análise de regressão linear múltipla com estudo relacionado a horas de máquinas paradas na linha de produção de uma indústria de calçados [manuscrito] / Humberto Moreira de Almeida. - 2014.
31 p. : il.

Digitado.
Trabalho de Conclusão de Curso (Graduação em Estatística) - Universidade Estadual da Paraíba, Centro de Ciências e Tecnologia, 2014.
"Orientação: Prof. Dr. Gustavo Henrique Esteves, Departamento de Estatística".

1. Análise de regressão. 2. Regressão linear. 3. Linha de produção. I. Título.

21. ed. CDD 519.5

Humberto Moreira de Almeida

Análise de regressão linear múltipla com estudo relacionado a horas de máquinas paradas na linha de produção de uma indústria de calçados

Trabalho de conclusão de curso apresentado ao curso de Bacharel em Estatística do Departamento de Estatística do centro de Ciências e Tecnologia da Universidade Estadual da Paraíba em cumprimento as exigências legais para obtenção do título de bacharel em Estatística

Aprovado em: 04/08/14

Banca Examinadora:

Gustavo H. S

Prof. Gustavo Henrique Esteves - Orientador
Universidade Estadual da Paraíba

Tiago Almeida de Oliveira

Prof. Tiago Almeida de Oliveira
Universidade Estadual da Paraíba

Ricardo Alves de Olinda

Prof. Ricardo Alves de Olinda
Universidade Estadual da Paraíba

Dedicatória

À Leide Targino Almeida, minha esposa, Antônio Camilo de Almeida e Francisca Moreira de Almeida, meus pais, que contribuíram direta e indiretamente para essa minha conquista.

Agradecimentos

Agradeço primeiramente ao meu senhor Deus por me guiar neste meu desafio e toda minha vida pessoal e profissional,

À minha querida esposa Leide Targino Almeida sempre muito compreensiva e meus filhos Leticia E. Targino Almeida e Hrek Ruan Targino Almeida por me dar a sustentação da família em todos os meus desafios.

Aos meus amigos de curso que contribuíram direta e indiretamente para que alcançasse meus objetivos nesta longa jornada de cinco anos de curso.

À empresa para qual eu trabalho por ter emprestado os livros que foram utilizados no meu curso.

E a todos os meus professores que sempre com muita humildade e paciência nos passaram todos os seus conhecimentos, pois sem eles nada teria sido possível.

Resumo

Neste trabalho foi estudada a análise de regressão linear múltipla, desde seu contexto histórico até um exemplo para aplicação da teoria. Para o estudo foi utilizado um banco de dados referente a registros de horas paradas da linha de produção de uma fábrica de calçados de Campina Grande-PB, durante o mês de Agosto de 2013. O objetivo é tentar explicar que tipo de problema/defeito tem gerado maior impactado na linha de produção ocasionando paradas de máquinas e gerando perda de produção. Os cálculos foram feitos com a ajuda do software estatístico R, versão 3.1.1 revelando as variáveis que estão relacionadas com a variável resposta.

Palavras-chave: Análise de Regressão, Horas de máquinas paradas, Testes de hipóteses.

Abstract

In this work we studied the multiple linear regression analysis since its historical context to an application example of the theory. For the study we used a database of records related to downtime of the production line of a shoe factory in Campina Grande-PB, during the month of August 2013. The main goal was to try to explain what kind of problem/defect has generated major impact on the production line causing machines stoppage and subsequent loss of production. Calculations were made using the R software, version 3.1.1, revealing the variables that are related to the response variable.

Keywords: Regression Analysis, Hours of idle machines, Hypotheses tests.

Sumário

Lista de Figuras

Lista de Tabelas

Lista de abreviaturas

ANOVA: Análise de variância

FV: Fonte de variação

Gl: Graus de liberdade

QM: Quadrados médios

SQ: Soma de quadrados

SQReg: Soma de quadrados de regressão

SQRes: Soma de quadrados de resíduos

SQTot: Soma de quadrados total

1 Introdução

A análise de regressão linear surgiu a partir da necessidade de cientistas descobrirem, através do cálculo de probabilidades, se algumas características físicas e psicológicas, entre membros de uma mesma família, poderiam estar associadas de forma que fosse possível explica-las através de um modelo matemático. Neste trabalho foi estudada a estimação dos parâmetros de regressão linear múltipla, o coeficiente de determinação, testes de hipóteses e somas de quadrados.

Para exemplificar os procedimentos teóricos e metodológicos da análise de regressão linear multipla, foi utilizado um banco de dados referentes a registros de horas paradas da linha de produção de uma fábrica de calçados de Campina Grande-PB durante o mês de Agosto de 2013. O objetivo é tentar explicar que tipo de problema/defeito tem gerado maior impacto na linha de produção ocasionando paradas de máquinas e gerando perda de produção.

Assim, o principal objetivo deste trabalho foi fazer um estudo minucioso da teoria da análise de regressão linear, com aplicação de um modelo de regressão linear múltipla para os dados coletados neste período.

As horas paradas foram registradas em livro de ocorrência conforme procedimento adotado por esta empresa, contendo 1,315 horas paradas no mês de agosto de 2013, que corresponde a 4,5% do total de 28,704 horas de trabalho disponível em cada mês considerando as 46 máquinas nos três turnos contendo todos os problemas que ocasionaram paradas de máquina gerando perda de produção. O *software* usado foi o programa estatístico R, na sua versão 3.1.1.

O estudo feito a partir destas informações traz resultados interessantes mostrando o turno os tipos de defeitos e as máquinas que tiveram maior tempo de paradas tendo como referência o mês de agosto de 2013 e que poderão ser utilizados pela gerência desta fábrica para evitar possíveis perdas de produção decorrentes destas paradas.

2 Fundamentação teórica

2.1 Origem histórica do termo Regressão

Francis Galton em 1886 verificou que, embora houvesse uma tendência de pais altos terem filhos altos e pais baixos terem filhos baixos, a altura média de filhos de pais de uma dada altura tendia a se deslocar ou "regredir" até a altura média da população como um todo. Em outras palavras, a altura dos filhos de pais extraordinariamente altos ou baixos tende a se mover para a altura média da população.

A lei de regressão universal de Galton foi confirmada por Karl Pearson em (1903), que coletou mais de mil registros das alturas dos membros de grupos de famílias e verificou que a altura média dos filhos de um grupo de pais altos era inferior a altura de seus pais, e que a altura média dos filhos de um grupo de pais baixos era superior a altura de seus pais.

Assim, tanto os filhos altos como baixos "regrediram" em direção a altura média de todos os homens (DEMÉTRIO e ZOCCHI, 2008).

Uma nova interpretação da regressão linear simples e múltipla

Estudo da dependência de uma variável explicativa (Y) que se tenha interesse em conhecer seu comportamento em relação a uma ou mais variáveis $(X_1, X_2 \cdots, X_k)$ com o objetivo de estimar e/ou prever valores para Y.

Desta forma, pode-se modelar e tentar descrever como as variáveis estão relacionadas entre si, devendo ter sempre a preocupação de criar modelos estatísticos que explicitem a estrutura do fenômeno em observação. Um dos métodos mais usados na estatística para investigar a relação entre variáveis é o modelo de regressão.

A regressão pode ser também definida como metodologia estatística que estuda (ou modela) a relação entre duas ou mais variáveis, conforme ilustrado pela Figura 1.

NAGHETTINI, M.; ANDRADE PINTO, E. J. de. **Hidrologia Estatística**. Belo Horizonte: [s.n.], 2007.

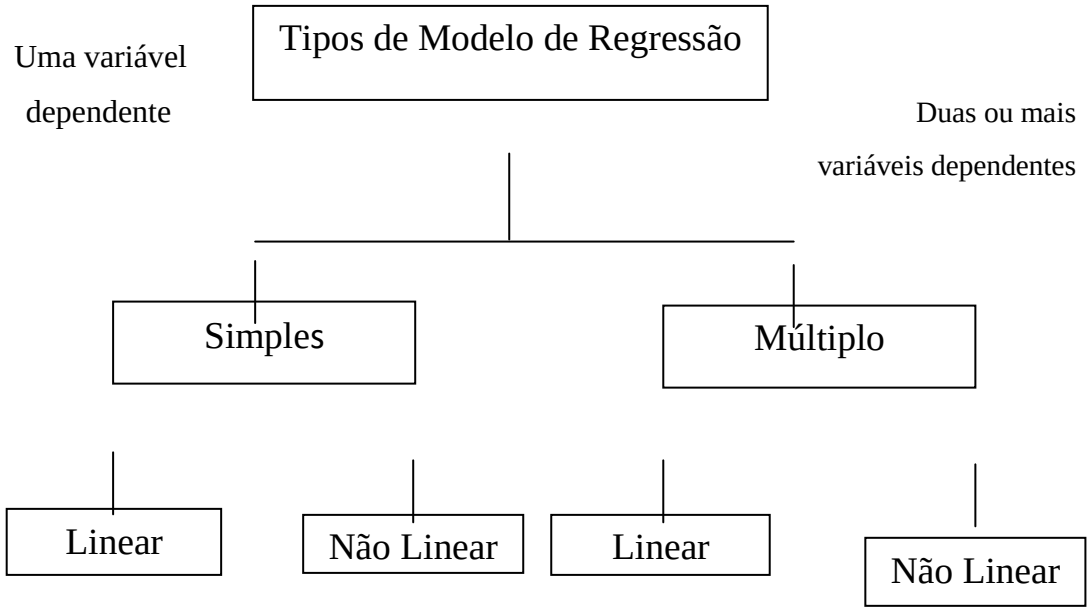

Figura 1: Fluxograma dos modelos de regressão.

SANTOS, A. M dos. Profa Alcione Miranda dos Santos. Departamento de Saúde Pública UFMA,(2007).

A utilização de modelos de regressão pode ter por objetivos:

i) Predição. Uma vez que se espera que uma parte (que se deseja que seja a maior) da variação de Y é explicada pelas variáveis $X_{j,}\ j = 1, 2, \cdots, k$, então, pode-se utilizar o modelo para obter valores de y correspondentes a valores de X_j.

ii) Seleção de variáveis. Frequentemente, não se tem ideia de quais são as variáveis que afetam significativamente a variação de Y. Para responder a esse tipo de questão, conduzem-se estudos onde está presente um grande número de variáveis. A análise de regressão pode auxiliar no processo de seleção de variáveis, eliminando aquelas cuja contribuição não seja importante.

iii) Estimação de parâmetros. Dado um modelo e um conjunto de dados (amostra) referente às variáveis resposta e preditoras, estimar parâmetros, ou ainda, ajustar o modelo aos dados, significa obter valores (estimativas) para os parâmetros, por algum processo, tendo por base o modelo e os dados observados.

iv) Inferência. O ajuste de um modelo de regressão tem, em geral, por objetivos básicos, além de estimar os parâmetros, realizar inferências sobre eles, tais como testes de hipóteses e intervalos de confiança.

2.2 Modelo de regressão linear simples

O modelo de regressão mais simples que pode ser definido é o modelo de regressão linear simples, dado pela expressão abaixo

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \qquad i = 1, \ldots, n,$$

Sendo:
y_i: valor da variável dependente (resposta) para o i-ésimo elemento da amostra;

x_i : valor (conhecido) da variável independente ou preditora para o *i*-ésimo elemento da amostra;

β_1 e β_2 : são parâmetros desconhecidos;
e_i: erro associado ao modelo.

Quanto à interpretação do modelo, o intercepto β_0 representa o ponto inicial de y quando o valor de x é igual a zero, x_i representa cada observação da variável explicativa x, o coeficiente angular β_1 representa o grau que a reta faz em relação ao eixo x e e_i é o erro associado a cada observação em relação à regressão linear.

2.3 Análise de Regressão Linear Múltipla

Tendo em vista que o problema deste trabalho envolve mais de uma variável. A partir deste momento a metodologia concentra-se no modelo de regressão linear múltipla, pois nos dá a condição de trabalhar com várias variáveis explicativas simultaneamente.

Para a definição do modelo de regressão linear múltipla, supõem-se que tem-se $X_{1,}\, X_2, \cdots, X_{P-1}$ variáveis preditoras e define-se como modelo de regressão linear múltipla, em termos destas variáveis preditoras, da seguinte forma:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_k x_{ki} + e_i, \qquad i = 1, 2, \ldots n.$$

Uma maneira de melhor visualizar o modelo de regressão linear múltipla é colocá-lo na sua forma matricial, dada por

$$\boldsymbol{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{e},$$

Em que:

$$\boldsymbol{y} = \begin{bmatrix} \mathbf{y_1} \\ \mathbf{y_2} \\ \vdots \\ \mathbf{y_n} \end{bmatrix}, \mathbf{X} = \begin{bmatrix} \mathbf{1} & \mathbf{x_{11}} & \mathbf{x_{21}} & \cdots & \mathbf{x_{k1}} \\ \mathbf{1} & \mathbf{x_{12}} & \mathbf{x_{22}} & \cdots & \mathbf{x_{k2}} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \mathbf{1} & \mathbf{x_{1n}} & \mathbf{x_{2n}} & \cdots & \mathbf{x_{kn}} \end{bmatrix}, \boldsymbol{\beta} = \begin{bmatrix} \boldsymbol{\beta_1} \\ \boldsymbol{\beta_2} \\ \vdots \\ \boldsymbol{\beta_k} \end{bmatrix}, \in = \begin{bmatrix} \boldsymbol{e_1} \\ \boldsymbol{e_2} \\ \vdots \\ \boldsymbol{e_n} \end{bmatrix}$$

Suposições sobre o modelo de regressão linear múltipla:

1- Existe relação linear entre y e $x_{\mathrm{j}}, \mathrm{j} = 1, 2, \cdots \mathrm{k}$,
2- Os valores dos x_{j} são fixos, logo, não são variáveis aleatórias;
3- As variáveis aleatórias e_i têm distribuição normal;
4- E $(e_i) = 0$, onde 0 representa o vetor nulo;
5- Var $(e_i) = \sigma^2$, para todos os valores de $i = 1,2, \dots, n$;
6- Os erros são não correlacionados dois a dois.

2.3.1 Estimação de parâmetros

De acordo com Morais (2010), para estimar os parâmetros do modelo de regressão múltiplo, é possível recorrer ao método dos mínimos quadrados, que permite encontrar uma reta que minimize a distância entre os pontos observados e a reta, fazendo, em média, a soma dos desvios quadráticos ser igual a zero.

Sejam β e e os vetores das estimativas e dos desvios do modelo, onde:

$$\boldsymbol{\beta} = \begin{bmatrix} \boldsymbol{\beta_1} \\ \boldsymbol{\beta_2} \\ \vdots \\ \boldsymbol{\beta_k} \end{bmatrix}, \boldsymbol{e} = \begin{bmatrix} \boldsymbol{e_1} \\ \boldsymbol{e_2} \\ \vdots \\ \boldsymbol{e_n} \end{bmatrix}$$

Temos que:

$$\boldsymbol{e} = \boldsymbol{y} - \mathbf{X}\boldsymbol{\beta}$$

A soma dos quadrados dos erros é dada por:

$$\boldsymbol{e^1\, e} \;=\; [\boldsymbol{e_1} \quad \boldsymbol{e_2} \quad \cdots \quad \boldsymbol{e_n}] \begin{bmatrix} \boldsymbol{e_1} \\ \boldsymbol{e_2} \\ \vdots \\ \boldsymbol{e_n} \end{bmatrix} \;=\; \boldsymbol{e_1\, e_1} \;+\; \boldsymbol{e_2\, e_2} \;+\; \cdots \;+\; \boldsymbol{e_n} \;=\; \sum_{\mathbf{i=1}}^{\mathbf{n}} \boldsymbol{e_i^2}.$$

Também pode-se mostrar que essa soma de quadrados ainda pode ser escrita como:

$$\mathbf{Z} = \boldsymbol{e'e} = \left(\mathbf{Y'} - \boldsymbol{\beta}'\mathbf{X'}\right)(\mathbf{Y} - \boldsymbol{\beta}\mathbf{X}) = \mathbf{Y'Y} - \mathbf{Y'X}\boldsymbol{\beta} - \boldsymbol{\beta}'\mathbf{X'Y} + \boldsymbol{\beta}'\mathbf{X}\boldsymbol{\beta},$$

como as matrizes $\boldsymbol{Y'X\beta - \beta'X'Y}$ são equivalentes, por uma ser a transposta da outra, então:

$$\mathbf{Z} = \mathbf{Y'Y} - \mathbf{2}\boldsymbol{\beta}'\mathbf{X'Y} + \boldsymbol{\beta}'\mathbf{X'X}\boldsymbol{\beta}.$$

A função Z deve ser igualada a zero para se obter o ponto de mínimo para os valores de β portanto:

$$\frac{\partial \mathbf{Z}}{\partial \boldsymbol{\beta}} = \partial\left(\mathbf{Y'Y} - \mathbf{2}\boldsymbol{\beta}'\mathbf{X'Y} + \mathbf{X'X}\boldsymbol{\beta}\right) = \mathbf{0}$$

$$\partial\left(\mathbf{Y'Y}\right) - \mathbf{2}\partial\left(\boldsymbol{\beta}'\mathbf{X'}\right)\mathbf{Y} + \left(\partial\boldsymbol{\beta}'\mathbf{X'}\right)\mathbf{X}\boldsymbol{\beta} = \mathbf{0}$$

$$-\mathbf{2}\left(\partial\boldsymbol{\beta}'\right)\mathbf{X'Y} + \left(\partial\boldsymbol{\beta}'\right)\mathbf{X'X}\boldsymbol{\beta} + \boldsymbol{\beta}'\mathbf{X'X}(\partial\boldsymbol{\beta}) = \mathbf{0}$$

Como $\left(\partial\boldsymbol{\beta}'\right)\mathbf{X'X}\boldsymbol{\beta} = \boldsymbol{\beta}'\mathbf{X'X}(\partial\boldsymbol{\beta})$ pois são matrizes simétricas com um único elemento, pode-se reescrever a equação acima de maneira que $-\mathbf{2}\left(\partial\boldsymbol{\beta}'\right)\mathbf{X'Y} + \mathbf{2}\left(\partial\boldsymbol{\beta}'\right)\mathbf{X'X}\boldsymbol{\beta} = \mathbf{0}$, e também:

$$\left(\partial\boldsymbol{\beta}'\left(\mathbf{X'X}\boldsymbol{\beta} - \mathbf{X'Y}\right) = \mathbf{0}\right),$$

para que $\partial\boldsymbol{\beta}' = \mathbf{0}$ é necessário que

$$\mathbf{X'X\beta = X'Y},$$

de onde se pode concluir que, a partir da matriz inversa de $\mathbf{X'X}$, pode-se chegar facilmente ao estimador dos parâmetros β, dados por

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X'X})^{-1}\mathbf{X'Y}.$$

2.3.2 Soma de quadrados e análise de variância da regressão linear Múltipla

Uma vez obtido o estimador para o vetor de parâmetros, deve-se observar a significância dos resultados obtidos. Para isso se faz necessário definir as somas de quadrados, que são descritas a seguir.

Soma de quadrados de resíduos é dada pela expressão:

$$\mathrm{SQRes} = \mathbf{Y'Y} - \widehat{\boldsymbol{\beta}}'\mathbf{X'Y}.$$

Já a soma de quadrados total é dada pela expressão:

$$\mathrm{SQTot} = \sum_{\mathbf{i=1}}^{\mathbf{n}} (\mathbf{y_i} - \hat{y})^{\mathbf{2}}.$$

E, por fim, a soma de quadrados da regressão é dada pela expressão:

$$\mathrm{SQReg} = \sum_{\mathbf{i=1}}^{\mathbf{n}} (\hat{y}_{\mathbf{i}} - \bar{y})^{\mathbf{2}}.$$

Sendo assim, após alguns cálculos relativamente simples, pode-se mostrar que a soma de quadrados de resíduos pode ser escrita como:

$$\mathrm{SQRes} = \mathrm{SQTot} - \mathrm{SQReg},$$

que é conhecida como a decomposição da soma de quadrados total.

Também é possível mostrar matematicamente que cada uma destas somas de quadrados segue distribuição de qui-quadrado, sendo que SQRes, SQReg e SQTot têm n-p, p-1 e n-1 graus de liberdade, respectivamente. A partir destes resultados é possível construir o quadro da análise de variância da regressão, dada pela Tabela 1, que é usada para calcular a estatística F, utilizada para verificar a validade do modelo estimado.

Através da estatística F pode-se testar as hipóteses $H_0: \beta_1 = \beta_2 = \cdots = \beta_k = 0$ versus $H_a: \text{pelo menos um } \beta_i \neq 0,\ i = 1,2,\ldots,k$. Ou seja, definimos o teste F para verificar se realmente existe uma relação linear entre Y e uma ou mais variáveis X_k, conforme é apresentado um pouco mais detalhadamente a seguir. Morais (2010),

Tabela 1: Esquema de análise de variância e estatística F para a regressão linear múltipla.

Causas de variação	G . L.	S. Q.	Q. M	F
Regressão Linear	k = p-1	SQReg	$\frac{SQReg}{p - 1}$	$\frac{QMReg}{QMRes}$
Resíduo	n - p	SQRes	$\frac{SQRes}{n - p}$	
Total	n – 1	SQTot	-	-

2.3.3 Coeficiente de determinação R^2

O coeficiente de determinação é uma estatística usada para medir a proporção da soma de quadrados que é explicada pela regressão linear múltipla. Este coeficiente pode ser obtido através da expressão:

$$R^2 = \frac{SQReg}{SQTot}.$$

2.3.4 Testes de hipóteses

Para verificar se existe regressão linear entre as variáveis do modelo é necessário que seja feito um teste de hipóteses, para o nosso caso o melhor teste a ser feito é o teste F, assim realiza-se o teste para dois casos.

Teste F para significância da equação de regressão linear múltipla

Aqui testa-se a existência da regressão linear no modelo, fazendo-se uso da estatística F citada na Tabela 1 vista anteriormente, que é:

$$F = \frac{QMReg}{QMRes},$$

utilizado para testar as hipóteses a seguir

$$\begin{cases} H_0: \beta_1 = \beta_2 = \beta_3 \cdots = \beta_k = 0 \\ H_1: \text{pelo menos um } \beta_j \neq 0,\ j = 1, 2, \ldots k \end{cases}$$

Se $F_{calculado}$ for maior que $F_{tabelado}$, onde F tem distribuição F de Snedecor com p-1 e n-p graus de liberdade, então rejeita-se H_0 e pode-se afirmar que, ao nível α de significância, pelo menos um $\beta_j \neq 0$, portanto, pode-se dizer que existe regressão linear entre as variáveis do modelo.

Teste F para as partes de um modelo de regressão linear

A influência de uma variável explicativa no modelo de regressão linear múltipla pode ser determinada pelo teste F parcial. Dessa forma, avalia-se a contribuição de uma variável explicativa para a soma dos quadrados devido à regressão, depois que todas as outras variáveis independentes foram incluídas no modelo. Assim, a influência desta variável x_k do modelo para a soma de quadrados da regressão será estimada pela diferença dada por:

$$SQReg(X_k) = SQReg_{(total)} - SQReg_{(total - x_k)}$$

As hipóteses que serão testadas são:

$$\begin{cases} H_0: \text{A variável } X_k \text{ não melhora significativamente o modelo} \\ H_1: \text{A variável } X_k \text{ melhora de forma significativa o modelo; } k = 1, 2, \cdots k \end{cases}$$

Uma forma que nos permite mostrar este teste é o $F_{calculado}$ dado pela expressão.

$$F_C = \frac{SQReg\ X_K}{QMRes}.$$

Teste t para significância de cada variável

O teste mais utilizado para medir a significância individual das variáveis do modelo e o teste t que nos mostra se existe alguma variável significativa no modelo. Assim a quantidade a ser testada para cada β_i será dada pela fórmula.

$$T_j = \frac{\hat{\beta}_j}{S(\hat{\beta}_j)}, \qquad j = 1, 2, \cdots, k,$$

Em que $S(\hat{\beta}_{i_i})$ representa o estimador da variância de β_i, e T_i segue distribuição t de Student com n-p graus de liberdade e que é usado para verificar as seguintes hipóteses

$$\begin{cases} H_0: \beta_j = 0 \\ H_1: \beta_j \neq 0, \qquad j = 1, 2, \cdots, k. \end{cases}$$

Portanto, se $|T_{calculado}| > T_{tabelado,}$ então rejeita-se H_0 e conclui-se que, ao nível α de significância, $\beta_j \neq 0$, $j = 1, 2, \cdots, n$ e assim pode-se afirmar que esta variável é importante para explicar a regressão linear. Caso contrário, esta variável não tem influência na regressão linear múltipla.

2.3.5 Análise de resíduos

Para uma confirmação a respeito das suposições necessárias para que haja regressão linear é fundamental fazer uma investigação no conjunto de dados para verificar a condição de normalidade através dos gráficos de probabilidade normal que são o PP-plot (Probabilidade acumulada esperada para a distribuição normal, em função

da probabilidade observada acumulada dos resíduos) e o QQ-plot (Quantil de probabilidade esperado para a distribuição normal, em função dos resíduos).

Após o esboço dos gráficos pode se verificar que, se os erros possuírem distribuição normal, os pontos devem estar mais ou menos alinhados sobre uma reta, caso contrário, os dados não apresentam indícios de normalidade.

Para melhor entendimento, podem-se observar os exemplos dados pelas Figuras 2 e 3 que seguem.

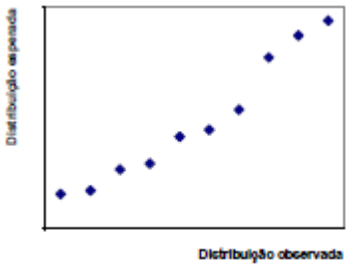

Figura 2: Gráfico PP-Plot, extraído de Morais (2010).

(Fonte: http://www.estv.ipv.pt/PaginasPessoais/psarabando/Estatistica CA 2009-2010/slides/regressão/Parte3 e Morais (2010).

A maioria dos pontos da Figura 1 concentram-se em torno de uma reta, o que dá indícios de normalidade dos dados.

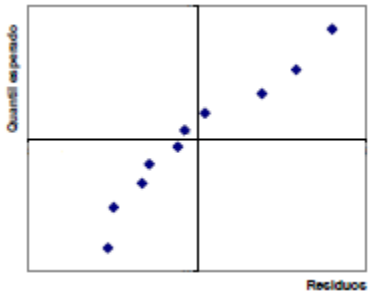

Figura 3: Gráfico QQ-plot, extraído de Morais (2010).

Na Figura 2, observa-se também que a maioria dos pontos está sobre uma reta, dando a entender que os erros seguem uma distribuição normal.

3 Aplicação

3.1 Banco de dados

Antes de qualquer estudo sobre o banco de dados, é necessário que seja feito um breve comentário sobre sua origem.

Os dados foram coletados da linha de produção de uma fábrica de calçados da região de Campina Grande em um dos setores produtivos onde existem 46 máquinas identificadas como prensa, as quais são numeradas de 1 a 46 e quando qualquer interferência gera perda de produção é registrado o tempo de parada em minutos e o motivo da parada em uma planilha do EXCEL contendo todos os motivos reais que causaram a perda de produção no total de 1.590 observações no mês de Agosto de 2013.

A variável de interesse é qual dos problemas citados na planilha gerou maior perda de produção em consequência do tempo que a máquina ficou parada.

De acordo com as informações registradas na planilha, conforme a Tabela 2, pode-se observar que no mês de Agosto foram 1.317 horas de maquinas paradas onde, 342h que representa 26,0% foi da turma A (06h00min às 14h00min), 379h que representa 28,8% foi da turma B (14h00min às 22h00min) e 594h que representa 45,2% foi da turma C (22h00min às 06h00min) distribuídos em 10 problemas, defeito mecânico, defeito elétrico, defeito elétrico e mecânico, varredura, falta de programação, falta de material, falta de sola, material com bolhas, material manchado e material atrasado. Na Tabela 3 constam os problemas que foram significativos após o ajuste do modelo de regressão e na Tabela 4 são apresentadas as máquinas que apresentaram valores significativos, também após o ajuste do modelo.

Tabela 2: resumo de horas de máquinas paradas por turma

Turma	Total de Horas	%
A	342	26,0
B	379	28,8
C	594	45,2
Total Geral	**1.317**	**100**

Tabela 3: resumo de horas paradas por problemas

Problema (código)	Total de Horas	% de horas
Defeito mecânico (1120)	268	20,5
Defeito elétrico (1121)	77	5,9
Falta de programação (2321)	17	1,3
Falta de material (2310)	838	64,0
Falta de sola (2322)	77	5,9
Material c / bolhas (3620)	32	2,4
Total Geral	**1309**	**100,0**

Tabela 4: resumo de horas paradas por máquina

Máquina	Total de Horas	% de horas
7	12	0,9
8	11	0,8
19	16	1,2
20	17	1,3
23	13	1,0
24	12	0,9
36	42	3,2
37	16	1,2
39	39	3,0

3.2 Análise de regressão

Para dar inicio no estudo de regressão linear entre as variáveis do banco de dados, devemos definir a variável de interesse ou variável resposta.

Conforme os dados coletados no setor de produção da fábrica no mês de agosto de 2013 deseja-se saber quais fatores tiveram contribuição para explicar o tempo que a máquina ficou parada gerando perda de produção, para isto será definida a variável resposta como sendo tempo da *i*-ésima parada.

As variáveis testadas para validar o modelo de regressão linear foram o tempo de máquina parada (variável resposta), as máquinas, os turnos de trabalho e os problemas que ocasionaram a parada.

O modelo de regressão linear múltipla a ser testado será dado por:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + e_i,$$

sendo y_i o tempo da parada do *i*-ésimo problema, x_{1i} está associado às máquinas, x_{2i} aos turnos de trabalho e x_{3i} aos problemas que ocasionaram as paradas. $\in_i$ é o erro para cada observação e β_j os parâmetros do modelo de regressão linear com $j = 1, 2, 3, \cdots, 1.590$.

Para verificar se existe regressão entre as variáveis x_{3i} do modelo, foi gerada a tabela de ANOVA, que está representada na Tabela 5. Nela pode-se observar as somas de (Q.M), (S.Q) os graus de liberdade (G.L), quadrados médios de regressão e de resíduos, com a inclusão do p-valor do teste F na última coluna.

Tabela 5: Análise da variância da variável x_{3i} do modelo

FV	SQ	GL	QM	F	P-valor
Regressão	312,279	9	34,698	18,7336	0,0001
Resíduo	292,826	1,581	1,852		
Total	605,100	1,590			

Observando-se a Tabela 5, pode-se concluir que, devido o P-valor do teste ser de 0,0001 é altamente significativo, ou seja, ao nível de 10%, 5% e até 1% de significância há fortes indícios contra H_0, ou seja, pode-se considerar que haja regressão linear para pelo menos uma variável tipo de defeito no modelo.

De acordo com os cálculos efetuados pelo software foram estimados os valores dos coeficientes de regressão linear para este modelo conforme Tabela 6, que das 46 máquinas somente 6 apresentaram valores significativos e a Tabela 7 mostra que a turma C foi quem apresentou maior significância, já na Tabela 8 tem-se os 5 problemas que foram significativos. Sendo o problema "falta de material" com maior significância, pois apareceu com 290 minutos a mais que o problema defeito mecânico com 59,7 minutos parado.

Tabela 6: Coeficientes de regressão linear do modelo, para as categorias da variável x_1 comparativamente à média do erro tipo máquina 1.

β_j	Estimativas	t	P - valor
08	- 28,0358	- 2,6530	0,0080
20	25,4524	-3,0150	0,0026
23	- 28,2820	- 2,7330	0,00063
24	- 29,1467	- 2,7150	0,0066
36	25,3804	2,9090	0,0036
39	55,2570	5,2400	<0,0001

Tabela7: Coeficientes de regressão linear do modelo, para as categorias da variável $x_2$comparativamente à média do erro turma A.

β_j	Estimativas	t	P - valor
B	7,8188	2, 6750	0,0001
C	13,9944	4,1960	< 0,0001

Tabela 8: Coeficientes de regressão linear do modelo, para as categorias da variável x_3, comparativamente à média do erro tipo 1120.

β_j	Estimativas	t	P - valor
1121	- 23,6882	- 4,3050	< 0,0001
2310	290,7743	11,1340	< 0,0001
2321	- 18,8383	- 5,2201	< 0,0001
2322	- 28,6199	- 5,5190	< 0,0001
3620	- 24,3600	- 3,5470	0,0004

4 Conclusão

Os modelos de regressão linear são bastante utilizados para explicar a associação entre duas ou mais variáveis. Esta técnica estatística nos permite escrever uma variável em função de outras variáveis independentes desde que estejam correlacionadas, podendo assim, explicar seu comportamento de acordo com valores estabelecidos para cada variável independente.

O uso da regressão feita neste trabalho nos mostrou a importância desta técnica estatística para modelar problemas do cotidiano nos mostrando onde possa haver uma relação causa- consequência.

No estudado realizado, deseja-se responder qual das variáveis em estudo tem maior impacto no resultado final de produção deste setor de trabalho em consequência do tempo em que a máquina ficou parada.

Conforme resultado dos dados coletados neste setor no mês de Agosto de 2013, pode-se concluir que, o turno com maior incidência de paradas é o turno C, enquanto que dos 10 problemas mencionados no banco de dados somente 6 foram significativos para explicar as perdas de produção devido o tempo em que as máquinas ficaram paradas mas dos 6 tiveram 2 com maior frequência o problema falta de material tem maior significância no total de horas parada em seguida vem defeito mecânico.

Entre as máquinas que apresentaram valores significativos a prensa 39 pode ser desconsiderada, pois neste período estava passando por uma alteração de manual para semiautomática, logo esta alteração teve um grande impacto no tempo parado desta máquina as demais poderão ser feito um estudo mais aprofundado por cada máquina.

Referências

NAGHETTINI, M.; ANDRADE PINTO, E. J. de. **Hidrologia Estatística**. Belo Horizonte: [s.n.], 2007.

DEMÉTRIO, C. G. B.; ZOCCHI, S. S. Clarice G.B. Demétrio & Silvio S. Zocchi

SANTOS, A. M dos. Profa Alcione Miranda dos Santos**. Departamento de Saúde Pública** UFMA,(2007).

MORAIS, N. F. **Análise de regressão linear com estudo de caso em acidentes de trânsito**. Monografia de TCC. Universidade Estadual da Paraíba: Campina Grande-PB, 2010.