Sistema FIEB | SENAI CIMATEC
PELO FUTURO DA INOVAÇÃO

Estudo estatístico
Análise de Regressão

Anderson Queiroz
Rodrigo Formiga

Senai Cimatec. CCRoSA - Centro de Competência em Robótica e Sitemas Autônomos
anderson.vale@fbter.org.br
rodrigo.farias@fbter.org.br

22 de setembro de 2020

1 Revisão literária

Serão apresentados na presente seção alguns pontos que, segundo a literatura, nortearão o estudo estatístico do trabalho em questão, abordando alguns pontos importantes para realização do planejamento e análise dos experimentos. Logo em seguida, serão aplicados os conceitos aqui apresentados.

1.1 Planejamento e análise de experimentos

Técnicas de planejamento e análises de experimentos são utilizáveis em empresa, processo de fabricação, desenvolvimento de um novo produto e até em fabricação de commodities. Geralmente a maioria destes processos possuem diversas variáveis controláveis, como temperatura, velocidade, dimensões entre outras. A variação destas variáveis, de modo isolado ou combinado, pode trazer melhorias ou prejuízos a qualidade do produto ou serviço. Com isso, faz-se necessário um estudo analítico de como cada variável pode influenciar o processo [1].

O planejamento e análise de experimentos (*Design of Experiments,* ***DOE***) é uma técnica de planejamento de experimentos, ou seja, como os experimentos devem proceder para garantir conclusões assertivas. A validade das conclusões obtidas após análise dos experimentos, são diretamente ligadas ao modo como os experimentos foram realizados, logo, o planejamento do experimento torna-se o ponto mais importante na hora de se analisar algum processo produtivo, bens ou serviços prestados. O planejamento corretamente executado também pode evitar desperdícios, além de isolar e determinar relações das variáveis [1].

A seguir serão apresentados alguns pontos importantes para análise estatística realizada neste presente trabalho, como: Princípios do planejamento, que irão nortear como deve ser condizido o planejamento de experimentos, bem como etapas para realização dos experimentos e análise dos mesmos [2].

1.1.1 Princípios do planejamento de experimentos

O planejamento de experimentos possui três principais princípios básicos, que são: Replicação, aleatóriedade e blocagem [2].

Replicação

A replicação consiste na obtenção de mais de uma unidade experimental, ou seja, replicar o teste para cada ponto experimental e assim permite que obtenha-se uma estimativa mais precisa, diminuindo a influência de variações indesejadas ou inevitáveis [2].

Aleatóriedade

Os métodos estatísticos requerem que as observações sejam variáveis aleatórias, de modo a garantir a distribuição igual dos fatores não esperados. Com isso garantem-se estimativas não tendênciosas dos efeitos e erros experimentais, bem como evitar influência sistemática de fatores não controláveis [2].

Blocagem

A blocagem é uma técnica extremamente importante, utilizada com o objetivo de aumentar a precisão de um experimento. Este princípio pode ser aplicado em alguns casos onde deseja-se isolar algum fator conhecido, analisando de forma individual sua influência sobre o processo. A mudança de pessoas no processo experimental ou a mudança de lote de um produto pode ser visto a princípio como um fator não desejado de ser analisado, logo, o princípio de blocagem permite que o mesmo seja isolado, porém, não ignorado[2].

1.1.2 Etapas para o desenvolvimento de experimentos

Coleman e Montgomery (1993) propõem as seguintes etapas para o desenvolvimento de um planejamento de experimentos [3]:

1. **Caracterização do problema:** A definição do problema que está sendo analisado é uma etapa essencial para se entender o estudo analítico. Para isso se faz necessário conhecimento sobre todo o processo para assim definir de forma clara o objetivo e relatar de forma específica o problema analisado.

2. **Escolha dos fatores de influência e níveis:** Para conduzir o experimento deve-se escolher os fatores variáveis, os intervalos sobre os quais esses fatores variarão e os níveis específicos. Para tanto é necessário conhecimento do processo, experiências práticas e teóricas para determinar tais fatores mais relevantes para análise experimental.

3. **Seleção das variáveis de resposta:** É também necessário ao planejar um experimento que se tenha clareza de qual variável-resposta será obtida como parâmetro para qualificar os experimentos. Variável de resposta que não atenda a necessidade do experimento pode levar a conclusões equivocadas.

4. **Determinação de um modelo de planejamento de experimento:** A escolha do planejamento envolve consideração sobre o tamanho da amostra (número de replicações) e determinar se há formação de blocos ou outras restrições de aleatorização.

5. **Condução do experimento:** A condução do experimento deve seguir todo o planejamento, evitando variações de ambiente, métodologia do experimento ou inserção de novas variações não previamente estabelecidas. O responsável pelo experimento e os equipamentos utilizados devem ser mantidos do início ao fim, exceto em casos onde estas variações sejam desejáveis.

6. **Análise dos dados:**

   Após todas as demais etapas anteriores, os resultados podem ser obtidos com uso de ferramentas e pacotes estatísticos para auxiliar na visualização de gráficos e dados. A verificação da validade do modelo é também um ponto importante para análisar.

7. **Conclusões e recomendações:**

   Após a obtenção dos resultados, o experimento deve apresentar conclusões práticas para proporcionar recomendações de ações em cima do processo analisado. As recomendações são frutos das etapas decorridas no planejamento do experimento, ou seja, planejamentos com falhas iram gerar ações equivocadas no processo.

1.2 Análise de regressão

Um método muito utilizado nos estudos estatísticos é a análise de regressão. Este método permite examinar a relação entre duas ou mais variáveis que se deseja-se estudar. O método permite por meio de modelos matemáticos avaliar quais as relações entre as variáveis, quais delas são importantes para o processo e quais apresentam pouca relevância [4].

As variáveis utilizadas podem ser divididas em duas: Variável Dependente, que é a variável que está sendo estudada, como uma sáida do processo (a influência das demais variáveis será analisada através das variáveis dependentes); e Variável Independente, que são as entradas do estudo, são variáveis que supostamente causam impactos ou certa influência na variável dependente [4].

A análise de regressão tem por finalidade chegar a algumas conclusões e direcionamentos sobre o processo estudado, as finalidades deste estudo são [4]:

- Predição dos dados;
- Seleção de variáveis influenciáveis no processo;
- Estimação de parâmetros;
- Realizar inferências sobre os parâmetros, como: testes de hipóteses e intervalos de confiança.

1.2.1 Tipos de regressão

A análise de regressão é dividida em alguns tipos, tendo cada uma sua própria especificidade, logo, todo analista de deve saber qual forma usar, variando sua escolha pelo tipo de dado e sua distribuição. Os tipos mais comuns são: Regressão Linear; Regressão Polinomial; Regressão de Poisson; Regressão de Ridge; Regressão Logística e Mínimos quadrados parciais (PLS) [5]. Para fins deste estudo será abordado de forma mais aprofundada a regressão linear.

Regressão linear

A regressão Linear é um modelo matemático que tem por objetivo observar a relação entre duas ou mais variáveis por meio de uma reta, e utilizar o resultado da função dessa reta para estimar valores e encontrar relações. Quando existe apenas uma variável independente e uma variável dependente, ela é chamada de **regressão linear simples**. Quando existe mais de uma variável independente, é chamado de **regressão linear múltipla**. Na equação de regressão linear múltipla, apresentada a seguir, o 'Y' é a variável dependente, 'X' são as variáveis independentes, β_i são os coeficientes de regressão e ϵ é o termo de erro[5].

$$Y = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + ... + \beta_i X_i + \epsilon \tag{1}$$

Como exemplo de uma análise de regressão linear no R (figura 1), será analisado os dados a seguir e entender as informações contidas no estudo.

Figura 1: Exemplo de regressão linear no R [5]

```
     (Intercept)      Agriculture      Examination        Education         Catholic
      66.9151817       -0.1721140       -0.2580082       -0.8709401        0.1041153
Infant.Mortality
       1.0770481
> summary(model)

Call:
lm(formula = Fertility ~ ., data = swiss)

Residuals:
     Min       1Q   Median       3Q      Max
-15.2743  -5.2617   0.5032   4.1198  15.3213

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)      66.91518   10.70604   6.250 1.91e-07 ***
Agriculture      -0.17211    0.07030  -2.448  0.01873 *
Examination      -0.25801    0.25388  -1.016  0.31546
Education        -0.87094    0.18303  -4.758 2.43e-05 ***
Catholic          0.10412    0.03526   2.953  0.00519 **
Infant.Mortality  1.07705    0.38172   2.822  0.00734 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7.165 on 41 degrees of freedom
Multiple R-squared:  0.7067, Adjusted R-squared:  0.671
F-statistic: 19.76 on 5 and 41 DF,  p-value: 5.594e-10
```

R múltiplo

Esse dado serve para medir a relação linear entre as variáveis dependente e independentes, ou seja, mede o quanto elas estão correlacionadas. O valor desejável é proxímo de 1 (100%) [5].

R-Quadrado

O R-quadrado é uma medida estatística de quanto os dados estão próximos em linha de regressão. Ele também é conhecido como coeficiente de determinação múltipla para a regressão múltipla. Este valor vaira de 0-1, onde o 0(zero) indica que o modelo não serve para explicar a variabilidade dos dados, enquanto 1(um) indica que o modelo consegue explicar toda variabilidade dos dados de resposta ao redor da média[5].

Coeficientes

São os valores que serão multiplicados pelas variáveis independentes para obter o valor esperado da variável dependente. Valores próximos de 1(um) indicam que a variável independente analisada interfere fortemente na variável de saída[5].

P-value

Servem para avaliar a hipótese núla. Em valores inferiores a 0.05 é rejeitado a hipótese núla e a variável não deve ser descartada ou ignorada para o estudo. No caso da figura 1, a variável "Education"apresenta p-value muito pequeno e menor que 0.05, logo, esta variável influencia fortemente a variável de saída[5].

2 Resultados

A partir dos testes realizados sob orientação do planejamento, anteriormente dito na sessão X, foram coletadas os resultados do tempo de vôo do helicóptero. As amostras dos testes coletados, como pode ser visto na tabela 1, possui os seguintes parâmetros: clip que pode ser com ou sem; AT que representa o adesivo no topo do helicóptero, sendo com ou sem; ADLat é o adesivo lateral, podendo ser colocado o adesivo do lado direito ou do lado esquerdo; AL é a altura de partida do helicóptero, em que foi definido como 1,30 metros e 2,10 metros e o tempo de queda em segundos. Dessa forma, com base nestes valores de tempo, a partir das disposições dos demais parâmetros, foi feito o teste de regressão para analisar a influência da relação entre os parâmetros com o tempo.

Tabela 1: Resultados das amostras coletadas.

	clip	AT	ADLat	AL	tempo
1	S	S	E	1.30	0.92
2	C	S	E	1.30	0.88
3	S	C	E	1.30	1.04
4	C	C	E	1.30	1.10
5	S	S	D	1.30	1.10
6	C	S	D	1.30	0.99
7	S	C	D	1.30	1.07
8	C	C	D	1.30	0.92
9	S	S	E	2.10	1.76
10	C	S	E	2.10	1.23
11	S	C	E	2.10	1.88
12	C	C	E	2.10	1.52
13	S	S	D	2.10	1.70
14	C	S	D	2.10	1.72
15	S	C	D	2.10	1.46
16	C	C	D	2.10	1.42

Para analise de regressão deste sistema foi divido em 2 modelos: primeiro e segunda ordem. Como o sistema possui 4 variáveis independentes, o modelo máximo atingido pode ser representado pela relação entre os 4 parâmetros. Porém, a partir do terceiro modelo não há mais significância entre os resultados obtidos, que será mostrado em diante, descartando assim a análise com mais relações.

2.1 Análise de Primeira Ordem - Modelo 1

A partir da análise linear aplicada aos resultados das amostras, aplicando a primeira ordem, o resultado é mostrado na tabela 2. O primeiro modelo dessa análise pode ser visualizada na equação 2. Esta equação mostra um modelo que consegue explicar o valor do tempo a partir da relação entre os parâmetros. Os valores dentro dos parâmetros são as variáveis que mais apresentaram valores significantes ao sistema, dentre as opções estabelecidas.

Tabela 2: Resultado da análise linear de primeira ordem do modelo 1.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	1.0622	0.0910	11.67	1.54e-07
clipC	-0.1419	0.0814	-1.74	0.1092
ATC	0.0119	0.0814	0.15	0.8866
ADLatD	0.0081	0.0814	0.10	0.9223
AL2.10	0.5844	0.0814	7.18	1.80e-05

$$\begin{aligned} tempo = 1.062187 - 0.141875.clip(C) + 0.011875.AT(C) \\ + 0.008125.ADLat(D) + 0.584375.AL(2.10) \end{aligned} \tag{2}$$

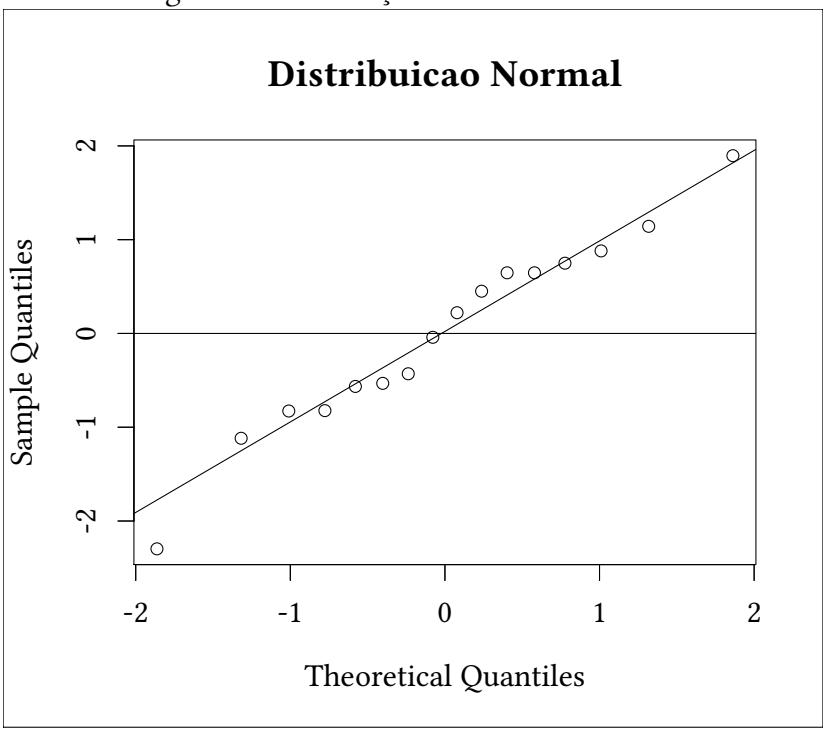
Com base na tabela 2, pode ser visto que estatisticamente os estimadores dos parâmetros clip, AT e ADLat são iguais a zero, pois, se assumir o nível de significância a 5% de probabilidade, o ρ valor desses coeficientes é superior a 0.05, aceitando a hipótese nula, ou seja, as estimativas desses parâmetros são números muitos próximos de zero no qual no teste t afirma que esses valores aceitam a hipótese H0. Portanto, se os parâmetros clip, AT e ADLat são estatisticamente iguais a zero, eles não precisam estar no modelo pois não vai ter nenhuma contribuição significante no valor de tempo de queda do helicóptero. Então, refazendo a análise linear do tempo de queda com relação apenas ao parâmetro AL, o novo modelo pode ser representado pela equação 3. O resultado dos coeficientes pode ser visto na tabela 3.

Tabela 3: Análise linear da relação do tempo com AL.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	1.0012	0.0577	17.35	7.29e-11
AL2.10	0.5844	0.0816	7.16	4.84e-06

$$tempo = 1.00125 + 0.58438.AL(2.10) \tag{3}$$

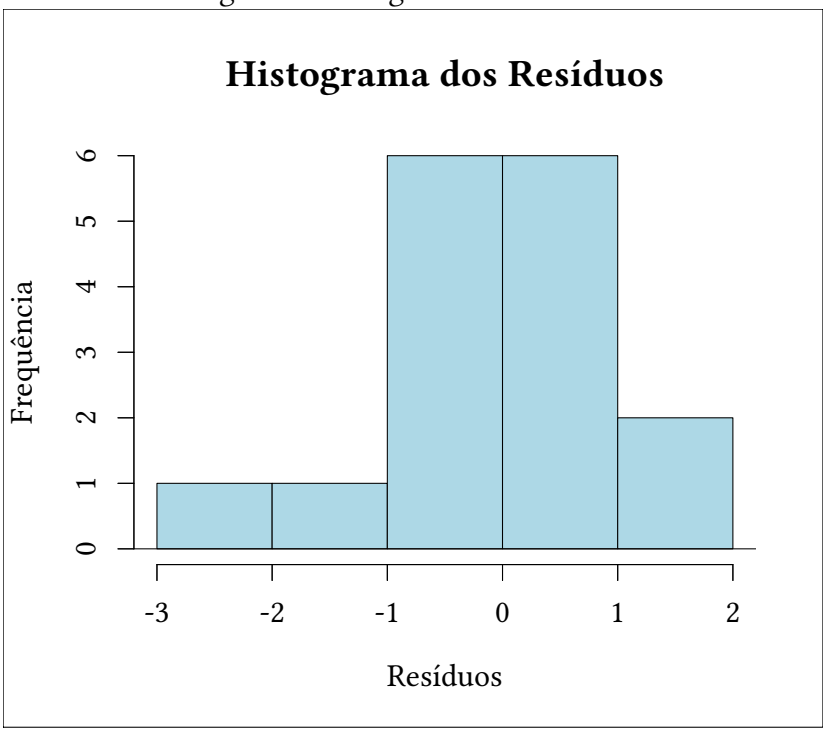
Figura 2: Distribuição normal dos resíduos.



O modelo apresentado na equação 2, que representa uma regressão múltipla, foi reduzido para uma regressão simples, representado pela equação 3. Desse modo, para explicar a equação reduzida pode-se dizer que para cada acréscimo a partir da altura de 2,10 metros, o tempo de queda aumenta em 0,58 segundos. Com base no valor do R^2, que foi de 78,56%, pode se afirmar a porcentagem do dados que são explicado pelo modelo da equação 3. O valor do R^2 ajustado será desconsiderado pois esse último modelo foi o único em que todos os estimadores dos parâmetros foram significativo, no caso dessa análise linear do modelo 1.

Analisando os resíduos do modelo 1 simplificado, foi aplicado o teste de normalidade de *Shapiro-Wilk*. Nesse teste pode-se afirmar que, a partir da verificação dos resíduos desse modelo mostrou-se que no teste da normalidade apresentou um valor da análise de 0.97528 com o ρ valor respectivo de 0,9152, adotando o nível de significância de 5%, não houve violação da normalidade dos resíduos. Então, pode-se concluir que esses erros (resíduos) tem uma distribuição normal, como pode ser visto nos gráficos das figuras 2 e 3. Dessa forma, assumisse a independência dos resíduos pelo fato que a estrutura de coleta declara essa independência. Portanto, não é preciso atribuir testes de independência.

Figura 3: Histograma dos resíduos.



2.2 Análise de Segunda Ordem - Modelo 2

No modelo 2 foi feito a análise linear de segunda ordem. O resultado dessa análise pode ser visto na tabela 4. A partir dos valores da tabela 4 foi calculado o modelo 2, representado pela equação 4.

$$\begin{aligned} tempo = {} & 0.93906 - 0.15.clip(C) + 0.21375.AT(C) + 0.1425.ADLat(D) \\ & + 0.74875.AL(2.10) + 0.04125.clip(C).AT(C) \\ & + 0.14375.clip(C).ADLat(D) - 0.16875.clip(C).AL(2.10) \\ & - 0.34875.AT(C).ADLat(D) - 0.09625.AT(C).AL(2.10) \\ & - 0.06375.ADLat(D).AL(2.10) \end{aligned} \tag{4}$$

Essa equação do modelo 2 consegue explicar o valor do tempo a partir da relação de segunda ordem dos parâmetros. Como pode ser visto na tabela 4, o parâmetro que possui os estimadores mais significantes, se assumir o nível de significância de 5%, é novamente o AL, ou seja, estatisticamente os valores do ρ valor são maiores que 0.05 em todos exceto o AL2.10. Como isso considera a hipótese nula em que as estimativas desses coeficientes possuem valores muito próximo de zero, podendo também ser reti-

Tabela 4: Resultado da análise linear de segunda ordem.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.9203	0.0206	44.73	0.0142
clipC	-0.0506	0.0281	-1.80	0.3227
ATC	0.1094	0.0281	3.89	0.1602
ADLatD	0.1744	0.0281	6.20	0.1018
AL2.10	0.8344	0.0281	29.68	0.0214
clipC:ATC	0.1262	0.0368	3.43	0.1806
clipC:ADLatD	-0.0438	0.0368	-1.19	0.4453
clipC:AL2.10	-0.4638	0.0368	-12.60	0.0504
ATC:ADLatD	-0.1288	0.0368	-3.50	0.1773
ATC:AL2.10	0.0162	0.0368	0.44	0.7353
ADLatD:AL2.10	-0.2238	0.0368	-6.08	0.1038
clipC:ATC:ADLatD	-0.1925	0.0425	-4.53	0.1383
clipC:ATC:AL2.10	0.0225	0.0425	0.53	0.6900
clipC:ADLatD:AL2.10	0.5675	0.0425	13.35	0.0476
ATC:ADLatD:AL2.10	-0.2475	0.0425	-5.82	0.1083

rado da equação da relação de segunda ordem, pois, estatisticamente esses parâmetros não apresentam contribuição significativa no valor do tempo de queda do helicóptero. Desta maneira, simplificando a equação 4 do modelo 2 irá ficar da mesma forma que a equação 3, propiciando o encerramento da análise de regressão para o modelo 2.

Referências

[1] MONTGOMERY, Donald C.; RUNGER, George **Estatística Aplicada e Probabilidade Para Engenheiros.** LTC; 5ª ed., 2012.

[2] TAHARA, Sayuri **Melhores Práticas - Planejamento de Experimentos (DOE)** 2008. Link: http://www.portaldeconhecimentos.org.br/index.php/por/Conteudo/Planejamento-de-Experimentos-DOE. Acessado em 17/09/2020 ás 13:30.

[3] COLEMAN, D. E.; MONTGOMERY, D. C. **A systematic approach to planning for a designed industrial experiment.**.Technometrics, v. 35, 1993.

[4] Portal Action **ANÁLISE DE REGRESSÃO**. Link: http://www.portalaction.com.br/analise-de-regressao. Acessado em 21/09/2020 ás 17:30.

[5] MARIA, Júlia **15 tipos de regressão mais frequentes**. Link: https://rpubs.com/JulhinhaM/395633. Acessado em 22/09/2020 ás 8:30.