



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

Nyedja Fialho Morais

Análise de regressão linear com estudo de caso em acidentes de trânsito

Campina Grande
Dezembro de 2010

Nyedja Fialho Morais

Análise de regressão linear com estudo de caso em acidentes de trânsito

Trabalho de conclusão de curso apresentado ao curso de Bacharelado em Estatística do Departamento de Estatística do Centro de Ciências e Tecnologia da Universidade Estadual da Paraíba em cumprimento às exigências legais para obtenção do título de bacharel em Estatística.

Orientador:
Gustavo H. Esteves, Dr.

Campina Grande
Dezembro de 2010

FICHA CATALOGRAFICA ELABORADA PELA BIBLIOTECA CENTRAL – UEPB

M827a Morais, Nyedja Fialho.
Análise de regressão linear com estudo de caso em acidentes de trânsito [manuscrito] / Nyedja Fialho Morais. – 2010.
46 f.: il. color.

Digitado.

Trabalho de Conclusão de Curso (Graduação em Estatística) – Universidade Estadual da Paraíba, Centro de Ciências e Tecnologias, 2010.

“Orientação: Prof. Dr. Gustavo Henrique Esteves, Departamento de Estatística”.

1. Estatística. 2. Acidente de Transito. 3. Regressão Linear. I. Título.

21. ed. CDD 519.5

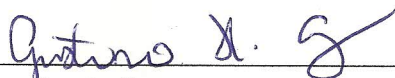
Nyedja Fialho Morais

Análise de regressão linear com estudo de caso em acidentes de trânsito

Trabalho de conclusão de curso apresentado ao curso de Bacharelado em Estatística do Departamento de Estatística do Centro de Ciências e Tecnologia da Universidade Estadual da Paraíba em cumprimento às exigências legais para obtenção do título de bacharel em Estatística.

Aprovado em: 06/12/10

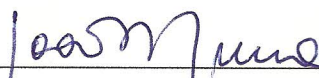
Banca Examinadora:



Prof. Gustavo H. Esteves, Dr.
Orientador



Profa. Diana Maia
Universidade Estadual da Paraíba



Prof. João Gil de Luna
Universidade Estadual da Paraíba

Dedicatória

À Maria da Salette Fialho Morais e Pedro de Morais Filho (In Memoriam), meus pais,
que contribuíram de forma significativa para que eu pudesse chegar até aqui.

Agradecimentos

Agradeço aos meus pais, que deram todo o seu amor e toda uma base cristã e ética, para que eu pudesse crescer em segurança, me ensinando a nunca desistir dos meus objetivos, e a ir em busca dos meus ideais com sabedoria, perseverança e humildade.

Aos meus irmãos, Érika Fialho Morais, que sempre esteve ao meu lado e foi minha grande companheira durante estes quatro anos de curso, e Pedro de Morais Júnior, que sempre me motivou a buscar a realização dos meus sonhos.

Ao meu namorado, Edson Américo, por me ajudar em todas as coisas, pelo seu amor e compreensão para comigo.

Aos meus amigos, pelo companheirismo e alegria que me proporcionaram.

Aos meus queridos professores pela dedicação com que transmitiram seus conhecimentos.

Agradeço a todos que me incentivaram a ir em busca da realização dos meus sonhos.

Mas o agradecimento mais especial vai para Jesus Cristo, meu Deus e meu tudo. A Ele toda a honra, toda a glória e todo louvor!

Resumo

Neste trabalho foi estudado a análise de regressão linear múltipla, desde seu contexto histórico até um exemplo para aplicação da teoria. Para o estudo foram utilizados dados verídicos extraídos dos boletins de ocorrência de acidentes de trânsito ocorridos na cidade de Campina Grande-PB durante o ano de 2009. O objetivo foi tentar, através da análise de regressão linear, explicar o número de vítimas envolvidas em acidentes de trânsito no perímetro urbano da localidade estudada. Os cálculos foram feitos com a ajuda do software estatístico SPSS, revelando as variáveis que estão relacionadas com a variável resposta.

Palavras-chave: Análise de Regressão Linear, Acidentes de trânsito, Estatística.

Abstract

In this work we studied the multiple linear, since its historical remarks until an application example to illustrate the theory. In this study we used real data obtained from accident reports occurred during the year of 2009 in Campina Grande town. The main objective was to try, using linear regression, to explain the variable number of victims from traffic accidents inside urban area at the studied city. The calculations was dane using SPSS statistical software, revealing the variables related with the response variable.

Key-words: Linear Regression analysis, Traffic Accidents, Statistics.

Sumário

Lista de Figuras

Lista de Tabelas

Lista de abreviaturas

1	Introdução	p. 13
2	Fundamentação Teórica	p. 14
2.1	Marco Histórico	p. 14
2.2	Análise de regressão linear simples	p. 16
2.2.1	Estimação de parâmetros	p. 17
2.2.2	Decomposição da soma de quadrados	p. 19
2.2.3	Coefficiente de determinação	p. 19
2.2.4	Testes de hipóteses	p. 20
2.2.5	Intervalo de confiança	p. 22
2.3	Análise de regressão linear múltipla	p. 22
2.3.1	Estimação de parâmetros	p. 23
2.3.2	Soma de quadrados e análise de variância da regressão linear múltipla	p. 27
2.3.3	Coefficiente de determinação	p. 30
2.3.4	Teste de hipóteses	p. 31
2.3.4.1	Teste F para significância da equação de regressão linear múltipla	p. 31

2.3.4.2	Teste F para as partes de um modelo de regressão linear múltipla	p. 31
2.3.4.3	Teste t para significância dos parâmetros	p. 32
2.3.5	Intervalo de confiança	p. 32
2.3.6	Regressões que se tornam lineares através de anamorfose	p. 35
2.3.7	Análise de resíduos	p. 35
3	Aplicação	p. 38
3.1	Banco de dados	p. 38
3.2	Análise de regressão	p. 39
4	Conclusão	p. 43
	Referências	p. 46

Lista de Figuras

1	Reta de regressão para a hereditariedade de Galton para pais e filhos	p. 16
2	Gráfico P P Plot	p. 36
3	Gráfico Q Q plot	p. 36
4	Exemplos de comportamento dos gráficos de resíduos	p. 37
5	Resíduos com variância não constante	p. 37

Lista de Tabelas

1	Análise de variância no caso simples	p. 21
2	Análise de variância no caso múltiplo	p. 30
3	Análise de variância do modelo	p. 40
4	Coefficientes de regressão linear do modelo	p. 41
5	Intervalo de confiança para os coeficientes de regressão linear do modelo	p. 42

Lista de abreviaturas

ANOVA: Análise de variância

BO: Boletim de Ocorrência

FV: Fonte de variação

GI: Graus de liberdade

MMQ: Método dos mínimos quadrados

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 quando cientistas tentavam descobrir, através do cálculo de probabilidades, se algumas características físicas e psicológicas, entre pais e filhos, poderiam estar associadas a tal ponto que se fosse possível explicá-las matematicamente através de uma função.

Neste trabalho foi estudado a estimação dos parâmetros de regressão, a decomposição das somas de quadrados, o coeficiente de determinação, ou explicação, além de testes de hipóteses e construção de intervalos de confiança para os coeficientes de regressão.

Para exemplificar os procedimentos teóricos e metodológicos da análise de regressão linear, foi utilizado um banco de dados referentes a registros de acidentes de trânsito ocorridos na cidade de Campina Grande-PB durante o ano de 2009. O objetivo é tentar explicar o número de vítimas de acidentes de trânsito em decorrência do perfil dos condutores, como idade, sexo e sobriedade.

Assim, o principal objetivo deste trabalho foi fazer um estudo minucioso da teoria da análise de regressão linear, com o intuito de aplicar um modelo de regressão múltipla para os dados em questão.

Os Boletins de Ocorrência para cada um dos acidentes foram registrados pela Superintendência de Trânsito e Transportes Públicos, STTP, Companhia de Policiamento de Trânsito, CPTRAN, Serviço de Atendimento Móvel de Urgência SAMU, sendo organizados no Software estatístico SPSS versão 18, contendo 3.486 observações e 66 variáveis, contendo informações do espaço físico, cronológico e perfil dos condutores e vítimas envolvidas nos acidentes de trânsito, sendo a maioria das variáveis do tipo categórica.

O estudo feito a partir deste banco de dados traz resultados interessantes que poderiam ser utilizados pelas políticas públicas de segurança, para que o número de vítimas decorrentes de acidentes de trânsito pudesse diminuir.

2 Fundamentação Teórica

2.1 Marco Histórico

Antes de qualquer coisa, é preciso entender como surgiram os primeiros estudos sobre a análise de regressão, portanto será retomada uma parte da história que tanto contribuiu e ainda contribui para o desenvolvimento da Estatística.

Através da necessidade de se desvendar os mistérios da hereditariedade, Mendel (1822-1884) chegou a um estudo matemático-probabilístico para explicar as características das ervilhas. O fato não foi muito aceito pelos seus contemporâneos, pois seus estudos envolviam cálculos complexos para seu tempo.

Outros pesquisadores também deram sua contribuição ao estudo da eugenia, concluindo que os caracteres herdados de uma geração para a outra o faziam por intermédio de fatores particulares que ocorriam aos pares (MANTIOLLI, 2001).

O pesquisador Francis Galton (1822-1911), a partir de um estudo com pares pais-filhos, propôs a **Lei de Regressão para mediocridade**.

Para Mendel, a partir dos resultados dos cruzamentos com ervilhas verdes e amarelas, percebia-se que algumas características eram dominantes sobre outras, e tudo era explicado com cálculos probabilísticos. Para Galton, as características se apresentavam em pares, e podiam ser representadas através de uma reta onde se teriam valores observados ao redor da média esperada.

Mesmo sem saber ao certo como se dava o mecanismo de transmissão das características, Galton sabia que podia comprovar suas crenças através de análises estatísticas dos registros de características de pais e filhos.

Ele se dispôs então a coletar os dados de interesse. Elaborou questionários para agregar as informações primordiais para o seu trabalho. No questionário tinham perguntas a respeito do antepassado da família, como por exemplo, características físicas e intelectuais.

Francis Galton chegou a oferecer dinheiro para quem quisesse participar da pesquisa. Em seus panfletos promocionais expunha o objetivo de seus estudos da seguinte forma:

1. Para uso daqueles que desejam ser medidos de diversas maneiras com exatidão, e também para conhecer a tempo defeitos remediáveis do desenvolvimento, e para conhecer os próprios poderes.
2. Para guardar um registro metódico das principais medidas de cada pessoa, do qual poderá, com algumas restrições razoáveis, obter no futuro uma cópia. Colocando suas iniciais e data de nascimento, mas não o seu nome. Os mesmos serão registrados em livro à parte.
3. Para obter informações sobre os métodos, práticas e usos das medidas humanas.
4. Para experimentação e investigação antropométricas e para obter dados para discussão estatística.

Ele conseguiu coletar 9000 registros familiares e como não possuía nenhum auxílio computacional naquela época, levou cerca de dez anos para analisar todos os dados.

Para Galton, a análise tanto das características fisiológicas quanto dos talentos, através da utilização de ferramentas estatísticas, revelaria que a frequência com que eram mantidas nas sucessivas gerações, em alguns casos, uma verdadeira dinastia de talentos, não poderia ser apenas uma bela coincidência ou obra do acaso, mas sim a evidência de uma regularidade natural ou biológica. (CONT, 2008)

Quando se analisava, por exemplo, as alturas dos indivíduos em uma população, percebia-se uma constante de regressão à média, indicando que os indivíduos em seus extremos deixaram descendências que tendiam ao valor médio. Essas disposições não estariam sujeitas às condições ambientais, tais como nutrição, clima, geografia, sendo, portanto, o resultado da herança de caracteres inatos, ou seja, transmitidos sem sofrerem influência das condições externas. (CONT, 2008)

Galton sabia que seus estudos possuíam um estrutura que permitia que se realizassem uma análise matemática para explicar o comportamento de seus dados.

Para dar continuidade e estrutura institucional às pretensões eugênicas galtonianas, liderados por Francis Galton e Karl Pearson, no final do século XIX, formou-se um grupo de cientistas conhecidos como biometristas. Esse grupo era constituído de evolucionistas

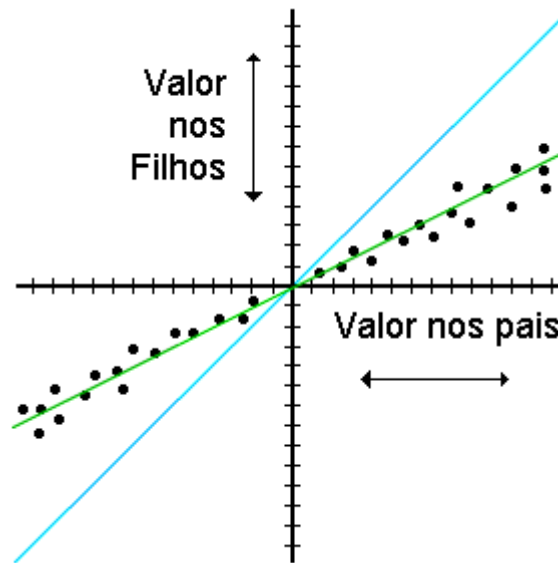


Figura 1: Reta de regressão para a hereditariedade de Galton para pais e filhos

preocupados em encontrar regularidades estatísticas que pudessem descrever a ocorrência de variações contínuas em uma dada população (CONT, 2008).

Na Figura 1 está representada a relação entre pais e filhos de uma variável métrica (por exemplo, altura). A linha azul representa o esperado se os filhos tivessem exatamente o valor da média dos pais. Os pais que apresentam valores maiores da característica têm descendência com um valor médio da característica menor que a média observada daquela medida entre os pais. Por outro lado, os pais que tem o valor menor da característica têm os filhos com valores maiores que aquele da média entre os pais. Por isso a lei foi chamada de **regressão para a média**. Como curiosidade, o método estatístico de ajuste de linhas pelo método dos mínimos quadrados é até hoje chamado de **regressão linear** por um dos seguidores de Galton, Pearson (MANTIOLLI, 2001).

2.2 Análise de regressão linear simples

Para saber que tipo de associação existe entre X e Y , é necessário encontrar um modelo matemático que explique, se existir, a dependência de Y em relação a X .

Y e X podem estar relacionadas de forma linear, polinomial, exponencial, logarítmica, etc. Uma forma simples de se avaliar o tipo de relação (ou associação) entre as duas variáveis é através do gráfico de dispersão bivariado entre Y e X .

Estando as variáveis relacionadas de forma linear, para se obter os valores estimados

dos parâmetros, basta resolver o sistema de equações lineares, que possui solução única pois os coeficientes são combinações lineares das observações. Já no modelo não linear, os parâmetros entram na equação de forma não linear, não podendo ser resolvido da mesma forma.

O objetivo da análise de regressão linear é encontrar a reta que melhor explique a dependência dos dados, para poder assim, estimar previsões para o comportamento da variável Y em decorrência dos acontecimentos da variável X . Para isto é necessário estimar os parâmetros do modelo.

Matematicamente, o modelo linear será apresentado da seguinte forma:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \quad i = 1, 2, \dots, n, \quad (2.1)$$

onde x_i representa cada observação da variável explicativa X ; β_0 representa o coeficiente linear da reta, ou seja, representa o ponto inicial para a variável Y , quando $X = 0$; β_1 representa o coeficiente angular da reta, ou seja, o grau que a reta faz com o eixo X , e define também o quanto aumenta, ou diminui, o valor de Y em relação a X ; e ϵ_i é o erro associado a cada observação em relação à reta de regressão linear.

Para que esse modelo seja admitido, é preciso que as seguintes hipóteses sejam verificadas:

1. Existe relação linear entre X e Y ;
2. X não é uma variável aleatória;
3. As variáveis aleatórias ϵ_i têm distribuição normal;
4. Todas as variáveis aleatórias ϵ_i têm média igual a zero;
5. A variância da variável aleatória ϵ_i é σ^2 , para todos os valores de X .
6. ϵ_i é não correlacionada com ϵ_j , $\forall i \neq j$.

As condições indicadas implicam que, $\epsilon_i \sim N(0; \sigma^2)$.

2.2.1 Estimação de parâmetros

Neste trabalho, não serão feitas demonstrações para a regressão simples, visto que o principal foco é o uso da regressão linear múltipla. Todavia é interessante apresentar

alguns resultados do caso simples para facilitar a compreensão da análise de regressão linear múltipla.

Para encontrar a reta que melhor represente a relação entre X e Y , deve-se estimar os valores β_0 e β_1 . Para isto, utiliza-se o método dos mínimos quadrados, que torna mínima a soma das distâncias entre a função linear e os pontos observados na amostra.

Quando a distribuição dos erros é Normal, as estimativas de máxima verossimilhança coincidem com as do Método dos Mínimos Quadrados. Neste caso, sob as condições básicas, as estimativas para os coeficientes da equação são as estimativas lineares não tendenciosas de variância mínima.

Aplicando esse método obtemos a equação que minimiza os erros e explica melhor a dependência entre as variáveis:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i, \quad (2.2)$$

onde $\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}}$ e $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$ são os estimadores de máxima verossimilhança de β_1 e β_0 respectivamente, sendo $S_{xx} = \sum_{i=1}^n (x_i - \bar{x})^2$ e $S_{xy} = \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})$.

Tanto β_1 quanto β_0 e \hat{y}_i tem distribuição normal, e seus respectivos parâmetros são:

$$\hat{\beta}_1 \sim N\left(\beta_1; \frac{\sigma^2}{S_{xx}}\right)$$

$$\hat{\beta}_0 \sim N\left(\beta_0; \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)\right)$$

$$\hat{y} \sim N\left[\beta_0 + \beta_1; \sigma^2 \left(\frac{1}{n} + \frac{(x_i - \bar{x})^2}{S_{xx}}\right)\right]$$

A estimativa da variância será dada pelo quociente da soma dos quadrados dos desvios pelo número de graus de liberdade da soma. Entendendo-se desvio pela diferença entre as observações na amostra e a reta estimada, \hat{y}_i , dada pela equação (2.2). Assim,

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

Após o desenvolvimento desta expressão chega-se ao seguinte resultado:

$$S^2 = \frac{Syy - \hat{\beta}_1 Sxx}{n - 2} \quad (2.3)$$

2.2.2 Decomposição da soma de quadrados

A partir dos valores observados na amostra, pode-se definir as seguintes somas de quadrados:

Soma de quadrados total:

$$SQ_{tot} = \sum_{i=1}^n (y_i - \bar{y})^2 \Rightarrow SQ_{tot} = S_{yy} \quad (2.4)$$

Soma de quadrados de regressão:

$$SQ_{reg} = \hat{\beta}_1 \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) \Rightarrow SQ_{reg} = \hat{\beta}_1 S_{xy} \quad (2.5)$$

Soma de quadrados de resíduo:

$$SQ_{res} = \sum_{i=1}^n (y_i - \bar{y})^2 - \hat{\beta}_1 \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) \Rightarrow SQ_{res} = SQ_{tot} - SQ_{reg}, \quad (2.6)$$

podendo ser escrito como $SQ_{tot} = SQ_{res} + SQ_{reg}$, que também é conhecido como a decomposição da soma de quadrados total.

As somas de quadrados tem distribuição χ^2 , com os seguintes graus de liberdade:

$$\begin{aligned} \frac{SQ_{tot}}{\sigma^2} &\sim \chi_{n-1}^2 \\ \frac{SQ_{reg}}{\sigma^2} &\sim \chi_1^2 \\ \frac{SQ_{res}}{\sigma^2} &\sim \chi_{n-2}^2 \end{aligned}$$

2.2.3 Coeficiente de determinação

O coeficiente de determinação, ou explicação, R^2 é uma medida que explica o grau de associação entre as variáveis, sendo seu valor calculado a partir da equação abaixo:

$$R^2 = \frac{SQ_{reg}}{SQ_{tot}} = \frac{\hat{\beta}_1 S_{xy}}{S_{yy}}, \text{ onde } 0 \leq R^2 \leq 1, \quad (2.7)$$

quanto mais próximo de zero, menor é a evidência de dependência linear entre X e Y , e quanto mais próximo de um, maior é a evidência de associação linear entre as variáveis.

2.2.4 Testes de hipóteses

Há dois testes que se pode aplicar para verificar se há regressão linear, o teste t de Student, e o F de Snedecor. A aplicação de cada teste vem a seguir.

- Teste t de student

Para testar se há existência de regressão linear entre as variáveis, é preciso confrontar as seguintes hipóteses:

$$\begin{cases} H_0 : \beta_1 = 0 \\ H_1 : \beta_1 \neq 0 \end{cases}$$

Lembrando que: $\hat{\beta}_1 \sim N\left(\beta_1; \frac{\sigma^2}{S_{xx}}\right)$

Ou seja, se padronizando $\hat{\beta}_1$, se obterá:

$$z = \frac{\hat{\beta}_1 - \beta_1}{\frac{\sigma}{\sqrt{S_{xx}}}} \quad (2.8)$$

Como não é conhecido o valor de σ^2 , deve-se substituir pelo seu estimador S^2 , e aplicar o teste t de student com $n - 2$ graus de liberdade, ao nível de significância α . Logo,

$$t_{n-2} = \frac{\hat{\beta}_1 - \beta_1}{\frac{S}{\sqrt{S_{xx}}}} \quad (2.9)$$

Depois de verificar o valor tabelado $t_{\frac{\alpha}{2}}$, e o valor calculado t , decide-se que:

- Se $-t_{\frac{\alpha}{2}} < t < t_{\frac{\alpha}{2}}$, **aceita-se H_0** e conclui-se que ao nível de significância α , não há regressão linear;
- Se $|t| > t_{\frac{\alpha}{2}}$, **rejeita-se H_0** e conclui-se que ao nível de significância α há indícios de regressão linear entre X e Y .

Com o teste t pode-se também testar se $\beta_0 = b_0$, assim as hipóteses de interesse são:

$$\begin{cases} H_0 : \beta_0 = b_0 \\ H_1 : \beta_0 \neq b_0 \end{cases}$$

lembrando que: $\hat{\beta}_0 \sim N\left(\beta_0; \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)\right)$

Padronizando β_0 , se obterá:

$$t = \frac{\hat{\beta}_0 - b_0}{S \sqrt{\frac{1}{n} + \frac{\bar{X}^2}{S_{xx}}}} \quad (2.10)$$

De maneira análoga, encontra-se um valor para t e conclui se H_0 será aceita ou rejeitada.

- TESTE F DE SNEDECOR

As hipóteses de interesse são as mesmas do teste t, ou seja,

$$\begin{cases} H_0 : \beta_1 = 0 \\ H_1 : \beta_1 \neq 0 \end{cases}$$

É preciso construir a tabela da análise de variância, que é representada pela Tabela 1, para encontrar a estatística F, e assim analisar se H_0 será aceita ou não.

Tabela 1: Análise de variância no caso simples

<i>FV</i>	Gl	SQ	QM	F
<i>Regressão</i>	1	$\hat{\beta}_1 S_{xy}$	$\frac{\hat{\beta}_1 S_{xy}}{1}$	$\frac{\hat{\beta}_1 S_{xy}}{S^2}$
<i>Resíduo</i>	$n - 2$	$S_{yy} - \hat{\beta}_1 S_{xy}$	$\frac{S_{yy} - \hat{\beta}_1 S_{xy}}{n-2} = S^2$	-
<i>TOTAL</i>	$n - 1$	S_{yy}	-	-

Após fixar o nível α de significância e olhar na tabela o valor de $F_{1,n-2}$, decide-se que:

- Se $F_{calculado} > F_{tabelado}$, rejeita-se H_0 ao nível de significância α , e conclui-se que não existe regressão.

- Se $F_{calculado} < F_{tabelado}$, aceita-se H_0 ao nível de significância α , e conclui-se que não há indícios de regressão.

2.2.5 Intervalo de confiança

Através das estimativas calculadas, podemos assegurar um intervalo de confiança que contenha o verdadeiro valor dos parâmetros β_0 , β_1 e Y ao nível α de significância.

- Para β_0 :

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

- Para β_1 :

$$IC(\beta_1) = \left[\hat{\beta}_1 \pm t_{\frac{\alpha}{2}} \frac{S}{\sqrt{S_{xx}}} \right]$$

- Para $F(X)$ ou $E(Y/X_i)$:

$$IC(E(y_i)) = \left[\hat{y}_i \pm t_{\frac{\alpha}{2}} S \sqrt{\frac{1}{n} + \frac{(x_i - \bar{x})^2}{S_{xx}}} \right]$$

- Para uma previsão y_i , dado um particular valor x_i de X :

$$IC(y_i) = \left[\hat{y}_i \pm t_{\frac{\alpha}{2}} S \sqrt{1 + \frac{1}{n} + \frac{(x_i - \bar{x})^2}{S_{xx}}} \right]$$

2.3 Análise de regressão linear múltipla

No caso da regressão linear simples sabemos que Y depende apenas de uma variável X . No caso múltiplo será estudada a relação de dependência de Y com mais de uma variável.

O modelo estatístico da regressão linear múltipla é dada pela seguinte equação:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \dots + \beta_k x_{ki} + \epsilon_i, \quad i = 1, 2, \dots, n \quad (2.11)$$

Ou seja,

$$y_i = \beta_0 + \sum_{j=1}^k \beta_j x_{ji} + \epsilon_i \quad (2.12)$$

Para facilitar a visualização da regressão linear múltipla, será feito uso da notação matricial, assim, o modelo é da seguinte forma:

$$Y = X\beta + \epsilon \quad (2.13)$$

onde

$$Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \quad X = \begin{bmatrix} 1 & x_{11} & x_{21} & \cdots & x_{k1} \\ 1 & x_{12} & x_{22} & \cdots & x_{k2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{1n} & x_{2n} & \cdots & x_{kn} \end{bmatrix}, \quad \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix}, \quad \epsilon = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

As suposições sobre o modelo de regressão são as mesmas descritas no modelos simples:

- Existe relação linear entre Y e X_j , $j=1,2,\dots,k$;
- Os valores dos X_j são sempre fixos, ou seja, eles não são variáveis aleatórias;
- As variáveis aleatórias ϵ_i têm distribuição normal;
- $E(\epsilon) = 0$, onde 0 que representa o vetor nulo;
- $Var(\epsilon) = \sigma^2$, para todos os valores de X_j ;
- Os erros são não correlacionados dois a dois.

2.3.1 Estimação de parâmetros

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 $\hat{\beta}$ e ϵ os vetores das estimativas e dos desvios do modelo, onde:

$$\hat{\beta} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \vdots \\ \hat{\beta}_k \end{bmatrix}, \epsilon = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

Temos que:

$$\epsilon = Y - X\beta$$

A soma dos quadrados dos erros é dada por:

$$\epsilon'\epsilon = \begin{bmatrix} \epsilon_1 & \epsilon_2 & \cdots & \epsilon_n \end{bmatrix} \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix} = \epsilon_1\epsilon_1 + \epsilon_2\epsilon_2 + \dots + \epsilon_n\epsilon_n = \sum_{i=1}^n \epsilon_i^2$$

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

$$Z = \epsilon'\epsilon = (Y' - \beta'X')(Y - X\beta) = Y'Y - Y'X\beta - \beta'X'Y + \beta'X'X\beta$$

Como as matrizes $Y'X\beta$ e $\beta'X'Y$ são equivalentes, pois uma é a transposta da outra, e ambas possuem um único elemento, então:

$$Z = Y'Y - 2\beta'X'Y + \beta'X'X\beta \quad (2.14)$$

A função Z deve ser diferenciada e igualada a zero para se obter o ponto de mínimo para os valores de β , logo:

$$\begin{aligned} \frac{\partial Z}{\partial \beta} &= \partial(Y'Y - 2\hat{\beta}'X'Y + \hat{\beta}'X'X\hat{\beta}) \equiv 0 \Rightarrow \\ \partial(Y'Y) - 2\partial(\hat{\beta}'X'Y) + \partial(\hat{\beta}'X'X\hat{\beta}) &\equiv 0 \Rightarrow \\ -2(\partial\hat{\beta}')X'Y + (\partial\hat{\beta}')X'X\hat{\beta} + \hat{\beta}'X'X(\partial\hat{\beta}) &\equiv 0 \Rightarrow \end{aligned}$$

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

ainda,

$$(\partial \hat{\beta}') (X'X \hat{\beta} - X'Y) \equiv 0$$

Para que $\partial \hat{\beta}' \equiv 0$ é necessário que:

$$X'X \hat{\beta} = X'Y \quad (2.15)$$

O sistema acima é denominado de sistema de equações normais, e sua solução nos fornece as estimativas dos parâmetros constituintes do vetor $\hat{\beta}$.

Pré multiplicando ambos os lados da igualdade pela matriz $(X'X)^{-1}$, que é a inversa da matriz $X'X$, desde que seja não-singular, será encontrado o seguinte resultado:

$$(X'X)^{-1}(X'X)\hat{\beta} = (X'X)^{-1}(X'Y)$$

Daí aplicando o produto de matrizes e fazendo as operações encontra-se:

$$\hat{\beta} = (X'X)^{-1}(X'Y) \quad (2.16)$$

Onde,

$$X'X = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ x_{k1} & x_{k2} & \cdots & x_{kn} \end{bmatrix} \begin{bmatrix} 1 & x_{11} & x_{21} & \cdots & x_{k1} \\ 1 & x_{12} & x_{22} & \cdots & x_{k2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{1n} & x_{2n} & \cdots & x_{kn} \end{bmatrix} =$$

$$= \begin{bmatrix} n & \sum_{j=1}^n x_{1j} & \sum_{j=1}^n x_{2j} & \cdots & \sum_{j=1}^n x_{kj} \\ \sum_{j=1}^n x_{1j} & \sum_{j=1}^n x_{1j}^2 & \sum_{j=1}^n x_{1j}x_{2j} & \cdots & \sum_{j=1}^n x_{1j}x_{kj} \\ \sum_{j=1}^n x_{2j} & \sum_{j=1}^n x_{1j}x_{2j} & \sum_{j=1}^n (x_{2j})^2 & \cdots & \sum_{j=1}^n x_{2j}x_{kj} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \sum_{j=1}^n x_{kj} & \sum_{j=1}^n x_{1j}x_{kj} & \sum_{j=1}^n x_{2j}x_{kj} & \cdots & \sum_{j=1}^n (x_{kj})^2 \end{bmatrix} = (X'X)'$$

Como $X'X$ é uma matriz simétrica, sabe-se que $X'X = (X'X)'$

E também,

$$X'Y = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ x_{k1} & x_{k2} & \cdots & x_{kn} \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} \sum_{j=1}^n y_j \\ \sum_{j=1}^n x_{1j}y_j \\ \sum_{j=1}^n x_{2j}y_j \\ \vdots \\ \sum_{j=1}^n x_{kj}y_j \end{bmatrix}$$

Lembrando que nos casos onde a matriz X tem posto incompleto, será usada a inversa generalizada condicional, $(X'X)^-$, trabalhando de forma análoga para todas as estimações feitas utilizando a inversa trivial, $(X'X)^{-1}$. Isso geralmente ocorrerá quando as variáveis explicativas forem qualitativas.

Da equação (2.15) pode-se ainda obter outros resultados importantes. Segue-se que:

$$X'X\hat{\beta} - X'Y = \mathbf{0},$$

onde $\mathbf{0}$ representa um vetor nulo, logo,

$$X'(X\hat{\beta} - Y) = \mathbf{0},$$

podendo ser escrita da seguinte maneira:

$$X'\epsilon = \mathbf{0}, \tag{2.17}$$

onde ϵ' é o vetor dos erros do modelo.

Essa relação significa que $\sum_{i=1}^n \epsilon_i = 0$ e $\sum_{i=1}^n x_{ij}\epsilon_i = 0$

De acordo com (HOFFMANN, 2006), a nulidade da soma dos desvios decorre do fato de o modelo ter um termo constante (β_0), fazendo com que a primeira coluna de X seja um vetor com todos os elementos iguais a 1.

Sendo nula a soma dos desvios, conclui-se que:

$$\sum_{i=1}^n y_i = \sum_{i=1}^n \hat{y}_i \quad (2.18)$$

Onde

$$\hat{y}_i = X\hat{\beta} \quad (2.19)$$

Seja H uma matriz simétrica e idempotente, também chamada de Matriz Chapéu tal que,

$$H = X(X'X)^{-1}X' \quad (2.20)$$

Então,

$$\begin{aligned} \hat{Y} &= X(X'X)^{-1}X'Y \Rightarrow \\ \hat{Y} &= HY \end{aligned}$$

Demonstrar que $\hat{\beta}$ é um estimador não tendencioso para β é muito simples.

$$\begin{aligned} \hat{\beta} &= (X'X)^{-1}X'Y = (X'X)^{-1}X'(X\beta + \epsilon) = \\ &= (X'X)^{-1}X'X\beta + (X'X)^{-1}X'\epsilon = \beta + (X'X)^{-1}X'\epsilon \end{aligned}$$

Aplicando a esperança matemática em ambos os lados da equação, tem-se:

$$E(\hat{\beta}) = E(\beta) + E((X'X)^{-1}X'\epsilon)$$

Como X não é variável aleatória, e $E(\epsilon) = 0$, a equação resultará em:

$$E(\hat{\beta}) = E(\beta) + 0 \Rightarrow E(\hat{\beta}) = \beta \quad (2.21)$$

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

Das equações (2.14) e (2.15) obtém-se que a soma de quadrados dos resíduos pode ser escrita como:

$$\epsilon'\epsilon = Y'Y - 2\hat{\beta}'X'Y + \hat{\beta}'X'Y,$$

ou ainda,

$$SQRes = Y'Y - \hat{\beta}'X'Y. \quad (2.22)$$

Sabe-se que a soma de quadrados total é dada pela expressão:

$$SQTot = \sum_{j=1}^n (y_j - \hat{y})^2 = \sum_{j=1}^n y_j^2 - \frac{(\sum_{j=1}^n y_j)^2}{n} = Y'Y - \frac{(\sum_{j=1}^n y_j)^2}{n}, \quad (2.23)$$

e a soma de quadrados de regressão é dada por:

$$\begin{aligned} SQReg &= \sum_{j=1}^n (\hat{y}_j - \bar{Y})^2 = \sum_{j=1}^n \hat{y}_j^2 - \frac{(\sum_{j=1}^n \hat{y}_j)^2}{n} = \\ &= \hat{Y}'\hat{Y} - \frac{(\sum_{j=1}^n \hat{Y}_j)^2}{n} = (X\hat{\beta})X\hat{\beta} - \frac{(\sum_{j=1}^n \hat{Y}_j)^2}{n} = \hat{\beta}'X'X\hat{\beta} \end{aligned}$$

Das equações (2.15) e (2.18) tem-se que:

$$SQReg = \hat{\beta}'X'Y - \frac{(\sum_{j=1}^n y_j)^2}{n}, \quad (2.24)$$

logo, a soma de quadrados total pode ser escrita da seguinte forma:

$$SQRes = SQTot - SQReg.$$

Para demonstrar que $E(SQRes) = (n-p)\sigma^2$, é necessário definir M como uma matriz simétrica e idempotente, dada por:

$$M = I - X(X'X)^{-1}X',$$

que, pela equação (2.20) pode ser escrita como:

$$M = I - H,$$

e sabe-se que

$$\begin{aligned} e &= Y - \hat{Y} = Y - X\hat{\beta} = Y - X(X'X)^{-1}X'Y = [I - X(X'X)^{-1}X']Y = \\ MY &= M(X\beta + \epsilon) = MX\beta + M\epsilon = [I - X(X'X)^{-1}X']X\beta + M\epsilon = \\ &= [X - X(X'X)^{-1}X'X]\beta + M\epsilon = [X - X]\beta + M\epsilon = 0 + M\epsilon \Rightarrow \end{aligned}$$

$$e = M\epsilon \tag{2.25}$$

Mas $SQRes = e'e$, então:

$$SQRes = (M\epsilon)'M\epsilon = \epsilon'M'M\epsilon$$

como M é simétrica, $M' = M$, logo,

$$SQRes = \epsilon'MM\epsilon = \epsilon'M^2\epsilon$$

sendo M idempotente, $M^2 = M$, portanto,

$$SQRes = \epsilon'M\epsilon \tag{2.26}$$

Como a matriz $e'e$ possui apenas um elemento, pode-se escrevê-la como sendo:

$$e'e = tr(\epsilon'M\epsilon) \text{ ou ainda, } e'e = tr(\epsilon'\epsilon M)$$

Assim,

$$E(SQRes) = E(e'e) = E(tr(\epsilon'\epsilon M)) = E(\epsilon'\epsilon tr(M))$$

aplicando a propriedade da esperança matemática se obterá

$$E(SQRes) = tr(M)E(\epsilon'\epsilon)$$

já que os erros são homocedásticos ,

$$E(SQRes) = tr(M)\sigma^2,$$

mas,

$$tr(M) = tr [I - X(X'X)^{-1}X'] = tr[I] - tr[X(X'X)^{-1}X'] = n - p.$$

Logo,

$$E(SQRes) = tr(M)\sigma^2 = (n - p)\sigma^2$$

Assim, podemos construir a tabela da análise de variância para a regressão múltipla de acordo com a Tabela 2.

Tabela 2: Análise de variância no caso múltiplo

<i>FV</i>	<i>Gl</i>	<i>SQ</i>	<i>QM</i>	<i>F</i>
<i>Regressão</i>	$k = p - 1$	<i>SQReg</i>	$\frac{SQReg}{p-1}$	$\frac{QMReg}{QMRes}$
<i>Resíduo</i>	$n - p$	<i>SQRes</i>	$\frac{SQRes}{n-p}$	-
<i>TOTAL</i>	$n - 1$	<i>SQTot</i>	-	-

A matriz das estimativas das variâncias e covariâncias dos estimadores de $\hat{\beta}$ é dada pela expressão:

$$\hat{V}(\hat{\beta}) = (X'X)^{-1}S^2 \quad (2.27)$$

onde $S^2 = QMRes$

2.3.3 Coeficiente de determinação

Como no caso de regressão linear simples, o coeficiente de determinação, ou explicação é uma estatística usada para medir a proporção da soma de quadrados que é explicada pela regressão múltipla.

O coeficiente pode ser obtido através da expressão:

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

Temos que:

$$1 - R^2 = \frac{SQRes}{SQTot},$$

e para aperfeiçoar essa medida, corrige-se este coeficiente dividindo-se pelos seus graus de liberdade de modo que:

$$1 - \bar{R}^2 = \frac{\frac{SQRes}{n-p}}{\frac{SQTot}{n-1}} = \frac{n-1}{n-p}(1 - R^2)$$

ou ainda

$$\bar{R}^2 = 1 - \frac{n-1}{n-p}(1-R^2) \quad (2.28)$$

2.3.4 Teste de hipóteses

Para verificar se, de fato, há regressão linear entre as variáveis do modelo é preciso que se faça um teste de hipóteses. O melhor teste a ser feito é o teste F . Por isto faremos uso deste teste para dois casos.

2.3.4.1 Teste F para significância da equação de regressão linear múltipla

Neste caso queremos testar se há regressão linear no modelo, fazendo uso da medida F que decorre da Tabela 2 vista anteriormente, onde:

$$F = \frac{QMReg}{QMRes}, \quad (2.29)$$

onde testam-se as hipóteses:

$$\begin{cases} H_0 : \beta_1 = \beta_2 = \dots = \beta_k, \\ H_1 : \text{Pelo menos um } \beta_i \neq 0, \quad i = 1, 2, \dots, k \end{cases}$$

Se $F_{calculado} > F_{tabelado}$, então rejeitamos H_0 e concluímos que, ao nível α de significância, pelo menos um dos $\beta_i \neq 0$, ou seja, podemos afirmar que há regressão linear entre as variáveis.

2.3.4.2 Teste F para as partes de um modelo de regressão linear múltipla

A contribuição de uma variável explicativa ao modelo de regressão linear múltipla pode ser determinada pelo critério do chamado ‘teste do F parcial’. De acordo com esse critério, avalia-se a contribuição de uma variável explicativa para a soma dos quadrados devido a regressão, depois que todas as outras variáveis independentes foram incluídas no modelo (NAGHETTINI; PINTO, 2007).

Assim, a contribuição de uma variável X_k do modelo para a soma de quadrados da regressão será estimada pela diferença dada por:

$$SQReg_{(X_k)} = SQReg_{(total)} - SQReg_{(total-X_k)} \quad (2.30)$$

As hipóteses que serão testadas são:

$$\begin{cases} H_0 : \text{A variável } X_K \text{ não melhora significativamente o modelo} \\ H_1 : \text{A variável } X_K \text{ melhora significativamente o modelo, } k = 1, 2, \dots, n \end{cases}$$

A estatística abaixo nos permite fazer a comparação do teste:

$$F_C = \frac{SQReg(X_k)}{QMRes} \quad (2.31)$$

Se $F_{calculado} > F_{tabelado}$, rejeita-se H_0 ao nível α de significância e assume-se que a variável X_k melhora significativamente o modelo.

2.3.4.3 Teste t para significância dos parâmetros

Um teste bastante utilizado para medir a significância individual das variáveis do modelo é o teste t . A quantidade a ser testada para cada β_i será dada pela expressão:

$$t_i = \frac{\hat{\beta}_i}{S(\hat{\beta}_i)} \sim t_{n-p}, \quad i = 1, 2, \dots, k, \quad (2.32)$$

o teste é executado para se verificar as seguintes hipóteses:

$$\begin{cases} H_0 : \beta_i = 0 \\ H_1 : \beta_i \neq 0, \quad i = 1, 2, \dots, n \end{cases}$$

Se $|t_{calculado}| > t_{tabelado}$, então rejeitamos H_0 e concluímos que, ao nível α de significância, $\beta_i \neq 0$, $i = 1, 2, \dots, n$, ou seja, esta variável é importante para explicar a regressão linear. Caso contrário, esta variável não influencia na regressão.

2.3.5 Intervalo de confiança

Para os modelos com erros normais, temos:

$$\frac{\hat{\beta}_k - \beta_k}{S(\hat{\beta}_k)} \sim t_{n-p}, \quad k = 1, 2, \dots, p-1,$$

a partir desta quantidade pivotal, podemos construir um intervalo de confiança para um certo β_k .

Sabe-se que:

$$P(-t_{n-p} < \frac{\hat{\beta}_k - \beta_k}{S(\hat{\beta}_k)} < t_{n-p}) = 1 - \alpha$$

Portanto, fazendo operações matemáticas necessárias podemos notar que:

$$\begin{aligned} P(-S\hat{\beta}_k t_{n-p} < \hat{\beta}_k - \beta_k < S\hat{\beta}_k t_{n-p}) &= 1 - \alpha \Rightarrow \\ P(-\hat{\beta}_k - S\hat{\beta}_k t_{n-p} < -\hat{\beta}_k + \hat{\beta}_k - \beta_k < -\hat{\beta}_k + S\hat{\beta}_k t_{n-p}) &= 1 - \alpha \Rightarrow \\ P(-\hat{\beta}_k - S\hat{\beta}_k t_{n-p} < -\beta_k < -\hat{\beta}_k + S\hat{\beta}_k t_{n-p}) &= 1 - \alpha \Rightarrow \\ P(\hat{\beta}_k + S\hat{\beta}_k t_{n-p} > \beta_k > \hat{\beta}_k - S\hat{\beta}_k t_{n-p}) &= 1 - \alpha \Rightarrow \\ P(\hat{\beta}_k - S\hat{\beta}_k t_{n-p} < \beta_k < \hat{\beta}_k + S\hat{\beta}_k t_{n-p}) &= 1 - \alpha \end{aligned}$$

Ou seja,

$$IC(\beta_k) = \left[\hat{\beta}_k \pm S(\hat{\beta}_K)t_{n-p} \right] \quad (2.33)$$

e este resultado significa que a probabilidade do intervalo de confiança conter o verdadeiro valor do parâmetro é de $1 - \alpha$.

Considerando o modelo de regressão linear múltiplo, $Y = X\beta + \epsilon_i$, temos que a estimativa de

$$E(y_h) = \beta_0 + \beta_1 x_{1h} + \beta_2 x_{2h} + \dots + \beta_k x_{kh} = x'_h \beta$$

é

$$\hat{y}_h = \hat{\beta}_0 + \hat{\beta}_1 x_{1h} + \hat{\beta}_2 x_{2h} + \dots + \hat{\beta}_k x_{kh} = X'_h \hat{\beta},$$

onde

$$X'_h = \left[1 \quad x_{1k} \quad x_{2k} \quad \dots \quad x_{kh} \right]$$

Sabendo também que

$$\hat{V}(\hat{Y}_h) = X'_h(X'X)^{-1}X_hS^2 \quad (2.34)$$

Com estes resultados podemos construir um intervalo de confiança para o valor esperado para um Y_h , dado um particular vetor X'_h .

De maneira análoga ao intervalo construído anteriormente, temos:

$$IC(E(Y_h)) = \left[X'_h \pm t_c \sqrt{\hat{V}(\hat{Y}_h)} \right], \quad (2.35)$$

ou seja, a probabilidade deste intervalo conter o verdadeiro valor do parâmetro é $1 - \alpha$.

Agora, se quisermos um intervalo de confiança para Y_h , é importante lembrar que o estimador de $Y_h = X'_h\beta + \epsilon_h$ é $\hat{Y}_h = X'_h\hat{\beta}$, e o erro de previsão é dado por:

$$\hat{Y}_h - Y_h = X'_h(\hat{\beta} - \beta) - \epsilon_i$$

A variância do erro de previsão é dada pela expressão:

$$V(\hat{Y}_h - Y_h) = V \left[X'_h(\hat{\beta} - \beta) + \sigma^2 \right] = \sigma^2 + X'_h(X'X)^{-1}X_h\sigma^2 \Rightarrow$$

$$V(\hat{Y}_h - Y_h) = \left[1 + X'_h(X'X)^{-1}X_h \right] \sigma^2, \quad (2.36)$$

como não sabemos o valor de σ , usaremos em seu lugar o S , ficando a expressão :

$$V(\hat{Y}_h - Y_h) = \left[1 + X'_h(X'X)^{-1}X_h \right] S^2,$$

podemos assim, de forma análoga ao intervalo anteriormente construído, encontrar um intervalo de confiança para Y_h , dado por:

$$IC(Y_h) = \left[X'_h\hat{\beta} \pm t_c \sqrt{V(\hat{Y}_h - Y_h)} \right]. \quad (2.37)$$

Significando que a probabilidade deste intervalo conter o verdadeiro valor do parâmetro é $1 - \alpha$.

2.3.6 Regressões que se tornam lineares através de anamorfose

Em alguns casos, não há interesse de se trabalhar com modelos não lineares, por isso se pode transformar um modelo não linear em linear, fazendo uma transformação matemática em suas variáveis. As transformações mais utilizadas são apresentadas a seguir:

- Transformação da raiz quadrada:

$$Y = \beta_0 + \beta_1\sqrt{X_1} + \beta_2\sqrt{X_2} + \dots + \epsilon$$

- Transformação logarítmica:

$$Y = \beta_0 + \beta_1\ln(X_1) + \beta_2\ln(X_2) + \dots + \epsilon$$

- Transformação recíproca:

$$Y = \beta_0 + \beta_1\frac{1}{X_1} + \beta_2\frac{1}{X_2} + \dots + \epsilon$$

Assim, um modelo do tipo multiplicativo pode sofrer uma anamorfose e tornar-se linear aplicando-se alguma transformação, como por exemplo, um modelo que é da forma $Y = \beta_0 X_1^{\beta_1} X_2^{\beta_2} \epsilon$ pode ser escrito linearmente como: $\ln Y = \ln \beta_0 + \beta_1 \ln X_1 + \beta_2 \ln X_2 + \ln \epsilon$, tornando-se linear.

2.3.7 Análise de resíduos

Para não haver nenhuma violação no que diz respeito às suposições necessárias para que haja regressão linear, é importante fazer uma investigação no conjunto de dados.

A condição de normalidade pode ser verificada usando um gráfico de probabilidade normal também conhecido como **Q-Q Plot**.

Os gráficos de probabilidade normal podem ser:

- **P-P Plot** : Probabilidade acumulada esperada para a distribuição normal, em função da probabilidade observada acumulada dos resíduos ;
- **Q-Q Plot** : Quantil de probabilidade esperado para a distribuição normal, em função dos resíduos .

Após esboçar os 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 tem indícios de normalidade.

Para melhor entendimento, pode-se observar os exemplos dados pelas Figuras 2 e 3 que seguem.¹

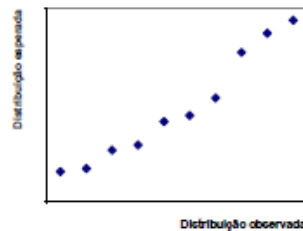


Figura 2: Gráfico P P Plot

A maioria dos pontos da Figura 2 concentram-se em torno de uma reta, o que dá indícios de que a distribuição dos erros é normal.

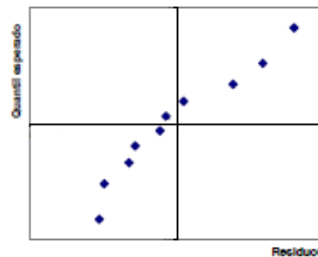


Figura 3: Gráfico Q Q plot

Da mesma forma, na Figura 3, observa-se que a maioria dos pontos estão sobre uma reta, dando a entender que os erros seguem uma distribuição normal.

O gráfico de resíduos é construído pelos valores esperados para a variável resposta Y , contra os resíduos. Nele, os pontos devem distribuir-se de forma aleatória em torno do zero, formando uma mancha de largura uniforme, para que os erros tenham variância constante. Quando os resíduos não se comportam de forma aleatória, a condição de homocedasticidade parece não ser satisfeita.

O gráfico tem no eixo das abcissas os valores estimados de Y , e no eixo das ordenadas os valores estimados de ϵ . A Figura 4 traz alguns exemplos para os possíveis comportamentos encontrados nos gráficos de resíduos.

¹Todos os exemplos de gráficos desta seção foram retirados de [http://www.estv.ipv.pt/PaginasPessoais/psarabando/Estatística CA 2009-2010/slides/regressão/Parte 3.pdf](http://www.estv.ipv.pt/PaginasPessoais/psarabando/Estatística_CA_2009-2010/slides/regressão/Parte3.pdf)

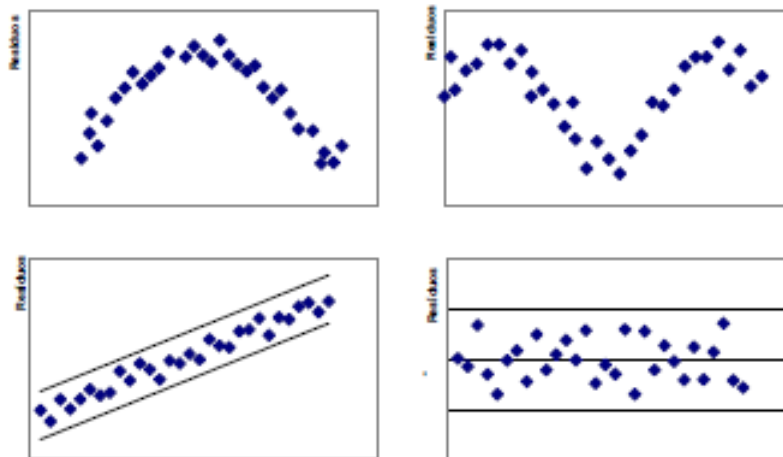


Figura 4: Exemplos de comportamento dos gráficos de resíduos

Nos 3 primeiros gráficos, não há indícios de linearidade, pois os pontos se distribuem nos gráficos de forma padronizada. E no último, os resíduos parecem estar distribuídos de forma aleatória, levando a acreditar que os modelo está bem ajustado.

Para saber se a variância é realmente constante também é feito um gráfico de resíduos. Observe na Figura 5 os casos onde não há homocedasticidade. As variâncias aumentam cada vez mais no primeiro caso, e diminuem no segundo caso.

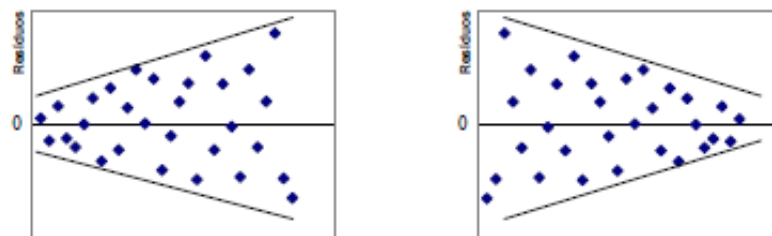


Figura 5: Resíduos com variância não constante

Assim sendo, quanto mais os pontos se comportam de forma aleatória em torno do zero no gráfico de resíduo, ou assumem um comportamento parecido com uma reta crescente nos gráficos Q-Q Plot e P-P-Plot, dá indícios de que os erros estão seguindo conforme as suposições necessárias.

3 Aplicação

3.1 Banco de dados

Antes de comentar sobre o banco de dados em si, é bom que haja uma contextualização sobre o local onde ocorreram os conflitos de trânsito.

A cidade em estudo é Campina Grande. Ela se localiza no interior do estado da Paraíba, sobre o Planalto da Borborema, com altitude média de 555 metros acima do nível do mar. A área do município abrange 599,6 km^2 .

Possui 383.764 habitantes, segundo estimativas do IBGE em 2009, e contém 49 bairros e 6 distritos.

Até dezembro de 2009, a frota veicular de Campina Grande era de 102.279 veículos, destes, 49.321 são automóveis e 37.324 são motocicletas, segundo dados do DENATRAN. Este número vem crescendo consideravelmente, e em consequência, o número de acidentes de trânsito no perímetro urbano também.

Dependendo do tipo de ocorrência são acionados a Superintendência de Trânsito e Transportes Públicos, **STTP**, que preza pela organização do trânsito no local do acidente, a Companhia de Policiamento de Trânsito, **CPTRAN**, que se responsabiliza pela perícia do acidente e pune os condutores não habilitados, embriagados, ou que estejam em alguma outra falta com a justiça, e o Serviço de Atendimento Móvel de Urgência **SAMU**, que se encarrega de socorrer as vítimas. E cada um destes órgãos emite um boletim de ocorrência (**BO**) para cada acidente.

A STTP se responsabiliza por recolher seus próprios BO's, bem como os da CPTRAN e SAMU, para poder organizar todas as informações sobre os acidentes.

É importante lembrar que em alguns acidentes apenas um dos órgãos é acionado, mas em alguns existem outros presentes, portanto, antes de montar o banco de dados é necessário que se faça uma investigação para averiguar se há mais de um registro do

referido acidente. Após feita a verificação, cada acidente é introduzido no banco de dados.

Todas as observações são introduzidas em planilhas do excel, onde cada linha representa um acidente, e as colunas são as variáveis envolvidas no acidente. Para facilitar a análise estatística, a planilha foi convertida em um arquivo do *SPSS* versão 18.

Para facilitar o estudo foram selecionados os acidentes de trânsito ocorridos em 2009 que continham um ou dois veículos envolvidos apenas, compondo um banco de dados de 3486 observações com 66 variáveis.

As variáveis são a data da ocorrência, o horário em que ocorreu o acidente, o dia da semana, a rua ou cruzamento, o bairro, um ponto de referência, o número de veículos envolvidos, as condições do tempo no momento do acidente (chuvoso ou não), o tipo acidente (colisão, atropelamento, choque, etc), a gravidade do acidente, o veículo de cada condutor, sexo dos condutores, idade dos condutores, habilitação do condutor, situação de sobriedade do condutor, placa do veículo, equipamento de segurança usado pelo condutor na hora do acidente, ação do condutor após o ocorrido, número de vítimas, tipo de vítima (condutor, passageiro, pedestre), sexo da vítima, idade da vítima, veículo em que a vítima estava na hora do acidente, uso de equipamento de segurança pela vítima, gravidade dos ferimentos, hospital para onde a vítima foi levado, e a equipe que registrou a ocorrência. Como se pode perceber, a maioria destas variáveis é do tipo nominal.

3.2 Análise de regressão

Antes de começar o estudo de regressão linear entre as variáveis do banco de dados, é necessário definir a variável dependente, também denotada de variável resposta.

A partir dos dados coletados dos boletins de ocorrência de acidentes de trânsito referentes ao ano de 2009 em Campina Grande, deseja-se saber quais fatores contribuem para explicar o número de vítimas ocasionadas por acidente. Para isto será definida a variável resposta como sendo o número de vítima do i -ésimo acidente.

As variáveis testadas para validar o modelo de regressão linear foram o número de veículos envolvidos no acidente, a condição do tempo no momento do acidente, o sexo dos condutores, a idade e o estado de sobriedade dos condutores.

Com a ajuda do *SPSS*, pode-se detectar que duas, dentre as oito variáveis escolhidas, não contribuía para explicar o número de vítimas, por isso, o *software* as excluiu automaticamente. As variáveis descartadas são o número de veículos e o sexo do

primeiro condutor. O *software* também excluiu todas as observações quem continham dados ausentes, ou seja, dos 3486 acidentes registrados, foram consideradas para o estudo apenas 879 observações.

O modelo de regressão testado para este caso particular será dado por:

$$y_i = \beta_o + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \beta_4 x_{4i} + \beta_5 x_{5i} + \beta_6 x_{6i} + \epsilon_i \quad (3.1)$$

Sendo y_i o número de vítimas na i -ésima observação, x_{1i} a condição do tempo na i -ésima observação, x_{2i} a idade do condutor 1 na i -ésima observação, x_{3i} o estado de sobriedade do condutor 1 na i -ésima observação, x_{4i} o sexo do condutor 2 na i -ésima observação, x_{5i} a idade do condutor 2 na i -ésima observação, x_{6i} é o estado de sobriedade do condutor 2 na i -ésima observação, ϵ_i o erro para cada observação e β_j os parâmetros do modelo de regressão, com $i = 1, 2, 3, \dots, 879$ e $j = 0, 1, \dots, 6$.

Como não se sabe o valor exato de cada parâmetro, e assumindo que os erros estão dentro das suposições do modelo de regressão linear, podemos estimar o modelo através da expressão:

$$\hat{y}_i = \hat{\beta}_o + \hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i} + \hat{\beta}_3 x_{3i} + \hat{\beta}_4 x_{4i} + \hat{\beta}_5 x_{5i} + \hat{\beta}_6 x_{6i} \quad (3.2)$$

Onde o valor de \hat{y}_i é a estimativa para o número de vítimas no acidente na i -ésima observação e os $\hat{\beta}_j$, $j = 0, 1, 2, \dots, 6$ são as estimativas dos valores dos coeficientes de regressão para a j -ésima variável.

Para verificar se existe regressão entre as variáveis foi gerada a tabela de ANOVA, que está representada na Tabela 3. Nela pode-se observar as somas de quadrados, os graus de liberdade, e os quadrados médios de regressão e de resíduos, que foram obtidas dos parâmetros estimados.

Tabela 3: Análise de variância do modelo

<i>FV</i>	<i>SQ</i>	<i>Gl</i>	<i>QM</i>	<i>F</i>	<i>P-valor</i>
<i>Regressão</i>	51,661	6	8,610	14,681	0,000
<i>Resíduo</i>	511,426	872	0,586	-	-
<i>TOTAL</i>	563,088	878	-	-	-

É possível verificar, observando a Tabela 3, que das 3.486 observações apenas 879 foram consideradas para o cálculo da ANOVA. Isto acontece porque o o SPSS ignora

as ocorrências onde tem dados perdidos. Note que o P-valor do teste é 0,000, mas não significa que o P-valor é nulo, e sim muito pequeno, inferior a 0,001, pois o *SPSS* só mostra três casas decimais após a virgula. Nos casos onde este valor se faz presente, pode-se afirmar que o P-valor do teste é altamente significativo porque aos níveis de 10%, 5% e até 1% de significância há fortes indícios contra H_0 , ou seja, pode-se considerar que haja regressão linear entre pelo menos duas variáveis do modelo. Das 3.486 observações apenas 879, isto porque o software não considerou os acidentes que tinham alguma variável com dado perdido.

A partir dos cálculos efetuados pelo programa utilizado foram estimados os valores dos coeficientes de regressão linear para este modelo. Os resultados encontram-se na Tabela 4.

Tabela 4: Coeficientes de regressão linear do modelo

β_j	Estimativa	t	<i>P-valor</i>
β_0	1,162	6,836	0,000
β_1	0,046	0,482	0,630
β_2	-0,004	-2,125	0,034
β_3	1,138	6,244	0,000
β_4	-0,142	-2,025	0,043
β_5	-0,008	-4,439	0,000
β_6	0,289	1,739	0,082

É possível ver na Tabela 4 os p-valores para cada variável, e analisar quais delas servem, ou não, para contribuir na estimação do número de vítimas. Se adotarmos $\alpha = 0,05$, rejeita-se β_1 e β_6 ao nível de 5% de significância já que seus P-valores foram respectivamente 0,630 e 0,082 e concluímos que aparentemente a condição do tempo na hora do acidente e o estado de sobriedade do condutor 2 não explicam o número de vítimas.

Esta evidência também pode ser constatada através do IC para cada parâmetro observando a Tabela 5. O intervalo foi construído ao nível de 95% de confiança.

Como os intervalos de confiança que contém o zero não são significativos, rejeitamos a hipótese de que as condições do tempo e a sobriedade do condutor 2 contribuam para explicar o modelo. Por isso, para melhor explicar o número de vítimas, o modelo será melhor estimado por:

$$\hat{y}_i = 1,162 - 0,004x_{2i} + 1,138x_{3i} - 0,142x_{4i} - 0,008x_{5i}$$

Tabela 5: Intervalo de confiança para os coeficientes de regressão linear do modelo

β_j	Limite inferior	Limite superior
β_0	0,828	1,495
β_1	-0,142	0,234
β_2	-0,008	-0,000
β_3	0,780	1,496
β_4	-0,280	-0,004
β_5	-0,012	-0,005
β_6	-0,370	0,616

Como se pode perceber, para este caso, as variáveis que realmente importam para explicar o fenômeno são a idade e a condição de sobriedade do condutor 1, além do sexo e da idade do condutor 2. As demais informações podem ser desconsideradas neste modelo pois não são estatisticamente significantes.

4 Conclusão

A análise de regressão é muito utilizada para explicar a associação entre variáveis. Com esta técnica estatística é possível 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.

A aplicação feita neste trabalho permitiu exemplificar a importância da análise de regressão para modelar problemas cotidianos onde se há a relação efeito causa-consequência. No caso estudado, desejava-se encontrar uma relação entre características dos condutores envolvidos em acidentes de trânsito e o número de vítimas decorrentes dos mesmos.

As informações foram retiradas dos Boletins de Ocorrência da STTP, SAMU e CP-TRAN, que coletaram dados dos acidentes ocorridos em Campina Grande-PB no ano de 2009, e organizadas no *SPSS*. O banco de dados original continha 3.627 observações, mas apenas 3.486 foram selecionadas, pois no restante o número de veículos envolvidos era maior do que 2, o que dificultava a análise. Do banco de dados selecionado podiam ser observadas 66 variáveis, mas só foram selecionadas 8, por parecer que pudessem influenciar sobre o número de vítimas geradas nos acidentes.

Após escolher as variáveis a serem testadas, foi montado o modelo inicial. Como o *SPSS* foi escolhido para efetuar os cálculos, as contas internas do software foram feitas utilizando a inversa generalizada, já que a matriz tem posto incompleto pelo fato de as variáveis respostas serem nominais. A aplicação da Análise de Regressão Linear neste banco de dados só foi possível porque a variável resposta é numérica, caso contrário não teria sentido fazer nenhum estudo de associação linear entre elas, seria mais indicado fazer a aplicação de um teste específico para variáveis categóricas.

O *SPSS* procurou na planilha os acidentes onde as variáveis selecionadas não tivessem dados ausentes ou perdidos para gerar os resultados de interesse, por isso, das 3.486 ocorrências válidas, apenas 879 foram escolhidas. Os cálculos feitos permitiram verificar que havia regressão entre pelo menos um par de variáveis, ao nível de 1% de significância.

Tendo sido aceita a hipótese de regressão, foi aplicado o teste t para verificar quais variáveis tinham potencial para explicar a variável dependente. A partir da Tabela 4 pôde-se estimar os valores para cada coeficiente de regressão e seus respectivos p-valores. Fixando o nível de significância $\alpha=0,05$, as hipóteses de que as variáveis tempo na hora do acidente e estado de sobriedade do segundo condutor contribuam para explicar o número de vítimas foram rejeitadas, pois seus p-valores foram, respectivamente, 0,630 e 0,082, por isso, as variáveis foram excluídas do modelo. Tais exclusões também podem ser confirmadas pelos intervalos de confiança dados pela Tabela 5, pois $0 \in IC(\beta_1)$ e $IC(\beta_6)$, ou seja, há indícios de que as mesmas não interfiram na variável resposta.

De acordo com o estudo feito, a função que é mais razoável para explicar o fenômeno é dada por:

$$\hat{y}_i = 1,162 - 0,004x_{2i} + 1,138x_{3i} - 0,142x_{4i} - 0,008x_{5i},$$

onde \hat{y}_i é a estimativa do número de vítimas na i-ésima observação, x_{2i} é a idade do condutor 1 na i-ésima observação, x_{3i} o estado de sobriedade do condutor 1 na i-ésima observação, x_{4i} o sexo do condutor 2 na i-ésima observação e x_{5i} a idade do condutor 2 na i-ésima observação.

A partir desta estimação, é possível identificar algumas características que podem determinar o número de vítimas de acidentes de trânsito observando os coeficientes de regressão. Deste modo, para este modelo, teremos que, para a idade dos condutores, quanto mais novo for o condutor, maior será o chance de que o número de vítimas aumente. Para o estado de sobriedade do condutor 1, se ele estiver embriagado, maior será a possibilidade de o número de vítimas ser maior. Quanto ao sexo do segundo condutor, se ele for do sexo masculino, provavelmente o número de vítimas será maior.

Neste trabalho foi possível exemplificar como é feita a análise de regressão linear através de um conflito real. Os cálculos foram feitos a partir do software estatístico SPSS, seguindo todas as suposições exigidas por esta técnica estatística. A partir da construção da ANOVA, feita internamente pelo SPSS, pôde-se verificar que havia regressão entre pelo menos um par de variáveis, e assim, achou-se necessário aplicar o teste t para verificar quais coeficientes contribuíam para explicar o fenômeno, e logo após foram construídos os intervalos de confiança individuais para os mesmos. Daí, excluindo-se as variáveis que não estavam relacionadas com a variável resposta, chegou-se à função estimada para o modelo de regressão linear múltiplo que sugere o número de vítimas a partir de informações sobre os condutores.

A análise de Regressão Linear é muito importante. Com ela podemos, a partir da amostra, analisar os dados de interesse e também estimar uma possível previsão do comportamento das variáveis a partir dos valores observados. Se as variáveis estão associadas linearmente, é possível encontrar a reta que consegue minimizar o erro das distâncias dos valores observados aos valores esperados. Esta técnica estatística tanto pode ser utilizada para um evento onde se queira explicar uma variável em função de outra, quanto para explicar em função de duas ou mais variáveis, e pode ser utilizada em vários estudos desde que as suposições a respeito dos resíduos sejam respeitadas.

Referências

- CONT, V. D. *Francis Galton: Eugenia e hereditariedade*. São Paulo, 2008. Disponível em: <<http://www.scielo.br/pdf/ss/v6n2/04.pdf> de Galton>. Acesso em: 11/03/2010.
- HOFFMANN, R. *Análise de regressão: uma introdução à Econometria*. 4. ed. São Paulo: Hucitec, 2006.
- MANTIOLLI, S. R. *Introdução ao Estudo dos QTLS (Locos de Caracteres Quantitativos): Aspectos Historicos*. Departamento de Biologia, Instituto de Biociências, Universidade de São Paulo: [s.n.], 2001. Disponível em: <<http://www.ib.usp.br/evolucao/QTL/historiaqtl.htm>>. Acesso em: 11 de março de 2010.
- NAGHETTINI, M.; ANDRADE PINTO, E. J. de. *Hidrologia Estatística*. Belo Horizonte: [s.n.], 2007. Disponível em: <www.cprm.gov.br/publique/cgi/cgilua.exe/sys/start.html?foid=9818.sid=36>. Acesso em: 21 de junho de 2010.