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

Nathielly Lima do Rêgo

Modelo de regressão linear múltipla com variável *dummy*: um estudo de caso

Campina Grande
Dezembro de 2012

Nathielly Lima do Rêgo

Modelo de regressão linear múltipla com variável *dummy*: um estudo de caso

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.

Orientadora:
Divanilda Maia Esteves

Co-orientador:
Tiago Almeida de Oliveira

Campina Grande
Dezembro de 2012

FICHA CATALOGRÁFICA ELABORADA PELA BIBLIOTECA CENTRAL – UEPB

R343m Rêgo, Nathielly Lima do.
Modelo de regressão linear múltipla com variável dummy [manuscrito] : um estudo de caso / Nathielly Lima do Rêgo . – 2012.
62f. : il. color.

Trabalho de Conclusão de Curso (Graduação em Estatística) – Universidade Estadual da Paraíba, Centro de Ciências e Tecnologia, 2012.
"Orientação: Profª. Dra. Divanilda Maia Esteves, Departamento de Estatística".

1. Estatística. 2. Análise de Regressão. 3. Diabetes. I. Título.

21. ed. CDD 519.536

Nathielly Lima do Rêgo

Modelo de regressão com variável *Dummy*: um estudo de caso

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: 14 / 12 / 12

Banca Examinadora:

Tiago Almeida de Oliveira

Prof. Tiago Almeida de Oliveira
Universidade Estadual da Paraíba - DE/CCT
Co-orientador

Ana Patricia Bastos Peixoto

Profª. Ana Patricia Bastos Peixoto
Universidade Estadual da Paraíba - DE/CCT
Examinadora

Ricardo Alves de Olinda

Prof. Ricardo Alves de Olinda
Universidade Estadual da Paraíba - DE/CCT
Examinador

Agradecimentos

Em primeiro lugar, agradeço a Deus pela sua infinita misericórdia, concedeu-me graça, disposição, inteligência, conhecimento, força, saúde, enfim, tudo proporcionou para que eu chegasse onde cheguei, pois sem ELE eu nada seria.

Aos meus pais Agnaldo Araújo do Rêgo e Maria Iolanda Lima do Rêgo, que deram de todo seu amor, de toda uma base familiar e ética, para que eu pudesse crescer e ser alguem com sabedoria e humildade.

As minhas irmãs Nathallie Lima do Rêgo e Nathalia Lima do Rêgo, que mesmo com muitas discussões e também muitas risadas, sempre me incentivaram na busca e realização dos meus sonhos.

A professora orientadora Drª Divanilda Maia Esteves pela dedicação oferecida para realização do presente trabalho, e principalmente ao professor Dr Tiago Almeida de Oliveira - coorientador e a professora Msª Ana Patricia Bastos Peixoto pelas horas e horas de tempo concedida para me ajudar na realização deste trabalho, pela total atenção, dedicação, compreensão e por seus exemplos de ética, compromisso e profissionalismo. A todos os professores(as) Ana Cristina, Castor, Fábio, Gil, Juarez, Kátia, Mauricio, Ricardo, Ruth, Victor Hugo, pela contribuição no meu aprendizado ao passar pela graduação.

Às colegas, Adriana, Arielly, Valneli, Wanessa e Samara, pelo incentivo, força, amizade, carinho que partilhamos durante nosso caminhar... nas viagens, nos pôsteres, nas conquistas, nas brigas, nas horas de muitas risadas, na espera do 333, no tempo bom que passei junto a vocês.

Obrigado a todas as pessoas que contribuíram para meu sucesso e para meu crescimento como pessoa. Sou o resultado da confiança e da força de cada um de vocês. **(Augusto Branco)**

Resumo

No presente trabalho foi estudado a análise de regressão linear múltipla com variável *dummy*, em que, sistematiza e analisa em detalhes duas situações típicas do uso desta técnica, de uma forma que pode-se generalizar a sua utilização a modelos mais complexos que incluam um número qualquer de variáveis regressoras. Para o estudo foram utilizados dados oferecidos pela biblioteca do R *Faraway* (*Diabetes survey on Pima Indians*) diabetes em 768 índias adulta Pimas que moram perto de *Phoenix* - *Arizona*, coletado segundo critérios da Organização Mundial de Saúde pelo Instituto Nacional de Diabetes e Doenças Digestivas e Renais. Com o objetivo de ajustar um modelo de regressão linear múltipla que melhor se adequasse aos dados, foi realizado uma seleção de variáveis regressoras segundo os critérios *stepwise*, AIC e BIC. Daí o modelo ajustado após o critério BIC, só foram incluidas as variáveis insulina, teste (quando 0 não diabética e 1 quando diabética), insulina ao quadrado, e idade, de modo que, todas foram consideradas estatísticamente significativas tanto pelo teste F quanto pelo teste t. Os cálculos foram realizados por meio do *software* R 2.15 (*R Development Core Team*), com a função *lm* da biblioteca *MASS*.

Palavra-chave: Análise de Regressão, Variável *Dummy*, Diabetes.

Abstract

In this work, we studied multiple linear models with dummies variables. Such variables are artificial variables, constructed to include in the model a categorical variable. These variables make it possible to observe whether there is, for example, intercepts different for different groups of observations. For the application, we used data Diabetes survey on Pima Indians provided by the library Faraway of the R software. The data are about 768 Pima Indians who live near Phoenix - Arizona, collected according to the criterion of World Health Organization by the National Institute of Diabetes and Digestive and Kidney Diseases. To find the best linear regression model adjusted to the data, we proceeded a selection of regressive variables according to the stepwise criteria, AIC and BIC. In the model chosen by the BIC criterion, only variables insulin test (dummy variable), insulin squared and age were included. So that, all those variables were considered statistically significant at both test F as the test t. The calculations were performed using the Software R 2.15 (*R Development Core Team*), with the function *lm* from the library *MASS*.

Key-words: Regression Analysis, Dummies Variables , Diabetes.

Sumário

Lista de Figuras

Lista de Tabelas

1 Introdução

A relação entre variáveis mensuradas em estudos de determinados fenômenos é um dos assuntos mais estudados e até mesmo investigados nas ciências. A ideia é estabelecer uma relação funcional entre variáveis, com o intuito de se predizer mudanças nos valores das variáveis que se estuda. A análise de regressão é um método utilizado para conhecer os efeitos que algumas variáveis exercem sobre outras. Até mesmo quando não existe uma relação casual entre as variáveis, elas podem se relacionar por meio de algumas expressões matemáticas, que são uteis para a estimação do valor de uma das variáveis, quando se tem conhecimento dos valores das outras variáveis (HOFFMANN, 2006).

Algumas vezes há interesse não apenas em saber se existe associação entre duas variáveis quantitativas X e Y, mas também em conhecer uma provável relação de causa e efeito entre variáveis. Deseja-se saber se Y depende de X. Neste caso, Y é chamado de variável dependente ou variável resposta e X é chamado de variável independente ou explanatória. A regressão é dita linear, quando considera-se que a relação da resposta às variáveis é uma função linear de alguns parâmetros. Os modelos de regressão que não são uma função linear dos parâmetros se chamam modelos de regressão não linear.

A análise de regressão linear simples é utilizada quando a predição da variável dependente é realizada em apenas uma variável independente, enquanto a análise de regressão linear múltipla diz respeito à predição da variável dependente com base em duas ou mais variáveis independentes. Na regressão, as variáveis independentes são influenciadas não apenas por variáveis quantitativas, mas também por variáveis qualitativas, em que, podem ser chamadas de variáveis *dummies*. Já quando a variável *dummy* é uma variável dependente toma o valor de 0 ou 1 para indicar a presença ou ausência de uma categoria.

O objetivo desse trabalho foi ajustar um modelo de regressão linear multipla com variável *dummy*, por meio de critérios de seleção de variáveis *stepwise*, AIC e BIC, aos dados oferecidos pela biblioteca do R (*Faraway*) *Diabetes survey on Pima Indians*.

2 Fundamentação Teórica

O conteúdo desta seção relata os principais aspectos da utilização dos modelos de regressão múltipla com variável *dummy*, por meio de artigos práticos e teóricos relacionados ao objetivo da pesquisa.

2.1 Marco Histórico

Sir Francis Galton (1822 - 1911) elaborou a sugestão de que a distribuição normal é completamente determinada pela mediana e o desvio semiquartílico, tendo usado preferencialmente a distribuição cumulativa de frequência. No entanto, a maior contribuição de Galton para a estatística foi a enumeração explícita e parcialmente quantitativa dos conceitos de regressão e correlação. Contudo, foi no estudo comparativo da estatura entre pais e filhos, em 1885, que Galton, observou que a altura dos filhos de pais com alturas extremas tendia a regredir na direção da altura média (da população de pais), ou seja, filhos de pais muito altos tendiam a ser mais baixos que seu pais, filhos de pais muito baixos tendiam a ser mais altos que seus pais. Na tábua de frequência bidimensional, representada pela altura dos filhos adultos e da média da altura (do pai com da mãe), onde a estrutura das mulheres foi multiplicada por 1,08; Galton observou que os contornos de igual frequência eram constituídos por elipses concêntricas semelhantes dispostas e traçou as linhas de regressão à mão, tendo achado que a declividade da linha de regressão dos pais em relação à dos filhos era metade da declividade da linha de regressão dos filhos em relação aos pais, uma vez que a média da altura do pai com a altura da mãe não eram correlacionadas, cada uma com a mesma dispersão populacional.

A Figura 1, representa a relação entre pais e filhos de uma variável métrica. A linha azul representa o valor esperado se os filhos tivessem exatamente o valor da altura 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

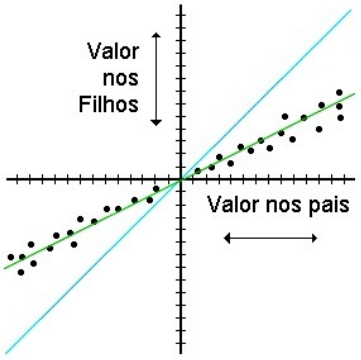

Figura 1: Gráfico de Dispersão da Lei de regressão para a mediocridade de Galton
Fonte: http://www.alea.pt/html/nomesEdatas/swf/biografias.asp?art=7

valores maiores que a média entre altura dos pais. Por isso a lei foi chamada de regressão para a média.

Ainda não havia ocorrido a Galton a expressão exata da correlação, pois desconhecia seu sinal, então Galton pediu ajuda para seu amigo J. D. Hamilton Dickson, professor de Matemática na Universidade de Cambridge, para achar a fórmula da superfície encontrada, que hoje corresponde à da função normal bidimensional. Ele expressou-se sobre a co-relação (só depois escrita correlação), como consequência das variações devidas a causas comuns. A letra r foi inspirada na letra inicialmente usada para designar a reversão. Os efeitos de posição e de escala das observações das variáveis foram eliminados com a padronização das variáveis por meio da centragem sobre a mediana e pela eliminação do efeito escala pela divisão pelo desvio semiquartílico. Contudo, essa padronização trazia a inconveniência de produzir valores de r maiores que a unidade. A fórmula por ele proposta foi modificada por Walter Frank Raphael Weldon (1860 - 1906), professor de Zoologia em Cambridge, muito ligado a Galton, que chegou à necessidade de se atribuir um sinal positivo ou negativo. Entretanto, a fórmula do coeficiente de correlação, como é hoje conhecida, só foi determinada em 1896, por Karl Pearson (1857 - 1936).

2.2 Análise de regressão linear

Na análise de regressão, há interesse de se avaliar o efeito que variáveis X_i, $i = 1, 2, \cdots, k$, chamadas de variáveis dependentes ou regressoras, têm sobre uma outra variável de interesse (Y), chamada de variável resposta, ou melhor, saber o quanto as variáveis independentes explicam o comportamento de Y. Matematicamente falando,

deseja-se encontrar uma função f de forma que

$$Y = f\left(X_1, X_2, \cdots, X_k\right).$$

No entanto, relações entre variáveis quase nunca são determinísticas, então, em lugar do modelo acima, costuma-se descrever a variável Y como a soma de uma quantidade determinística e uma quantidade aleatória. A parte aleátoria é denominada erro e pode representar inúmeros fatores, que em conjunto, podem interferir na variável resposta. Logo a função pode ser reescrita dessa forma

$$Y = f\left(X_1, X_2, \cdots, X_k\right) + \varepsilon, \tag{2.1}$$

em que, $f\left(X_1, X_2, \cdots, X_k\right)$ é a parte sistemática, e ε é a componente aleatória do modelo. Logo, pode-se dizer que o erro provoca uma distorção sobre a parte determinística.

Em geral, não há dados suficientes para estimar f diretamente e por isso, uma alternativa, é assumir que f tem uma forma mais simples. Um caso especialmente importante e bastante usado é o dos modelos lineares, em que considera-se f como sendo uma função linear. Neste caso, tem-se um modelo de regressão linear, que é representado como

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_n X_k + \varepsilon,$$

sendo $\beta_0, \beta_1, \cdots, \beta_k$ são parâmetros desconhecidos. Vale lembrar que o modelo é linear nos parâmetros e não necessariamente nas variáveis X_i, ou seja, o modelo é linear quando a derivada parcial do modelo em relação aos parâmetros não depende de nenhum dos parâmetro, caso contrário, o modelo é considerado não linear. Por exemplo os modelos:

$$Y = \beta_0 + \beta_1 \ln X_1 + \varepsilon \quad \text{e} \quad Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$$

são lineares, enquanto que

$$Y = \beta_0 \beta_1 + X_2^{\beta_2} + \beta_3 X_3 + \varepsilon \quad \text{e} \quad Y = \frac{\beta_1 X_1}{1 + \beta_0 X_1^{\beta_2}}$$

são não lineares.

Na prática, tem-se n observações sobre Y e $X_1, X_2, \cdots, X_k$, então tem-se o modelo

$$Y_i = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_k X_{ik} + \varepsilon_i, \quad i = 1, 2, \cdots, n. \tag{2.2}$$

E então, para facilitar cálculos e notação, usa-se representação matricial para o modelo

2.2:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

sendo que

$$\mathbf{Y} = \begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix}, \quad \mathbf{X} = \begin{bmatrix} 1 & X_{11} & \dots & X_{1k} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & X_{n1} & \dots & X_{nk} \end{bmatrix}, \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix}, \quad \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix}.$$

2.2.1 Modelos de regressão linear simples (MRLS)

Para saber que tipo de associação existe entre as variáveis X e Y, é necessário encontrar um modelo matemático que explique, se existe, a dependência da variável resposta Y em relação a uma variável independente X. Para isto é necessário estimar os parâmetros do modelo. O modelo estatístico da regressão linear simples é apresentado da seguinte forma:

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad i = 1, 2, \cdots, n, \tag{2.3}$$

em que,

i) x_i representa cada observação da variável explicativa X;

ii) β_0 é chamado de intercepto ou coeficiente linear representa o coeficiente linear da reta, ou seja, o ponto onde a reta corta o eixo Y, quando x=0 ;

iii) 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;

iv) ϵ_i é o erro associado a cada observação em relação à reta de regressão linear.

2.2.1.1 Estimação dos parâmetros do modelo

Como o foco principal deste trabalho é o uso da regressão linear múltipla com variável *dummy*, não serão realizadas demonstrações para regressão linear simples. Porém, é muito importante apresentar alguns resultados para melhor entendimento de análises realizadas pela regressão linear múltipla.

Supondo-se que a relação linear entre as variáveis Y e X é satisfatória, pode-se estimar a linha de regressão e resolver alguns problemas de inferência. Deseja-se, portanto, estimar

β_0 e β_1, tais que, torna-se mínima a soma das distâncias entre a função linear e os pontos observados na amostra, ou seja, segundo Holffman (2006) adota-se como estimativa dos parâmetros os valores que minimizam a soma de quadrado dos desvios. Esse método é chamado **método dos mínimos quadrados** é uma eficiente estratégia de estimação dos parâmetros da regressão.

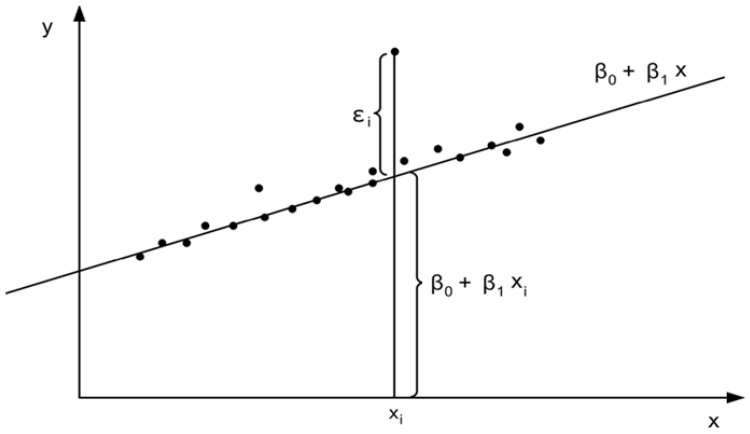

Figura 2: Reta de Regressão
Fonte: Portal Action

Aplicando-se o método de mínimos quadrados obtem-se a os valores de β_0 e β_1 que minimizam a soma dos quadrados dos erros, que são $\hat{\beta}_1$ e $\hat{\beta}_0$, os estimadores de mínimos quadrados de β_1 e β_0 respectivamente. Assim, obtem-se a reta que melhor explica a relação linear entre as variáveis, que é

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i,$$

sendo

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \quad \text{e} \quad \hat{\beta}_1 = \frac{S_{xy}}{S_{xx}},$$

com

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

Além disso, por meio das suposições do modelo, pode-se mostrar que as seguintes relações são satisfeitas

$$\hat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma^2}{S_{xx}}\right) \quad \text{e} \quad \hat{\beta}_0 \sim N\left[\beta_0, \sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)\right],$$

e ainda

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

Outro método para estimação dos parâmetros é o de **máxima verossimilhança**, em que, determinam-se as estimativas de máxima verossimilhança dos parâmetros β_0 e β_1 do modelo de regressão dado pela Equação (2.3). Levando-se em consideração as suposições consideradas para o modelo tem-se que:

$$\varepsilon_i \sim N\left(0, \sigma^2\right),$$

E consequentemente, para um certo valor y_i, a função densidade de probabilidade será:

$$f(y_i) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}\left[y_i - (\beta_0 + \beta_1 x_i)\right]^2\right\}.$$

Observando-se uma amostra de n pares de valores (x_i, y_i), sendo os valores de x_i fixos e as observações são independentes, a função de verossimilhança da amostra será, neste caso, definida por

$$L(x_1, \cdots, x_n; \beta_0, \beta_1, \sigma^2) = \left(2\pi\sigma^2\right)^{-n/2} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left[y_i - (\beta_0 + \beta_1 x_i)\right]^2\right\}. \qquad (2.4)$$

Os estimadores de máxima verossimilhança de β_0, β_1 e σ^2 são aqueles que maximizam a função de verossimilhança $L(\beta_0, \beta_1, \sigma^2 | x_1, \cdots, x_n)$. Uma vez que β_0 e β_1 só aparecem no expoente negativo da Equação (2.4), conclui-se que o máximo da função corresponde ao mínimo de:

$$\sum_{i=1}^{n}\left[y_i - (\beta_0 + \beta_1 x_i)\right]^2.$$

Desta forma, as estimativas de máxima verossimilhança dos parâmetros β_0 e β_1 são equivalentes as estimativas de mínimos quadrados, levando em consideração que a distribuição dos erros seja normal (HOFFMANN, 2006). Usando-se o método de máxima verossimilhança, obtem-se também um estimador para a variância dos erros, σ^2, que é

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}\varepsilon_i^2.$$

Enfim, estima-se β_0 e β_1 por dois métodos diferentes (mínimos quadrados e de máxima verossimilhança) e em ambos os casos encontra-se os mesmos estimadores para β_0 e para β_1 e ainda encontra-se um possível estimador para $\hat{\sigma}^2$ que é dado pela média dos quadrados dos resíduos. Como os estimadores de máxima verossimilhança de β_0 e β_1 são os mesmos do método de mínimos quadrados, eles tem as mesmas propriedades de todos os estimadores de mínimos quadrados. As demonstrações dos estimadores podem ser encontradas com detalhes em Charnet et al. (2008).

Vale lembrar que esse trabalho será desenvolvido com o uso de análise de regressão linear múltipla com variável *dummy*, por esse detalhe, algumas técnicas e suas respectivas demonstrações não serão discutidas nessa seção sobre análise de regressão linear simples.

2.2.2 Modelos de regressão linear múltipla (MRLM)

Na regressão linear múltipla admite-se que a variável dependente Y é função de duas ou mais variáveis regressoras, ou seja, descreve a variável de interesse Y, com sendo uma soma da parte determinística mais geral e a parte aleatória, em que, o seu valor esperado pode ser expresso como função de várias variáveis regressoras e como função de polinômio de maior grau de uma única variável regressora.

O modelo estatístico de uma regressão linear múltipla com k variáveis regressoras $(X_1, X_2, \cdots, X_k)$, é representada por

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_k x_{ki} + \varepsilon_i, \quad i = 1, 2, \cdots, n.$$

ou ainda

$$y_i = \beta_0 + \sum_{j=1}^{k} \beta_j x_{ji} + \varepsilon_i,$$

em que, y_i é a variação dos valores das variáveis regressoras x_k, ε_i são os erros associados ao modelo, k é o número de variáveis explicativas para o modelo e n é o tamanho da amostra.

Segundo Charnet et al. (2008), o modelo estatístico polinomial com uma variável regressora é definido por

$$y_i = \beta_0 + \beta_1 x + \cdots + \beta_k x^k + \varepsilon_i,$$

em que, x é considerado o valor fixo para a variável regressora X, os parâmetros $\beta_0 + \beta_1 x + \cdots + \beta_k$ são os coeficientes do polinomio de grau k, que define a esperança de Y.

O modelo hiperplano com três variáveis regressoras é representado por

$$y_i = \beta_0 + \beta_1 x + \beta_2 x_2 + \beta_3 x_3 + \varepsilon_i,$$

em que, x_j é o valor fixo da variável regressora X_j com (j = 1, 2, 3); os parâmetros β_j com (j = 1, 2, 3) pode ser interpretado como a mudança esperada em Y devido ao aumento

de uma unidade de X_j estando-se as outras variáveis X_k com $k \neq j$. A esperança de Y é definido por

$$\beta_0 + \beta_1 x + \beta_2 x_2 + \beta_3 x_3,$$

se aumentar a variável X_1 em uma unidade, e considerar fixa as outras variáveis regressoras, a esperança de Y, será:

$$\beta_0 + \beta_1 (x_1 + 1) + \beta_2 x_2 + \beta_3 x_3.$$

Assim, a diferença entre as duas esperanças será:

$$\beta_0 + \beta_1 x + \beta_2 x_2 + \beta_3 x_3 - (\beta_0 + \beta_1 (x_1 + 1) + \beta_2 x_2 + \beta_3 x_3) = \beta_1.$$

Pode ser visto, desde que as regressoras sejam consideradas fixas as alterações ocorridas na esperança de Y é correspondente ao coeficiente β_1 (CHARNET et al. , 2008).

Já o modelo de duas variáveis regressoras e interação é definido por

$$\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \varepsilon_i,$$

em que, x_i é o valor fixo da variável regressora X_i com (i = 1, 2).

De acordo com Charnet et al. (2008), a expressão linear nos modelos descritos acima, deve-se ao fato que a esperança de Y, para valores fixos das variáveis regressoras são função linear dos parâmetros.

Em notação matricial, o modelo de regressão linear múltipla pode ser escrito na forma:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{2.5}$$

sendo,

$$\mathbf{Y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \quad \mathbf{X} = \begin{bmatrix} 1 & x_{11} & x_{21} & \cdots & x_{k1} \\ 1 & x_{12} & x_{22} & \cdots & x_{k2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{1n} & x_{2n} & \cdots & x_{kn} \end{bmatrix}, \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix} \quad \text{e} \quad \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

, em que, $\boldsymbol{Y}$ é um vetor, de dimensão $n \times 1$, da variável aleatória $\boldsymbol{Y}$, $\boldsymbol{X}$ é a matriz, de dimensões, $n \times p$, denominada matriz do modelo, $\boldsymbol{\beta}$ é o vetor, de dimensão $p \times 1$, de parâmetros desconhecidos, $\boldsymbol{\varepsilon}$ é o vetor, de dimensão $n \times 1$ e de variáveis aleatórias não observáveis. Tal representação simplifica a notação e os cálculos a serem realizados

futuramente. (HOFFMANN, 2006)

2.2.2.1 Presuposições para modelo

De forma semelhante ao que foi visto em regressão linear simples, têm-se as seguintes suposições:

1. A variável Y é função linear das variáveis explicativas $X_j, \quad j = 1, 2, \cdots, k$;
2. Os valores das variáveis explicativas X_j são consideradas fixas;
3. $E(\varepsilon_i) = \mathbf{0}$, ou seja, $E(\boldsymbol{\varepsilon}) = \mathbf{0}$, sendo $\mathbf{0}$ um vetor de zeros de dimensão $n \times 1$;
4. Os erros são homocedásticos, isto é, $Var(\varepsilon_i) = E(\varepsilon_i^2) = \sigma^2$;
5. Os erros são independentes, isto é, $Cov(\varepsilon_i, \varepsilon_j) = E(\varepsilon_i \varepsilon_j) = 0$, $i \neq j$;
6. Os erros têm distribuição normal.

A suposição de normalidade dos erros não utiliza-se na estimação, e sim na elaboração dos testes de hipóteses e na obtenção de intervalos de confiança.

2.2.2.2 Estimação dos parâmetros do modelo

O número de parâmetros a serem estimados é $p = k + 1$. Se existirem p observções, a estimação dos parâmetros reduz-se a um problema matemático de resolução de um sistema de p equações e p incógnitas, não sendo possível fazer qualquer análise estatística. Então, deve-se ter $n > p$. (HOFFMANN, 2006)

Usando-se a notação matricial para a Equação (2.5), o vetor dos erros, será:

$$\boldsymbol{\varepsilon} = \mathbf{Y} - \mathbf{X}\boldsymbol{\beta}. \tag{2.6}$$

De uma forma geral, melhor será o modelo quanto menor for o comprimento de $\boldsymbol{\varepsilon}$. Usando-se a norma Euclidiana para o comprimento de $\boldsymbol{\varepsilon}$, tem-se que a soma dos mínimos quadrados é definida por

$$Z = ||\boldsymbol{\varepsilon}||^2 = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} \varepsilon_i \varepsilon_i = \varepsilon_1\varepsilon_1 + \varepsilon_2\varepsilon_2 + \cdots + \varepsilon_n\varepsilon_n = \begin{bmatrix} \varepsilon_1 & \varepsilon_2 & \cdots & \varepsilon_n \end{bmatrix} \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix} = \varepsilon'\varepsilon$$

Substituindo-se a Equação (2.6) na Equação acima e usando-se propriedades de matrizes, obtem-se

$$\begin{aligned} Z = \boldsymbol{\varepsilon}'\boldsymbol{\varepsilon} &= (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}) \\ &= (\boldsymbol{Y}' - \boldsymbol{\beta}'\boldsymbol{X}')(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}) \\ &= \boldsymbol{Y}'\boldsymbol{Y} - \boldsymbol{Y}'\boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{\beta}'\boldsymbol{X}'\boldsymbol{Y} + \boldsymbol{\beta}'\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta} \end{aligned}$$

em que, $\boldsymbol{Y}'\boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{\beta}'\boldsymbol{X}'\boldsymbol{Y}$, pois uma matriz é transposta da outra, e possuem um único elemento, logo

$$Z = \boldsymbol{Y}'\boldsymbol{Y} - 2\boldsymbol{\beta}'\boldsymbol{X}'\boldsymbol{Y} + \boldsymbol{\beta}'\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta} \tag{2.7}$$

O ponto de mínimo da função Z para os valores de $\boldsymbol{\beta}$, é aquele que torna a equação diferencialvel de Z

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

identicamente nula, ou seja:

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

como, $(\partial\boldsymbol{\beta}')\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta}$ e $\boldsymbol{\beta}'\boldsymbol{X}'\boldsymbol{X}(\partial\boldsymbol{\beta})$ são matrizes transposta uma da outra e ambas têm apenas um elemento, $(\partial\boldsymbol{\beta}')\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{\beta}'\boldsymbol{X}'\boldsymbol{X}(\partial\boldsymbol{\beta})$. Daí

$$\frac{\partial Z}{\partial \boldsymbol{\beta}} \equiv 0 \quad \Longleftrightarrow \quad 2(\partial\boldsymbol{\beta}')(-\boldsymbol{X}'\boldsymbol{Y} + \boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta}) \equiv 0.$$

Segue que o vetor $\hat{\boldsymbol{\beta}}$ dos estimadores dos parâmetros deve ser solução do sistema de equações

$$\boldsymbol{X}'\boldsymbol{X}\beta - \boldsymbol{X}'\boldsymbol{Y} = 0, \tag{2.8}$$

que é chamado **sistema de equações normais**. Em outras palavras, a solução do sistema de equações normais, quando existir, nos fornecerá os estimadores $\hat{\boldsymbol{\beta}}$ dos parâmetros $\boldsymbol{\beta}$. O sistema de equações normais definido em (2.8) pode ser ainda reescrito como

$$\boldsymbol{X}'\boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{X}'\boldsymbol{Y}, \tag{2.9}$$

em que,

$$\boldsymbol{X'X} = \begin{bmatrix} n & \sum_{i=1}^{n} X_{i1} & \sum_{i=1}^{n} X_{i2} & \cdots & \sum_{i=1}^{n} X_{ik} \\ \sum_{i=1}^{n} X_{i1} & \sum_{i=1}^{n} X_{i1}^2 & \sum_{i=1}^{n} X_{i1}X_{i2} & \cdots & \sum_{i=1}^{n} X_{i1}X_{ik} \\ \sum_{i=1}^{n} X_{i2} & \sum_{i=1}^{n} X_{i1}X_{i2} & \sum_{i=1}^{n} X_{i2}^2 & \cdots & \sum_{i=1}^{n} X_{i2}X_{ik} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \sum_{i=1}^{n} X_{ik} & \sum_{i=1}^{n} X_{i1}X_{ik} & \sum_{i=1}^{n} X_{i2}X_{ik} & \cdots & \sum_{i=1}^{n} X_{ik}^2 \end{bmatrix},$$

$$\hat{\beta} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \vdots \\ \hat{\boldsymbol{\beta}}_k \end{bmatrix} \quad \text{e} \quad \boldsymbol{X'Y} = \begin{bmatrix} \sum_{i=1}^{n} Y_i \\ \sum_{i=1}^{n} X_{i1}Y_i \\ \sum_{i=1}^{n} X_{i2}Y_i \\ \vdots \\ \sum_{i=1}^{n} X_{ik}Y_i \end{bmatrix}.$$

No caso em que a matriz $\boldsymbol{X'X}$ for não singular, e portanto invertível, conclui-se que o estimador para o vetor de parâmetros $\hat{\boldsymbol{\beta}}$ é definido por

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

De acordo com Hoffmann (2006), alguns resultados são extremamente importantes, sobre sistema de equações normais:

$$\begin{aligned} \boldsymbol{X'Y} - \boldsymbol{X'X\beta} &= 0 \\ \boldsymbol{X'}(\boldsymbol{Y} - \boldsymbol{X\beta}) &= 0 \\ \boldsymbol{X'\varepsilon} &= 0. \end{aligned}$$

Por essa relação matricial significa dizer que:

$$\sum_{i=1}^{n} \varepsilon_i = 0.$$

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

$$\sum_{i=1}^{n} \boldsymbol{Y}_i = \sum_{i=1}^{n} \hat{\boldsymbol{Y}}_i. \tag{2.10}$$

Ao invés de calcular esperanças e variâncias separadamente para cada elemento de $\hat{\beta}$,

usando notação matricial obtêm-se simultaneamente, então :

$$\begin{aligned}
E(\hat{\boldsymbol{\beta}}) &= E[(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}] \\
&= E[(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'(\boldsymbol{X}\boldsymbol{\beta}+\boldsymbol{\varepsilon})] \\
&= E[\underbrace{(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{X}}_{I}\boldsymbol{\beta} + (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\varepsilon}] \\
&= E[\boldsymbol{I}\boldsymbol{\beta}] + E[(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\varepsilon}] \\
&= \boldsymbol{\beta} + (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\underbrace{E[\boldsymbol{\varepsilon}]}_{0} \\
E(\hat{\boldsymbol{\beta}}) &= \boldsymbol{\beta}.
\end{aligned}$$

$$\begin{aligned}
Var(\hat{\boldsymbol{\beta}}) &= Var[(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}] \\
&= (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\underbrace{Var(\boldsymbol{Y})}_{\sigma^2}[(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}']' \\
&= \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'[(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}']' \\
&= \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}\underbrace{\boldsymbol{X}'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}}_{\boldsymbol{I}} \\
Var(\hat{\boldsymbol{\beta}}) &= \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}.
\end{aligned}$$

Então,

$$\hat{\boldsymbol{\beta}} \sim N(\boldsymbol{\beta}, \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}).$$

As estimativas das variâncias destes coeficientes de regressão são obtidas substituindo-se σ^2 pelo seu estimador. O erro padrão estimado de $\hat{\beta}_j$ é encontrado e a raiz quadrada da variância estimada para o j-ésimo coeficiente de regressão, então

$$S(\hat{\beta}_j) = \sqrt{\hat{\sigma}^2(\boldsymbol{X}'\boldsymbol{X})^{-1}}.$$

Seja x_i' a i-ésima linha da matriz $\boldsymbol{X}$, isto é, o vetor linha com valores das variáveis explicativas, em que, o primeiro elemento é igual a 1. O estimador de mínimos quadrados da esperança de $\boldsymbol{Y}$, correspondente aos valores de x_i', é definido por

$$\hat{y}_i = x_i'\hat{\boldsymbol{\beta}}.$$

Este estimador tem distribuição Normal univariada, com esperança e variância respecti-

vamente iguais a

$$E(x_i'\hat{\boldsymbol{\beta}}) = x_i' E[\hat{\boldsymbol{\beta}}] = x_i'\boldsymbol{\beta}.$$

$$\begin{aligned} Var(x_i'\hat{\boldsymbol{\beta}}) &= x_i' Var[\hat{\boldsymbol{\beta}}] x_i \\ &= x_i'\sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}x_i \\ Var(x_i'\hat{\boldsymbol{\beta}}) &= \sigma^2 x_i'(\boldsymbol{X}'\boldsymbol{X})^{-1}x_i. \end{aligned}$$

Logo,

$$\hat{y}_i = x_i'\hat{\boldsymbol{\beta}} \sim N(x_i'\boldsymbol{\beta}, \sigma^2 x_i'(\boldsymbol{X}'\boldsymbol{X})^{-1}x_i).$$

Os resultados descritos acima podem ser encontrados em Charnet et al. (2008).

2.2.2.3 Soma de quadrados

i) Soma de quadrados de resíduos (SQ_{Res})

Para calcular soma de quadrados dos desvios, ou soma de quadrados residual, é necessário relembrar a Equação (2.7), em que

$$\begin{aligned} \boldsymbol{\varepsilon}'\boldsymbol{\varepsilon} &= \boldsymbol{Y}'\boldsymbol{Y} - 2\boldsymbol{\beta}'\boldsymbol{X}'\boldsymbol{Y} + \boldsymbol{\beta}'\underbrace{\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta}}_{\boldsymbol{X}'\boldsymbol{Y}} \\ &= \boldsymbol{Y}'\boldsymbol{Y} - 2\boldsymbol{\beta}'\boldsymbol{X}'\boldsymbol{Y} + \boldsymbol{\beta}'\boldsymbol{X}'\boldsymbol{Y} \\ SQ_{Res} &= \boldsymbol{Y}'\boldsymbol{Y} - \boldsymbol{\beta}'\boldsymbol{X}'\boldsymbol{Y}. \end{aligned}$$

ii) Soma de quadrados total (SQ_{Total})

A soma de quadrado total, mede a variação total das observações em torno da média. Tem-se a expressão

$$SQ_{Total} = \sum_{i=1}^{n} y_i^2 = \sum_{i=1}^{n} Y_i^2 - \frac{\left(\sum_{i=1}^{n} Y_i\right)^2}{n} = \boldsymbol{Y}'\boldsymbol{Y} - \frac{\left(\sum_{i=1}^{n} Y_i\right)^2}{n}.$$

iii) Soma de quadrados de regressão (SQ_{Reg})

A soma de quadrado de regressão, mede a quantidade de variação da variável dependente explicada pela equação de regressão linear múltipla. Então a expressão é

definida por

$$SQ_{Reg} = \sum_{i=1}^{n}(\hat{\boldsymbol{Y}}_i - \bar{\boldsymbol{Y}})^2 = \sum_{i=1}^{n}\hat{y}_i^2 = \sum_{i=1}^{n}\hat{Y}_i^2 - \frac{\left(\sum_{i=1}^{n}\hat{Y}_i\right)^2}{n} = \hat{\boldsymbol{Y}}'\hat{\boldsymbol{Y}} - \frac{\left(\sum_{i=1}^{n}\hat{Y}_i\right)^2}{n}$$
$$= (\boldsymbol{X}\hat{\boldsymbol{\beta}})'\boldsymbol{X}\hat{\boldsymbol{\beta}} - \frac{\left(\sum_{i=1}^{n}\hat{Y}_i\right)^2}{n} = \hat{\boldsymbol{\beta}}'\boldsymbol{X}'\boldsymbol{X}\hat{\boldsymbol{\beta}} - \frac{\left(\sum_{i=1}^{n}\hat{Y}_i\right)^2}{n} = \hat{\boldsymbol{\beta}}'\boldsymbol{X}'\boldsymbol{Y} - \frac{\left(\sum_{i=1}^{n}\hat{Y}_i\right)^2}{n}.$$

De acordo com a Equação (2.10), então:

$$SQ_{Reg} = \hat{\boldsymbol{\beta}}'\boldsymbol{X}'\boldsymbol{Y} - \frac{\left(\sum_{i=1}^{n}Y_i\right)^2}{n}.$$

2.2.2.4 Testes de hipótese e intervalo de confiança

Segundo Queiroz (2011), após a estimação dos parâmetros, em geral, realizam-se testes a fim de determinar se hipóteses realizada sobre tais parâmetros são suportadas por evidências obtidas por meio de dados amostrais. Ou melhor, é importante avaliar se existe uma boa correlação entre a variável resposta e a variável explicativa. Por exemplo, se o aumento da variável explicativa acarretará em uma mudança significativa ou não no valor esperado da variável resposta. Há dois testes que podem ser aplicados para verificar a tal mudança significativa, o teste t de Student e o F de Snedecor.

i) Teste de significância para o modelo de regressão (Teste F)

O teste F é utilizado para verificar se as variáveis independentes conjuntamente, contribuem significativamente para explicar a variação da variável resposta. Definindo-se às hipóteses

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

A estatística teste será

$$F = \frac{QM_{Reg}}{QM_{Res}} \sim F_{(k,n-p)},$$

em que, k é o número de váriaveis independentes e $p = k+1$. Então, após encontrar o valor F calculado, o F tabelado, e atribuir o nível α de siginificância, pode-se decidir que:

- Se $F_{calculado} > F_{tabelado}$, rejeita-se a hipótese H_0 e conclui-se ao nível α de significância que há regressão;

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

Pode-se resumir o procedimento descrito em uma Tabela da Análise de Variância (ANOVA), conforme representado na Tabela 1.

Tabela 1: Análise de variância para o modelo de regressão linear múltipla

Fonte de Variação	GL	SQ	QM	F
Regressão	k	SQ_{Reg}	$\frac{SQ_{reg}}{k}$	$\frac{QM_{Reg}}{QM_{Res}}$
Resíduo	$n-p$	SQ_{Res}	$\frac{SQ_{reg}}{n-p}$	-
Total	$n-1$	SQ_{Tot}	-	-

ii) Teste de significância para os coeficientes de regressão (Teste t-student)

Muitas vezes é de interesse do pesquisador testar hipóteses acerca dos coeficientes de regressão, para determinar o potencial de cada regressor no modelo. Segundo Charnet *et al.* (2008), para medir a significância das variáveis do modelo individualmente, para cada $j = 1, 2, \cdots, k$, testa-se

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

Neste caso, a estatística do teste é

$$T = \frac{\hat{\beta}_j - \beta_j}{S(\hat{\beta}_j)} \sim t_{n-p} \quad j = 1, 2, \cdots, k, \tag{2.11}$$

e a regra de decisão é

– Se $t_{cal} > t_{tab}$, rejeita-se a hipótese H_0, e conclui-se ao nível α de significância, que a variável não pode ser eliminada do modelo, pois explica bem a regressão linear;

– Se $t_{cal} < t_{tab}$, aceita-se a hipótese H_0, e ao nível α de significância, conclui-se que a variável pode ser eliminada do modelo sem tanto dano para a explicação da regresssão linear.

Intervalo de confiança para os coeficientes de regressão

Outra forma de se avaliar a significância dos parâmetros do modelo é por meio da construção de intervalos de confiança. Podendo-se encontrar um intervalo de confiança

que contenha o verdadeiro valor do parâmetro β_j com $j = 1, 2, \cdots, k$, a um certo nível de significância α, que se queira.

Considerando-se a estatística teste dada em (2.11), um intervalo com $100(1-\alpha)\%$ de confiança para o coeficiente da regressão β_j, $j = 1, 2, \cdots, k$, é definido por

$$IC(\beta_j) = \left[\hat{\beta}_j - t_{(\frac{\alpha}{2}, n-p-1)}\sqrt{S(\hat{\beta}_j)}; \hat{\beta}_j + t_{(\frac{\alpha}{2}, n-p-1)}\sqrt{S(\hat{\beta}_j)}\right]$$

2.2.2.5 Análise de resíduos no MRLM

Para que os resultados de uma análise de regressão sejam confiáveis, tanto no MRLS quanto MRLM, é fundamental que as suposições do modelo ajustado sejam válidas. Se as suposições são violadas, têm-se falhas sistemáticas, ou seja, não linearidade, não normalidade, heterocedasticidade, não independência dos erros, e presença de pontos atípicos, e então, levando-se à análises com conclusões duvidosas.

Desta forma a análise de resíduos desempenha um papel fundamental, pois, oferece técnicas que nos ajudam a verificar a presença de indícios da adequabilidade do modelo por meio dos resíduos.

Como visto anteriormente, o vetor de resíduos é definido por

$$\boldsymbol{\varepsilon} = \boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}.$$

e lembrando alguns resultados importantes:

$$\boldsymbol{y} \sim N(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2 I),$$

$$\hat{\boldsymbol{\beta}} \sim N(\boldsymbol{\beta}, \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}),$$

$$\hat{y}_i = x_i'\hat{\boldsymbol{\beta}} \sim N(x_i'\boldsymbol{\beta}, \sigma^2 x_i'(\boldsymbol{X}'\boldsymbol{X})^{-1}x_i).$$

então, a esperança e a variância dos resíduos são definidas respectivamente por

$$E[\boldsymbol{\varepsilon}] = E[\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}] = 0.$$

e

$$\begin{aligned} Var[\boldsymbol{\varepsilon}] &= Var[\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}] \\ &= \sigma^2[\boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}']. \end{aligned}$$

o que pode ser reescrito da seguinte forma:

$$\boldsymbol{\varepsilon} \sim N(0, \sigma^2[\boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}']).$$

Segundo Hoffmann (2006), a matriz $\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$ é considerada matriz de projeção $\boldsymbol{H}$, em que, é simétrica e idempotente e os valores da diagonal principal da matriz $\boldsymbol{H}$ são h_{ii}, com $0 < h_{ii} < 1$ e $i = 1, 2, \cdots, n$. Em que, h_{ii} é o valor observado da influência de x_i a $\bar{x}$.

Utiliza-se algumas técnicas para verificar as suposições do modelo, podem ser informais (como gráficos) ou formais (como testes), em que, são mais indicadas para a tomada de decisão. O ideal é combinar as técnicas disponíveis, para o diagnóstico de problemas nas suposições do modelo. Para cada suposição do modelo, descreve-se com detalhes as técnicas para diagnóstico.

1. **Diagnóstico de normalidade**

   A normalidade nos resíduos é uma suposição muito importante para que sejam confiáveis os resultados a respeito do ajuste do modelo de regressão linear. Essa suposição é verificada por

   i) **Gráfico de probabilidade normal - Q-Q Plot** - Quantil de probabilidade esperado para a distribuição normal, em função dos resíduos.

   ii) **Teste Shapiro-Wilk** - Segundo Ferreira (2009) o teste de Shapiro-Wilk é baseado em estatísticas de ordem da distribuição normal e de seus respectivos valores esperados. Supondo-se que a partir de uma população normal sejam retirada amostras aleatória de tamanho n, com $(X_1, X_2, ..., X_n)$, em que, os valores das amostras são ordenadas em forma crescente.

   Testando-se as hipóteses

   $$\begin{cases} H_0\text{: A amostra provém de uma população Normal;} \\ H_1\text{: A amostra não provém de uma população Normal} \end{cases}$$

   A estatística do teste de Shapiro-Wilk (1965), é representado pela seguinte expressão:

   $$W = \frac{\left[\sum_{i=1}^{n} a_i X_{(i)}\right]^2}{(n-1)S^2},$$

   em que, a_i é o melhor estimador linear não-viesado normalizado do valor esperado das estatísticas de ordem da distribuição normal padrão e $X_{(i)}$ os valores

das amostras ordenadas de forma crescente $(X_{(1)}, X_{(2)}, ..., X_{(n)})$. Realiza-se o teste escolhendo o nível α de probabilidade.

2. **Diagnóstico de homoscedasticidade**

A falta de homoscedasticidade é chamado de heteroscedasticidade, ou seja, quando há heteroscedasticidade as variâncias não são constantes em diferentes observações, daí o modelo sofre alguns efeitos no seu ajuste. A suposição de homoscedasticidade é testada pelas técnicas a seguir:

i) **Gráfico dos resíduos *versus* valores ajustados** - Por meio de alguma tendência nos pontos pode-se identificar se há detecção de heteroscedasticidade da variância dos erros, se os pontos estão aleatoriamente distribuídos em torno do 0, sem nenhum comportamento, há indícios de que a variância dos resíduos é homoscedástica.

ii) **Teste de Goldfeld-Quandt** - A exigência do teste de Goldfeld-Quandt é de que a amostra seja relativamente grande.

Segundo Rodrigues e Diniz (2006) as n observações são ordenadas de acordo com os valores da variável regressora, divide-se a amostra ordenada em 3 partes, em que, a parte do meio deve ter 25% dos dados, a 1^{o} contendo os menores valores da variável explicativa e a 3^{o} contendo os maiores valores da variável explicativa, em que deve-se apresentar praticamente a mesma quantidade de dados. De posse dessas três partes, ajusta-se dois modelos de regressão, um com os dados da 1^{o} parte e outro com os dados da 3^{o} parte.

Por fim, utilizando-se o teste F, testando-se as seguintes hipóteses:

$$\begin{cases} H_0 : \sigma_1^2 = \sigma_3^2; \\ H_1 : \sigma_3^2 > \sigma_1^2 \end{cases}$$

em que, σ_i^2 com $i = 1, 2, 3$ é a variância dos resíduos dos três modelos de regressão.

A estatística de teste é definido por

$$F_{GQ} = \frac{SQ_{Res^3}/(n_3 - (p+1))}{SQ_{Res^1}/(n_1 - (p+1))},$$

em que, SQ_{Res^1} e SQ_{Res^3}, são as somas de quadrados dos resíduos dos modelos de 1^{o} e 3^{o} parte, n_1 e n_3 é o número de observações da 1^{o} e 3^{o} parte dos valores da variável regressora e p o número de observações da 2^{o} parte. O $F_{tabelado} = F_{(n_3(p+1), n_1-(p+1))}$, então rejeita-se a hipótese nula se $F_{GQ} > F(\alpha)$.

3. **Diagnóstico de independência**

   Independência dos erros é um acontecimento aleatório que ocorre em um determinado período de tempo, em que, um resíduo não afeta nos resíduos seguintes. Esse diagnóstico é verificado da seguinte forma:

   i) **Gráfico dos resíduos *versus* a ordem de coleta** - Ao avaliar o gráfico e perceber alguma tendência nos pontos, ou seja, se os pontos repetem-se em um determinado ambiente do gráfico há indícios de dependência dos resíduos.

   ii) **Teste de Durbin-Watson** - De acordo com Montgomery (2003) o teste de Durbin-Watson é utilizado para a detectar a presença de autocorrelação nos resíduos de um modelo de regressão. Testa-se a presença de autocorrelação por meio da hipótese:

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

   A estatística teste é representada por:

   $$DW = \frac{\sum_{i=2}^{n} (e_i - e_{i-1})^2}{\sum_{i=1}^{n} e_i^2},$$

   em que, $e_i = y_i - \hat{y}_i$ é o resíduo associado a i-ésima observação. O valor da estatística deve variar de $0 \leq DW \leq 4$, para a tomada de decisão compara-se o valor da estatística DW com os valores críticos D_L e D_U, daí toma-se a decisão de acordo com:

   - Se $DW < D_L$ rejeita-se a hipótese $H_0 : \rho = 0$;
   - Se $DW > D_L$ aceita-se a hipótese $H_0 : \rho = 0$;
   - Se $D_L < DW < D_U$ o teste é inconclusivo.

4. **Diagnóstico de *outliers***

   Outlier é uma observação com o comportamento diferente das demais. Desta forma, pode ser um *outlier* em relação a Y ou aos X, e pode ou não ser um ponto influente.

   i) ***Outliers* com relação a variável X**

   Para identificar um *outliers* em X, utilizam-se os valores h_{ii} da matriz de projeção, observa se há valor extremo do h_{ii} em um box-plot

   ii) ***Outliers* com relação a variável Y**

   Os resíduos são definidos no modelo (2.6), para detectar melhor *outliers* na

variável Y, os resíduos foram modificados por

- Resíduos padronizados

  O resíduo padronizado não tem boas propriedades, por não ter variância constante, muda cada valor de X_i. Se os erros seguem uma distribuição normal, 95% dos resíduos padronizados devem está no intervalo entre (-3,3), se não, podem indicar a presença de *outlier*. O resíduo padronizado é definido por

$$d_i = \frac{\varepsilon_i}{\sqrt{QM_{Res}}}, \quad i = 1, 2, \cdots, n.$$

- Resíduos estudentizados

  Os resíduos estudentizados tem variância constante e igual a 1, ajudando-se a encontrar com maior facilidade *outliers*. Desta forma, os resíduos estudentizados são definidos por:

$$r_i = \frac{\varepsilon_i}{\sqrt{QM_{Res}(1 - h_{ii})}}, \quad i = 1, 2, \cdots, n.$$

Se após a realização da análise de resíduo, constata-se que não foi possível satisfazer presuposição para o modelo linear clássico, é possível que uma transformação não linear dos dados possa produzir a homogeneidade da variância e a distribuição aproximadamente normal dos resíduos.

A transformação Box-Cox identifica uma transformação a partir de uma família de transformação de potência de Y, a fim de encontrar a transformação que estabilize ou reduza a vareabilidade existente e normalidade dos resíduos.

Box e Cox (1964) propuseram uma família de transformação definida por

$$Y(\lambda) = \begin{cases} \frac{Y^{\lambda}-1}{\lambda}, & \lambda \neq 0; \\ log(Y), & \lambda = 0 \end{cases} \tag{2.12}$$

em que, λ é o parâmetro de transformação e Y a variável resposta. Quando $\lambda = 1$ não é necessário a realização de transformação e quando $\lambda = 0$ utiliza-se a transformação logarítimica.

Nas observações $(Y_i, \boldsymbol{x}_i')$, $\quad i = 1, 2, ..., n \quad$ e $\quad \boldsymbol{x}_i' = (X_{1i}, X_{2i}, ..., X_{ki})$, tem-se que

$$Y_i(\lambda) \sim N(\boldsymbol{x}_i'\boldsymbol{\beta}, \sigma^2), \quad i = 1, 2, ..., n.$$

Para escolha da melhor potência para λ consideram-se valores no intervalo de [-2, 2], conforme descrevem Draper e Smith (1998). Se no gráfico da verossimilhança perfilhada o

valor 0 estiver contido no intervalo, é indicado a utilização da transformação logarítimica da variável, pois os resultados serão bem próximos dos obtidos com a transformação préviamente adotada.

Ao realizar a transformação na variável Y, as estimações e predições são expressas em novas unidades, de acordo com cada transformação admitida. Portanto é um problema que não pode ser esquecido, o retorno à escala normal.

Para facilitar esse retorno, Miller (1984) sugere o estimador $E[Y|x]$, definido por

$$E[Y|x] = exp(\hat{\beta}_0 + \hat{\beta}_1 x_1 + ... + \hat{\beta}_k x_k) exp\left(\frac{\hat{\sigma}^2}{2}\right),$$

em que, $\hat{\beta}_0 + \hat{\beta}_1 x_1 + ... + \hat{\beta}_k x_k$ é o ajuste do modelo na escala transformada e $\hat{\sigma}^2$ é o quadrado médio do resíduo também na escala transformada.

Se o valor de $\hat{\sigma}^2$ for pequeno, há uma outra linha de desenvolvimento, em que, Taylor (1986) propôs

$$E[Y|x] = exp(\hat{\beta}_0 + \hat{\beta}_1 x_1 + ... + \hat{\beta}_k x_k) exp\left(\frac{1 + \hat{\sigma}^2}{2}\right),$$

dado que o $\hat{\sigma}^2 \approx 0$ os dois estimadores acima praticamente coincidirão.

Outras transformações são adotadas de acordo com a necessidade dos dados, algumas delas são destacadas a seguir:

- Raiz quadrada ($\hat{Y} = \sqrt{Y}$) é utilizada para estabilizar a variância quando é proporcional à média dos $Y's$. O estimador para $E[Y|x]$ proposto por Miller (1984), será

$$E[Y|x] = (\hat{\beta}_0 + \hat{\beta}_1 x_1 + ... + \hat{\beta}_k x_k)^2 + \hat{\sigma}^2.$$

- Transformação recíproca ($\hat{Y} = Y^{-1}$) é utilizada para estabilizar a variância, minimizando possíveis alto valores da variável Y. O estimador para $E[Y|x]$ proposto por Miller (1984), será

$$E[Y|x] = (\hat{\beta}_0 + \hat{\beta}_1 x_1 + ... + \hat{\beta}_k x_k)^{-1} + \frac{\hat{\sigma}^2}{(\hat{\beta}_0 + \hat{\beta}_1 x_1 + ... + \hat{\beta}_k x_k)^2}.$$

2.2.3 Seleção de variáveis regressoras

Em modelos de regressão múltipla é necessário determinar um subconjunto de variáveis independentes que melhor explique a variável resposta, isto é, dentre todas as variáveis regressoras disponíveis, deve-se encontrar um subconjunto de variáveis importantes para

o modelo. Com o objetivo de diminuir a variância da estimativa por meio de um modelo com menor número possível de variáveis. Para isto, utiliza-se a técnica, denominada de seleção de variáveis. Existem vários critérios para a seleção de um subconjunto de variáveis regressoras a serem encorporadas ao modelo.

2.2.3.1 Todas as regressões possíveis

i) **Coeficiente de determinação e coeficiente de determinação ajustado**

Segundo Hoffmann (2006), o coeficiente de determinação do modelo de regressão linear múltipla é definido por

$$R^2 = \frac{SQ_{Reg}}{SQ_{Tot}} = 1 - \frac{SQ_{Res}}{SQ_{Tot}}, \quad \text{com} \quad 0 \leq R^2 \leq 1, \tag{2.13}$$

Nem sempre um grande valor de R^2 implica necessariamente que o modelo de regressão tenha um bom ajustamento, ou seja, a medida que adiciona-se uma variável ao modelo, tem-se o aumento na soma de quadrados de regressão, mesmo que a variável seja ou não estatisticamente significativa. Como o (R^2) não obtem um valor esperado de Y confiável, então é preferível que utilize o coeficiente de determinação ajustado (R^2_{ajust}) para graus de liberdade, representado por

$$1 - R^2_{ajust} = 1 - \frac{\frac{SQ_{Res}}{n-p}}{\frac{SQ_{Tot}}{n-1}} = \frac{n-1}{n-p}(1 - R^2),$$

ou melhor

$$R^2_{ajust} = R^2 - \frac{p-1}{n-p}(1 - R^2), \tag{2.14}$$

em que, n é o tamanho da amostra e p é o número de parâmetros. O R^2_{ajust} dá uma ideia da proporção da variação de Y explicada pelo método de regressão uma vez que leva em conta o número de regressores. Ao contrário do que acontecia no coeficiente não ajustado, o ajustado nem sempre aumenta a medida que é adicionada uma nova variável ao modelo.

ii) Critério de informação de Akaike (AIC)

O Critério de Informação de Akaike é bastante utilizado para diferentes estruturas de covariâncias, relaciona-se a discrepância que existe entre o modelo verdadeiro e o aproximado, por meio da máxima verossimilhança. Seleciona-se uma combinação de variáveis exploratórias a modelos para a função de correlação que minimize o valor de AIC (Olinda, 2010). Sendo assim, o Critério de informação de Akaike é

definido por

$$AIC_p = -2log(L_p) + 2[(p+1)+1], \tag{2.15}$$

em que, L_p é a função de máxima verossimilhança e p é o número de variáveis regressoras utilizadas no modelo.

iii) Critério de informação de Bayes BIC

O Critério de informação de Bayes (BIC) penaliza mais fortemente modelos com um maior número de parâmetros do que o AIC. Más tanto o AIC quanto o BIC aumentam conforme SQ_{Res} aumenta.

Ao estimar os parâmetros do modelo usando estimativa de máxima verossimilhança, é possível aumentar a probabilidade de adicionar parâmetros. Daí, menor valor do BIC indica o melhor ajuste do modelo (Olinda, 2010). Desta forma o Critério de informação de Bayes é definido por

$$BIC_p = -2log(L_p) + [(p+1)+1]log(n), \tag{2.16}$$

em que, L_p é a função de máxima verossimilhança e p é o número de variáveis regressoras utilizadas no modelo.

2.2.3.2 Seleção automática

i) Método do passo atrás (*backward*)

Segundo Charnet *et al.* (2008), o método é caracterizado por incorporar todas as variáveis auxiliares em um MRLM e percorrer etapas, em que, uma ou mais variável pode ser eliminada, a que tenha a menor correlação parcial com a resposta Y. Assim, a estatística do teste é definida por

$$F_{parcial} = \frac{SQ^c_{Reg} - SQ^r_{Reg}}{\hat{\sigma}^2}, \tag{2.17}$$

em que, SQ^c_{Reg} e $\hat{\sigma}^2$ são calculados pelo modelo completo, e SQ^r_{Reg} calculada pelo modelo reduzido.

Desta forma, os passos a seguir são baseados no teste F parcial:

1º Passo: Ajustar o modelo completo de m variáveis e calcular SQ^c_{Reg} e $\hat{\sigma}^2$;

2º Passo: Para cada uma das m variáveis do modelo completo do 1º Passo, e com a retirada de uma variável considera-se o modelo reduzido, e calcula-se SQ^r_{Res},

para encontrar o valor de $F_{parcial}$;

3º Passo: Encontra o menor valor de $F_{parcial}$ dos m valores, onde, é denominado de F_{min};

4º Passo: Dado que F_{out} é o quantil especificado da distribuição F, tem-se que;

* Se $F_{min} > F_{out}$ não é eliminado nenhuma variável e para o processo, optando pelo modelo completo com m variáveis;
* Se $F_{min} < F_{out}$ elimina a variável com F_{min} e volta ao 1º Passo com um novo modelo completo com $(m-1)$ variáveis.

ii) Método do passo a frente (*forward*)

Inicialmente considera-se um MRLS, utilizando-se como variável auxiliar a variável com maior coeficiente de correlação para a variável resposta Y. Como utiliza-se um MRLS, então denomina-se modelo reduzido da etapa, em que, compara-se com os modelos que já tenham novas variáveis acrecentadas. Então, o procedimento é constituido das seguintes etapas:

1º Passo: Ajustar o modelo reduzido de m variáveis e calcular SQ^r_{Reg};

2º Passo: Para cada variável que não pertence ao modelo reduzido, então, considera-se como o modelo completo quando adiciona-se variáveis ao modelo reduzido, e calcula-se SQ^c_{Res} e $\hat{\sigma}^2$, para encontrar o valor de $F_{parcial}$;

3º Passo: Encontra o maior valor de $F_{parcial}$ dos m valores, em que, é denominado de F_{max};

4º Passo: Dado que F_{in} é o quantil especificado da distribuição F, tem-se que;

* Se $F_{max} > F_{in}$ inclui-se a variável com o F_{max} e volta ao 1º Passo iniciando-se uma nova etapa cujo modelo reduzido tem $(m+1)$ variáveis;
* Se $F_{max} < F_{in}$ não inclui a variável ao modelo interrompendo o processo.

iii) Método do passo a passo (*stepwise*)

O método passo a passo é uma generalização do método passo a frente, em que, cada passo todas as variáveis do modelo são previamente verificadas pelo valor de $F_{parcial}$. Uma variável adicionada no modelo no passo anterior pode ser redundante para o modelo pelo seu relacionamento com as outras variáveis. As etapas de eliminação e a adição das variáveis são realizadas como nos métodos anteriores. O processo é encerrado quando nenhuma variável é descartada ou incluida no modelo. Então, este procedimento é realizado conforme os seguintes passos:

1º Passo: Ajustar o modelo reduzido de m variáveis e calcular SQ^r_{Reg};

2^{o} Passo: Para cada variável que não pertence ao modelo reduzido, então, considera-se como o modelo completo, calcula-se SQ^{c}_{Res} e $\hat{\sigma}^2$;

3^{o} Passo: Encontra-se o maior valor de $F_{parcial}$ dos m valores, onde, é denominado de F_{max};

4^{o} Passo: Comparar F_{max} com F_{in};

- Se $F_{max} > F_{in}$ inclui-se a variável com o F_{max} e passa-se ao 5^{o} Passo com o modelo completo de $l = m + 1$ variáveis;
- Se $F_{max} < F_{in}$ não inclui a variável com o F_{max} e passa-se ao 5^{o} Passo com o modelo completo de $l = m$ variáveis.

5^{o} Passo: Ajustar o modelo completo com l variáveis, sendo $l = m$ ou $l = m+1$ e calcular SQ^{c}_{Reg} e $\hat{\sigma}^2$

6^{o} Passo: Para cada l variáveis do modelo completo do 4^{o} Passo, considerar o modelo reduzido, com a retirada de uma variável e então calcular a SQ^{r}_{Reg}

7^{o} Passo: Encontra o minimo dos l valores de $F_{parcial}$ encontrados no 6^{o} Passo, que é denominado de F_{min}

8^{o} Passo: Comparar F_{out} com F_{min};

- Se $F_{min} > F_{out}$ não é eliminado nenhuma variável e passa para o 1^{o} Passo iniciando-se uma nova etapa com o novo modelo reduzido com $l = m$ variáveis, e só para o processo no 4^{o} Passo se nenhuma variável for incluída;
- Se $F_{min} < F_{out}$ elimina-se a variável com F_{min} e volta ao 1^{o} Passo com um novo modelo reduzido com $(m = l - 1)$ variáveis.

2.2.4 Variáveis *dummies* na regressão linear

Muitas vezes há necessidade de incluir no modelo de regressão variáveis qualitativas ou categóricas (*dummies*), inclusive em fenômenos econômicos, região, etc. As variáveis quantitativas são de fácil utilização na análise de regressão, o que não acontece com as variáveis qualitativas, uma vez que indicam ausência ou presença de uma qualidade ou tributo. Deste modo, um método de quantificar esses atributos é a construção de variáveis artificias assumindo-se valores binários, ou seja, 0 e 1, em que, será indicada a ausência ou presença do atributo.

De acordo com Missio e Jacobi, (2007) as variáveis *dummies* não precisam ser neces-

sariamente os valores 0 e 1, o par $(0, 1)$ pode ser transformado pela função linear

$$Z = a + bD, \quad b \neq 0,$$

em que, a e b são constantes e D pode ser atribuido os valores 0 ou 1, os valores quantificados da ausência ou presença do atributo.

Quando $D = 1$, $Z = a + b$

Quando $D = 0$, $Z = a$

Então, o par ordenado se torna $(a, a + b)$.

A aplicação da variável *dummy*, pode ser feita em um modelo simples, no qual a variável explicativa é a própria variável *dummy*, como em modelos múltiplos com apenas variáveis *dummies* e modelos múltiplos em que pode combinar variáveis *dummies* com variáveis quantitativas, em que deve-se uma atenção redobrada pois, pode-se ter mudanças no intercepto, na declividade da função, fazendo com que se tenha mudança na estrutura do modelo.

É nitidamente observado que, um simples cálculo de média entre dois valores implica a comparação de dois intervalos, então, exigem que a medida tenha escala intervalar ou cardinal. Para a escala nominal pode-se determinar a moda (não a mediana ou a média da variável). Para uma escala ordinal pode-se determinar tanto a moda quanto a mediana (não a média da variável)(HOFFMANN, 2006).

2.2.4.1 Modelo de regressão com variável *dummy*

Seja um modelo, que assuma uma relação linear entre a variável dependente, e uma ou mais variáveis quantitativas seja estável para todas as observações de uma dada amostra. Tem-se a equação correspondente

$$y_i = \beta_0 + \beta_1 X_i + \varepsilon_i; \quad i = 1, 2, \cdots, n. \tag{2.18}$$

em que, X_i é a variável quantitativa e $\varepsilon_i \sim N(0, \sigma^2)$.

Segundo Valle e Rabelo (2002), para a formulação de uma equação de regressão com paramêtros que variam, precisa-se da tomada de posição acerca da forma como se processam as variações. Tem-se que levar em consideração que, embora X constitua uma variável importante para determinar o comportamento da variável dependente (Y), ainda existe uma parte do comportamento desta variável que não é explicada pelo modelo. As-

sim nos leva a concluir que o modelo descrito acima encontra-se mal especificado por incorreta omissão de variáveis explicativas.

Para melhor entendimento da técnica, apresenta-se um exemplo com observações de um estudo hipotético, em que, num primeiro momento, existem três grupos, em que, cada grupo tem uma "qualidade" com ausência ou presença do atributo. A relação entre a variável dependente e a variável explicativa é estável para todas as observações da amostra. Assim a Figura abaixo expressa a relação entre a variável regressora (X) e a varíavel dependente (Y), então

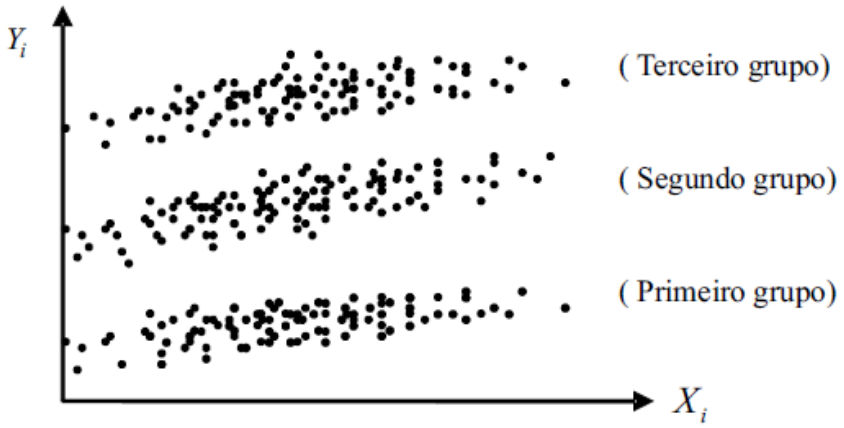

Figura 3: Gráfico de dispersão com obeservações hipotéticas
Fonte: Valle e Rabelo (2002)

Pode-se obeservar na Figura (3), que existe uma relação linear positiva entre as variáveis regressora (X) e dependente (Y). Parece existir uma relação distinta entre as duas variáveis para as observações que pertencem a cada um dos grupos. Então, uma forma de solucionar esse problema seria separar o modelo descrito anteriormente em grupos de observações, onde, cada grupo tem seu modelo de regressão, tornando-se um modelo de regressão em três novos modelos distintos.

As retas de regressão podem ter inclinações semelhantes, no entanto, a Figura 3 mostra que os interceptos β_0's diferem nos três grupos, e a inclinação β_1's das retas parece não diferir. Então, os três modelos podem ser descritos desta forma

i) $y_i = \beta_{01} + \beta_1 X_i + \varepsilon_i$; primeiro grupo

ii) $y_i = \beta_{02} + \beta_1 X_i + \varepsilon_i$; segundo grupo

iii) $y_i = \beta_{03} + \beta_1 X_i + \varepsilon_i$; terceiro grupo

Vale lembrar que é quase impossível que os três modelos tenham o mesmo valor para a inclinação (β_1). Mas se for razoável assumir que os grupos têm uma certa similaridade para a variação em X, o melhor a se fazer será, juntar os três modelos em um modelo de regressão, com uma única estimativa para o coeficiente de inclinação e três termos independentes diferentes.

Então, assim define-se as variáveis *dummies*, para esse caso

$$D_{2i} = \begin{cases} 1, & \text{se a observação verifica a característica que define o segundo grupo;} \\ 0, & \text{caso contrário} \end{cases}$$

$$D_{3i} = \begin{cases} 1, & \text{se a observação verifica a característica que define o terceiro grupo;} \\ 0, & \text{caso contrário} \end{cases}$$

Daí, ajusta-se o novo modelo de regressão,

$$y_i = \beta_{0_1} + (\beta_{0_2} - \beta_{0_1})D_{2i} + (\beta_{0_3} - \beta_{0_1})D_{3i} + \beta_1 X_i + \varepsilon_i \tag{2.19}$$

em que, $i = 1, 2, \cdots, n \quad \varepsilon_i \sim N(0, \sigma^2)$, a mesma estimativa para β e interceptos diferentes.

Nota-se que, quando $D_{2i} = D_{3i} = 0$, na Equação (2.19) tem-se

$$y_i = \beta_{0_1} + \beta_1 X_i + \varepsilon_i; \quad \text{primeiro grupo}$$

quando $D_{2i} = 1$ e $D_{3i} = 0$, tem-se

$$y_i = \beta_{0_2} + \beta_1 X_i + \varepsilon_i; \quad \text{segundo grupo}$$

quando $D_{2i} = 0$ e $D_{3i} = 1$, tem-se

$$y_i = \beta_{0_3} + \beta_1 X_i + \varepsilon_i; \quad \text{terceiro grupo}$$

ou seja, a Equação (2.19) é uma forma alternativa de representar os modelos dos três grupos. Ao agrupar os três modelos num só modelo (2.19) não tem apenas o mesmo β_1, mas também o mesmo vetor de erros $\boldsymbol{\varepsilon_i}$, ou seja, tem a mesma distribuição para os três grupos.

Podendo-se reescrever o modelo de regressão (2.19), em que, o coeficiente da variável *dummy* D_2 representa a diferença entre os interceptos do primeiro e segundo grupo, da mesma forma o coeficiente da variável *dummy* D_3 que representa a diferença dos inter-

ceptos entre o primeiro e o terceiro grupo. O modelo descrito desta forma

$$y_i = \beta_{0_1} + \gamma_2 D_{2i} + \gamma_3 D_{3i} + \beta_1 X_i + \varepsilon_i \quad i = 1, 2, \cdots, n. \tag{2.20}$$

em que, $\varepsilon_i \sim N(0, \sigma^2)$ e $\gamma_2 = (\beta_{0_2} - \beta_{0_1})$ e $\gamma_3 = (\beta_{0_3} - \beta_{0_1})$ são as distâncias entre as retas no intercepto, assim para cada grupo tem-se

$$y_i = \beta_{0_1} + \beta_1 X_i + \varepsilon_i; \quad \text{quando } D_{2i} = D_{3i} = 0$$

$$y_i = (\beta_{0_1} + \gamma_2) + \beta_1 X_i + \varepsilon_i; \quad \text{quando } D_{2i} = 1 \text{ e } D_{3i} = 0$$

$$y_i = (\beta_{0_1} + \gamma_3) + \beta_1 X_i + \varepsilon_i; \quad \text{quando } D_{2i} = 0 \text{ e } D_{3i} = 1$$

De acordo com Valle e Rabelo (2002), as variáveis do intercepto devem ser aditivas, ou seja, o intercepto é somado ao efeito de cada fator qualitativo e o efeito de qualquer variáveil binária é independente de outro valor qualitativo. A estimativa do modelo (2.20), pode ser representado graficamente na forma

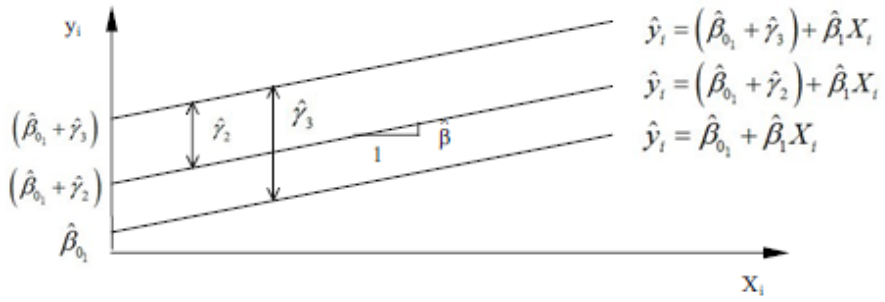

Figura 4: Gráfico da estrutura estimada do modelo 2.20
Fonte: Valle e Rabelo (2002)

Agora será realizado o estudo da regressão sobre a variável quantitativa e as variáveis qualitativas, utilizando-se o modelo (2.19) adimitindo-se que $[E(\varepsilon_i) = 0]$, tem-se

- O valor esperado para o primeiro grupo

$$E(Y_i | X_i, D_{1i} = D_{2i} = 0) = \beta_{01} + \beta_1 X_i,$$

- O valor esperado para o segundo grupo

$$E(Y_i | X_i, D_{1i} = 1, D_{2i} = 0) = (\beta_{02} - \beta_{01}) + \beta_1 X_i,$$

- O valor esperado para o terceiro grupo

$$E(Y_i|X_i, D_{1i} = 1, D_{2i} = 0) = (\beta_{03} - \beta_{01}) + \beta_1 X_i,$$

De acordo com Valle e Rabelo (2002) as retas de regressão podem assumir um valor igual no intercepto com o eixo Y, com coeficientes de inclinações diferentes. Dessa forma os problemas que foram enunciados nos modelos com inclinações semelhantes se aplica também nesse modelo. Assim os modelos de regressão para cada grupo que tenha as especificações acima, serão

i) $y_i = \beta_0 + \beta_{11} X_i + \varepsilon_i$; primeiro grupo

ii) $y_i = \beta_0 + \beta_{12} X_i + \varepsilon_i$; segundo grupo

iii) $y_i = \beta_0 + \beta_{12} X_i + \varepsilon_i$; terceiro grupo

Mas uma vez o indicado é ajustar um único modelo de regressão a partir dos modelos dos três grupos definidos anteriormente, onde haverá uma estimativa para o termo independente e três coeficientes de inclinação diferentes. O modelo será

$$y_i = \beta_0 + \beta_{11} X_i + (\beta_{12} - \beta_{11})(D_{2i} X_i) + (\beta_{13} - \beta_{11})(D_{3i} X_i) + \varepsilon_i, \tag{2.21}$$

em que, $i = 1, 2, \cdots, n \quad \varepsilon_i \sim N(0, \sigma^2)$. De forma equivalente, obtem-se o modelo de regressão

$$y_i = \beta_0 + \beta_{11} X_i + \delta_2 (D_{2i} X_i) + \delta_3 (D_{3i} X_i) + \varepsilon_i \quad i = 1, 2, \cdots, n, \tag{2.22}$$

em que, $\varepsilon_i \sim N(0, \sigma^2)$ e $\delta_2 = (\beta_{12} - \beta_{11})$ e $\delta_3 = (\beta_{13} - \beta_{11})$ são os coeficientes de inclinação dos dois últimos grupos, tomado pela diferença entre o segundo e terceiro grupo com o primeiro.

Desta forma é permitido a determinação dos modelos dos grupos separadamente, por meio de valores adequados para as variáveis *dummies*. Então os modelos de regressão para cada grupo será

Quando $D_{2i} = D_{3i} = 0$,

$$y_i = \beta_0 + \beta_{11} X_i + \varepsilon_i; \quad \text{primeiro grupo}$$

Quando $D_{2i} = 1$ e $D_{3i} = 0$,

$$\begin{aligned} y_i &= \beta_0 + \beta_{12} X_i + \varepsilon_i \\ y_i &= \beta_0 + (\beta_{11} + \delta_2) X_i + \varepsilon_i; \quad \text{segundo grupo} \end{aligned}$$

Quando $D_{2i} = 0$ e $D_{3i} = 1$,

$$\begin{aligned} y_i &= \beta_0 + \beta_{13} X_i + \varepsilon_i \\ y_i &= \beta_0 + (\beta_{11} + \delta_3) X_i + \varepsilon_i; \quad \text{terceiro grupo} \end{aligned}$$

A estimativa do modelo (2.21), pode ser representado graficamente desta forma

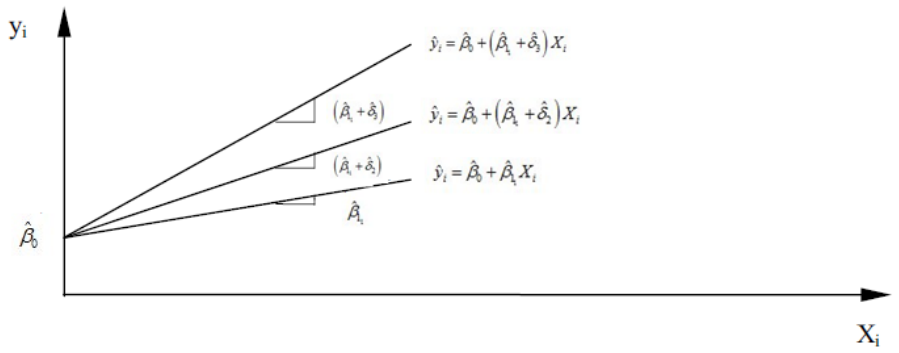

Figura 5: Gráfico da estrutura estimada do modelo (2.21)
Fonte: Valle e Rabelo (2002)

3 Aplicação

Encontra-se nesta seção as principais técnicas que serviram de base para este trabalho, tanto na parte da descrição dos modelos quanto nas inferências.

3.1 Um breve comentário sobre a diabetes

A diabetes é uma disfunção do metabolismo, ou seja, o jeito com que o organismo faz a digestão dos alimentos para a produção de energia. A maioria dos alimentos que se ingerem são quebradas em pequenas partículas de glicose, um tipo de açúcar encontado no sangue. Que após a digestão passa para a corrente sanguínea. No entanto, para que a glicose possa adentrar nas células, ela precisa da ajuda da insulina. A insulina é um hormônio produzido no pâncreas, uma glândula localizada por trás do estômago. Quando nos alimentamos, o pâncreas produz automaticamente a quantidade certa de insulina necessária para mover a glicose do sangue para as células do corpo. Em pessoas com diabetes, o pâncreas produz pouca insulina, então as células não respondem da forma esperada à insulina produzida. Daí, a glicose fica no sangue aumentando o que se chama de glicemia (concentração de glicose), ou vai direto para a urina (não sendo aproveitada pelas células).

3.2 Dados utilizados

Para o desenvolvimento das técnicas expostas no capítulo anterior foi utilizado o banco de dados da biblioteca do R *Faraway* (*Diabetes survey on Pima Indians*), em que, o Instituto Nacional de Diabetes e Doenças Digestivas e Renais realizou um estudo em 768 índias Pimas adultas que vivem perto de *Phoenix - Arizona*. Os dados foram coletados segundo critérios da Organização Mundial de Saúde, em que, as variáveis utilizadas foram as seguintes:

i) **Partos**: Número de vezes que as índias ficaram grávidas

ii) **Glicose**: A concentração plasmática de glicose a 2 horas em um teste oral de tolerância à glicose

iii) **Diastólica**: A pressão arterial diastólica (mm Hg)

iv) **Triceps**: Espessura cutânea triciptal (mm)

v) **Insulina**: 2 Horas de insulina no soro (mu U / ml)

vi) **IMC**: Índice de massa corporal (peso em kg / (altura em metros ao quadrado))

vii) **Diabetes**: Diabetes função da genealogia

viii) **Idade**: Idade em anos

ix) **Teste**: Teste de sinais de diabetes nos pacientes (categorizada 0 se negativo, 1 se positivo)

Os dados são descritos na Tabela 2,

Tabela 2: Dados da biblioteca do R (*Diabetes survey on Pima Indians*)

Obs.	Partos	Glicose	Diastólica	Triceps	Insulina	IMC	Diabetes	Idade	Teste
1	6	148	72	35	0	33,60	0,63	50	1
2	1	85	66	29	0	26,60	0,35	31	0
3	8	183	64	0	0	23,30	0,67	32	1
4	1	89	66	23	94	28,10	0,17	21	0
5	0	137	40	35	168	43,10	2,29	33	1
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
768	1	93	70	31	0	30,40	0,32	23	0

3.2.1 Análise descritiva das variáveis

De início optou-se por fazer uma análise descritiva dos dados, na Figura 6 têm-se os histogramas para quem teve a variável teste assumindo-se (1 considerado com sinais de diabétes e 0 para quem não apresentava sinais de diabétes)

De forma que pela Figura (6a e 6b) as índias não diabéticas tiveram índices de partos, entre 1 e 5 partos. Em compensação índias diabéticas foram as que tiveram uma maior quantidade de partos, podendo ser uma grande influência na causa da diabetes entre elas.

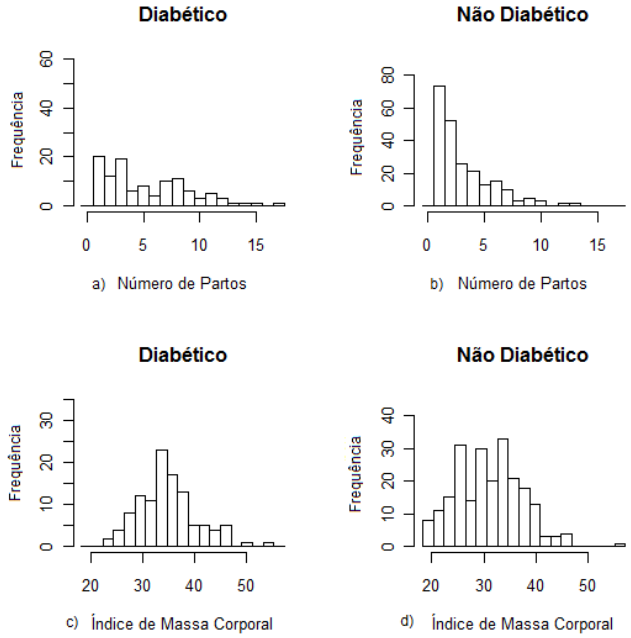

Figura 6: Análise descritiva das variáveis índice de massa corporal e número de partos

Em relação a Figura (6c e 6d) as índias diebéticas tiveram uma maior concentração em índice de massa corpórea entre 30 e 40 (kg), já as não diabéticas apresentaram uma maior dispersão em relação ao índice de massa corporal.

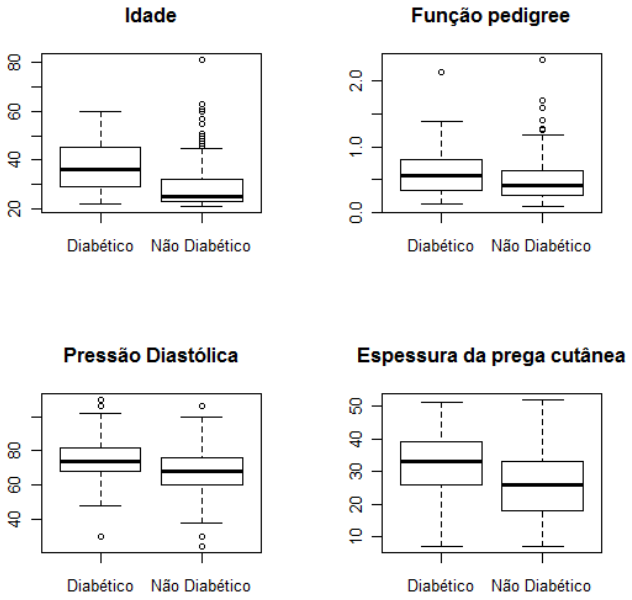

Figura 7: Análise descritiva das variáveis idade, função da genealogia, pressão diastólica e espessura da prega cutânea (triceps)

De acordo com a Figura 7, tem-se que,

i) **Idade:** As índias diabéticas têm as idade variando-se de 21 a 60 anos, já nas não diabéticas pôde-se observar alguns pontos discrepantes e uma boa parte das índias não diabéticas têm idade de 20 a 45 anos, e algumas chegam a 81 anos podendo considerar como sendo possíveis *outliers*.

ii) **Função de genealogia:** De forma geral, percebe-se que os traços de diabétes se manifestam a partir dos 20 anos com forte concentração na idade fértil da índias. Manifestação de traços de diabétes em função da genealogia nas índias não é perceptível, pois no boxplot para as índias que não apresentaram traços de diabétes o mesmo foi notado.

ii) **Pressão diastólica:** As índias diabéticas e não diabéticas têm praticamente a mesma variabilidade em relação a pressão diastólica

iv) **Espessura da prega cutânea:** Nota-se que índias diabéticas têm a espessura da prega cutânea variando de 25 a 40 (mm), já em relação às não diabéticas estão em sua grande maioria variando de 18 a 30 (mm).

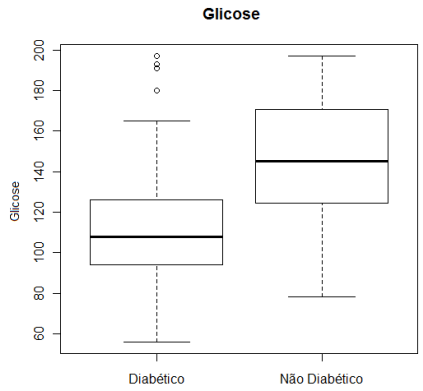

Figura 8: Análise descritiva da variável glicose

Por meio da Figura 8, nota-se que índias não diabéticas têm uma maior concentração plasmática de glicose que as índias diabéticas, tendo em vista que algumas (poucas) índias diabéticas têm uma alta concentração plasmática de glicose.

Após a descrição do banco de dados faz-se uma caracterização gráfica por meio da Figura 9, em que utilizou-se a variável glicose como base e comparou-se à dispersão da mesma com as demais. Esta análise serviu de base para a sugestão das variáveis que

pertenceriam ao modelo, por exemplo, a relação entre glicose e insulina, sugeriu-se que a relação era de um polinômio de segundo grau.

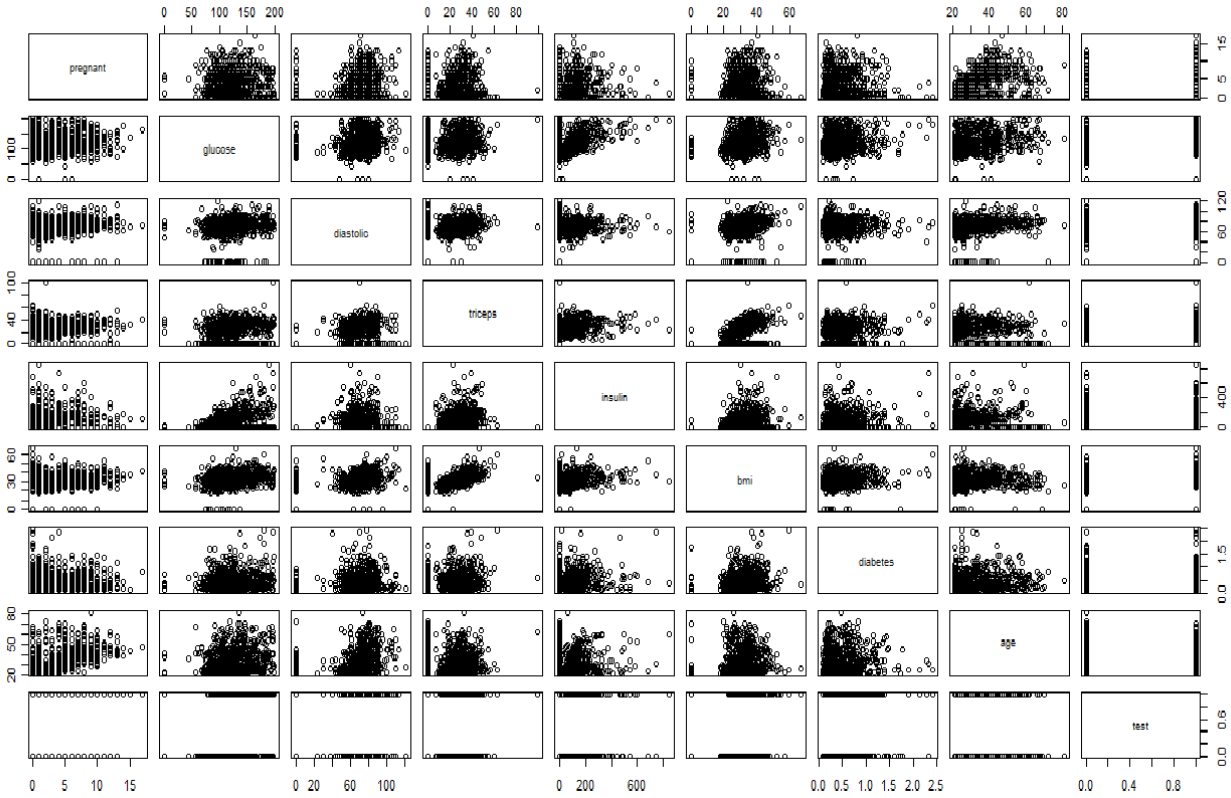

Figura 9: Scatter Plot para as variáveis glicose, partos, diastólica, teste, triceps, bmi, insulina, insulina^2, idade, idade^2, diabete e diabete^2

Propõe-se um modelo de regressão linear múltipla com onze variáveis regressoras envolvidas:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 D_1 + \beta_3 x_2 + \beta_4 x_3 + \beta_5 x_4 + \beta_6 x_5 + \beta_7 x_5^2 + \beta_8 x_6 + \beta_9 x_6^2 + \beta_{10} x_7 + \beta_{11} x_7^2 + \varepsilon \quad (3.1)$$

em que, foi considerado glicose como variável resposta (Y) e partos, diastólica, teste, triceps, bmi, insulina, insulina^2, idade, idade^2, diabete e diabete^2 como variáveis regressoras, respectivamente representados por: $x_1, D_1, x_2, x_3, x_4, x_5, x_5^2, x_6, x_6^2, x_7, x_7^2$ e seus respectivos coeficientes.

Utilizando-se a função *lm* da biblioteca *MASS* do software R 2.15 (*R DEVOLOPMENT CORE TEAM*, 2012). Sejam as hipóteses:

$$\begin{cases} H_0: & \beta_j = 0 \\ H_1: & \beta_j \neq 0 \quad \text{para pelo menos um,} \quad j = 1, 2, \cdots, k \end{cases}$$

tem-se que a análise de variância para a decomposição das somas de quadrados é definido

por

Tabela 3: Análise de variância para o modelo ajustado em (3.2)

Causas de Variação	GL	SQ	QM	F_{cal}	Pr(>F)
Partos	1	60,37	60,37	0,13	0,7221
Diastólica	1	1188,78	1188,78	2,49	0,1152
Teste	1	16931,20	16931,20	35,53	0,0000
Triceps	1	88,03	88,03	0,18	0,6676
IMC	1	974,06	974,06	2,04	0,1538
Insulina	1	30424,95	30424,95	63,84	0,0000
Insulina^2	1	8669,57	8669,57	18,19	0,0000
Idade	1	485,25	485,25	1,02	0,3137
Idade^2	1	115,32	115,32	0,24	0,6231
Diabetes	1	781,16	781,16	1,64	0,2014
Diabetes^2	1	602,14	602,14	1,26	0,2618
Resíduos	324	154407,13	476,56	-	-
Total	335	214727,96	60797,39	-	-

Analisando-se as variáveis teste, insulina e insulina^2, nota-se que o valor P $= P[F_{1,324} > F_0] < 0,001$, sendo assim há fortes indícios para rejeitar a hipótese H_0 ao nível de 5% de probabilidade para algumas variáveis, ou seja, há evidência de que os coeficientes das variáveis, teste, insulina e insulina^2 são estatísticamente diferentes de zero.

Tabela 4: Estimativas dos parâmetros com respectivos erros padrão e estatística t para as variáveis partos, diastólica, teste, triceps, bmi, insulina, insulina^2, idade, idade^2, diabete e diabete^2

Efeitos	Estimativas	Erro Padrão	Valor de t	Pr(>|t|)
Intercepto	71,5064	14,6452	4,88	0,0000
Partos	-0,1903	0,5348	-0,36	0,7221
Diastólica	0,1697	0,1074	1,58	0,1152
Teste	18,2249	3,0576	5,96	0,0000
Triceps	0,0666	0,1549	0,43	0,6676
IMC	-0,3678	0,2573	-1,43	0,1538
Insulina	0,2362	0,0296	7,99	0,0000
Insulina^2	-0,0002	0,0000	-4,27	0,0000
Idade	0,7466	0,7398	1,01	0,3137
Idade^2	-0,0044	0,0090	-0,49	0,6231
Diabétes	-13,0969	10,2296	-1,28	0,2014
Diabétes^2	6,9238	6,1597	1,12	0,2618

Os erros padrão em geral foram baixos, sendo altos apenas para as variáveis que não foram significativas. Devido ao fato do teste T^2 de *Hotelling* ser aquivalente ao teste F, as significâncias da Tabela 3 e Tabela 4 foram iguais.

Desta forma obtem-se o seguinte modelo ajustado:

$$\begin{aligned}\hat{y} &= 71,51 - 0,1903x_1 + 0,1697x_2 + 18,22D_1 + 0,0665x_3 - 0,3678x_4 + \\ &+ 0,2362x_5 - 0,0002x_5^2 + 0,7466x_6 - 0,0044x_6^2 - 13,10x_7 + 6,924x_7^2\end{aligned} \quad (3.2)$$

Pelo método de seleção de variável regressora passo a passo (*stepwise*) sobre o critério de informação de Akaike (AIC) e o critério de informação de Bayesiano (BIC), pode-se selecionar quais variáveis realmente vão continuar no modelo. Para tanto, ajustou-se inicialmente o modelo reduzido, aquele apenas com o intercepto e vão sendo incluidas as variáveis regressoras, até que o menor valor de AIC fosse obtido, deste modo partiu-se de um $AIC = 2303,95$ e obteve-se ao final um $AIC = 2076,1$ com o modelo 3.3:

$$\hat{y} = 73,09 + 0,2309x_1 + 18,03D_1 - 0,0001x_2^2 + 0,3583x_3 + 0,1853x_4 - 0,3028x_5 \quad (3.3)$$

A análise de variância com o modelo 3.3, é definida na Tabela 5:

Tabela 5: Análise de Variância após a seleção de variáveis pelo critério AIC

Causas de Variação	GL	SQ	QM	F_{cal}	Pr(>F)
Insulina	1	109522,78	109522,78	231,74	0,0000
Teste	1	36574,40	36574,40	77,39	0,0000
Insulina^2	1	8362,36	8362,36	17,69	0,0000
Idade	1	5470,08	5470,08	11,57	0,0008
Diastiólica	1	1022,08	1022,08	2,16	0,1424
IMC	1	1037,75	1037,75	2,20	0,1393
Resíduo	329	155488,26	472,61	-	-
Total	335	317477,71	162462,06	-	-

Como o valor P foi maior que o nível de significância nas variáveis insulina, teste, insulina^2 e idade, então rejeita-se a hipótese H_0 ao nível de 5% de probabilidade, ou seja, há indícios de que os coeficientes das variáveis são estatísticamente diferentes de zero.

Porém o método de seleção baseado no AIC, selecionou variáveis que individualmente não apresentavam significância estatística, em função disto, realizou-se o critério de BIC, e por meio dele, também podemos selecionar quais variáveis regressora vão continuar no modelo, e se vão ser as mesmas variáveis selecionadas pelo critério AIC, com o intuito de comparar os critérios na busca de selecionar o modelo que melhor se ajusta aos dados. No BIC, ajustou-se o modelo reduzido como no AIC, em que, encontrou-se um $BIC = 2307,77$ e chegou ao $BIC = 2095,61$. Tendo em vista que a tabela de análise de variância

é dado por:

Tabela 6: Análise de variância após a seleção de variáveis pelo critério BIC, em que as variáveis selecionadas foram insulina, teste, insulina^2 e idade

Causas de Variação	GL	SQ	QM	F_{cal}	Pr(>F)
Insulina	1	109522,78	109522,78	230,10	0,0000
Teste	1	36574,40	36574,40	76,84	0,0000
Insulina^2	1	8362,36	8362,36	17,57	0,0000
Idade	1	5470,08	5470,08	11,49	0,0008
Resíduo	331	157548,08	475,98	-	-
Total	335	317477,70	160405,6	-	-

em que, todas as variáveis que foram selecionadas pelo critério BIC foram significativas, ou seja, ao nível de 5% de probabilidade os coeficientes das variáveis, insulina, teste, insulina^2 e idade são estatísticamente diferentes de zero. Percebe-se que apesar da perda de graus de liberdade nos resíduos do modelo selecionado pelo AIC, não houve significativa redução na soma de quadrados de resíduos, se comparada a soma de quadrados do modelo 3.4, em que, utilizou-se o BIC.

Tabela 7: Estimativas dos parâmetros com respectivos erros padrão e estatística t para as insulina, teste, insulina^2 e idade

Efeitos	Estimativas	Erro Padrão	Valor de t	Pr(>|t|)
Intercepto	74,9843	4,5134	16,61	0,0000
Insulina	0,2260	0,0285	7,94	0,0000
Teste	17,8012	2,9111	6,11	0,0000
Insulina^2	-0,0002	0,0000	-4,15	0,0000
Idade	0,4234	0,1249	3,39	0,0008

Na Tabela 7 encontram-se as estimativas dos parâmetros, a partir daí obteve-se o modelo ajustado 3.4.

$$\hat{y} = 74,98 + 0,2259x_1 + 17,80D_1 - 0,0001x_2^2 + 0,4234x_3 \tag{3.4}$$

Nota-se que o modelo selecionado pelo critério BIC, foi aquele cujo as variáveis foram as mesmas do critério AIC a menos das variáveis que não foram significativas na seleção do AIC, pois o critério BIC leva em consideração o tamanho da amostra. De posse do modelo ajustado em (3.4), verifica-se as suposições para validação do modelo, por meio de uma análise gráfica dos resíduos em que, podemos observar se os resultados acima são confiáveis ou não:

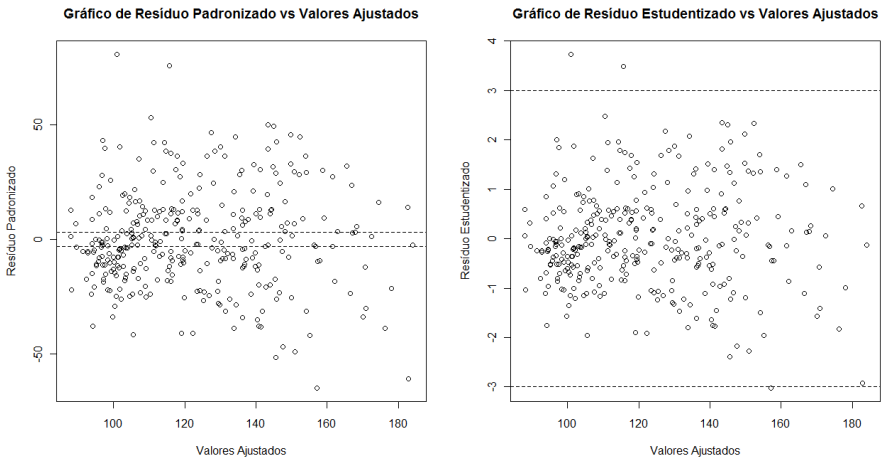

Figura 10: Análise gráfica de resíduos

Na Figura 10 observa-se que não há tendência nos resíduos e os pontos estão aleatoriamente distribuídos em torno do 0, levando-se a supor que a variância dos resíduos é homoscedástica.

O teste de Goldfeld-Quandt foi utilizado a fim de constatar se realmente a variância dos resíduos é homocedástica, para o mesmo encontrou-se $GQ = 0,9027$, o valor $P = 0,743 > \alpha = 0,05$, então há indicios para aceitar a hipótese de que as variâncias dos resíduos é homocedásticas ao nível de 5% de probabilidade.

Utiliza-se o gráfico de probabilidade normal envelopado para uma análise visual da normalidade dos resíduos.

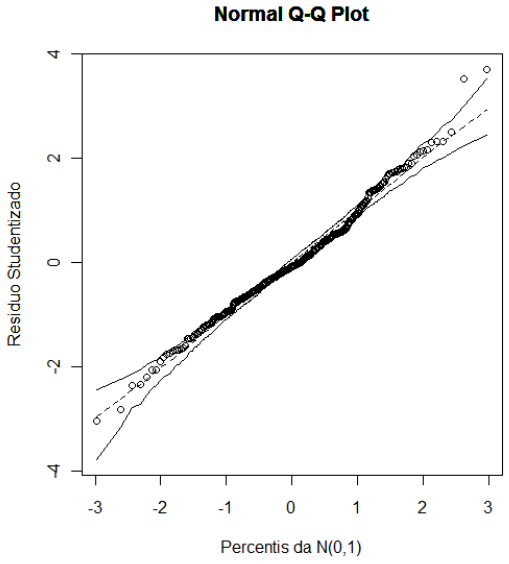

Figura 11: Gráfico de probabilidade normal envelopado

Como existem muitos dados fora do envelope, foi necessário realizar um teste que seja preciso para identificação da normalidade dos resíduos. Para tanto, utilizou-se o teste Shapiro-Wilk e obteve-se $W = 0,9881$ com valor $P = 0,0074 < \alpha = 0,05$, então conclui-se ao nível de 5% de probabilidade que não há evidências para dizer que a distribuição dos resíduos seja normal.

Como não foi possível satisfazer a uma das suposições do modelo ajustado, então é necessário buscar alternativas para que se possa modelar e tirar inferências de forma confiável, uma delas é a transformação dos dados. Optou-se por transformar a variável resposta para que houvesse a possibilidade de encontrar homogeneidade nas variâncias e uma distribuição normal dos resíduos. Para tanto utilizou-se a transformação Box-Cox da biblioteca car do R. O gráfico de perfil de verossimilhança encontra-se na Figura 12:

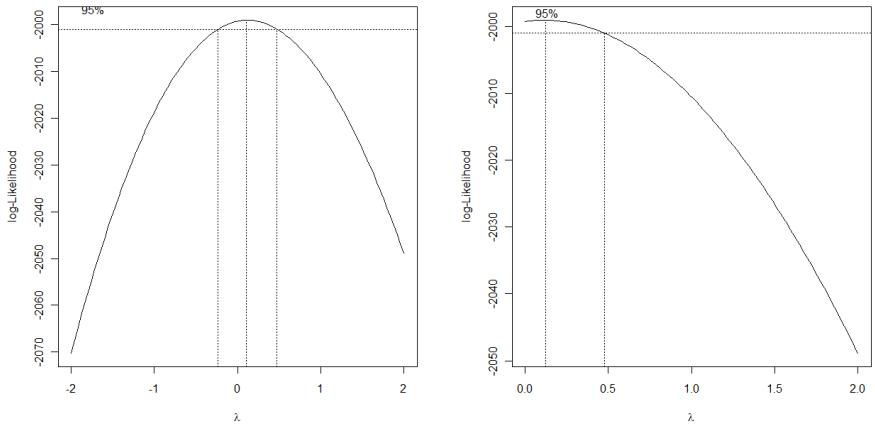

Figura 12: Gráfico do perfil de verossimilhança para o modelo de transformação de Box-Cox

Como pode ser visto na Figura 12, o valor que maximiza o logaritmo da função de verossimilhança foi aproximadamente $\lambda = 0,01 \approx 0$, e no intervalo de confiança para λ o valor 0 está contido, ou seja, pode-se utilizar a transformação logaritimica na variável dependente, desta forma a transformação realizada será

$$y(\lambda) = log(y)$$

Após a variável resposta ser transformada, devem-se realizar todas as técnicas anteriormente executadas. Tendo em vista que o modelo inicialmente ajustado foi o (3.2), utilizou-se novamente o método de seleção de variáveis *stepwise* sobre o critério (AIC) e o critério (BIC) após a variável resposta ser transformada e encontrou-se um $AIC = -928,85$ quando o modelo era apenas com o intercepto. Quando realizado todo procedi-

mento de seleção encontrou-se um $AIC = -1155,69$ e o seguinte modelo:

$$\hat{y}(\lambda) = 4,3530 + 0,0021x_1 + 0,1336D_1 - 0,000001x_2^2 + 0,0028x_3 + 0,0016x_4 - 0,0024x_5 \quad (3.5)$$

A Tabela de análise de variância para o modelo 3.5:

Tabela 8: Análise de variância para o modelo após transformação da variável resposta e seleção de variáveis, pelo critério de AIC

Causas de Variação	GL	SQ	QM	F_{cal}	$Pr(> F)$
Insulina	1	7,05	7,05	224,49	0,0000
Teste	1	2,24	2,24	71,31	0,0000
insulina^2	1	0,90	0,90	28,55	0,0000
Idade	1	0,37	0,37	11,62	0,0007
Diastolica	1	0,08	0,08	2,53	0,1130
IMC	1	0,07	0,07	2,25	0,1348
Resíduo	329	10,34	0,03	-	-
Total	335	21,05	10,74	-	-

Pelo que pode ser observado na Tabela 8 o valor P das variáveis insulina, teste, I(insulina^2) e idade, foi menor que $\alpha = 0,05$, rejeita-se a hipótese H_0 ao nível de 5% de probabilidade, ou seja, há indícios de que os coeficientes dessas variáveis são estatísticamente diferentes de zero. O valor estimado de (σ^2) foi de 0,03 é bem próximo de zero o que favorece o uso da transformação conforme pode ser visto em (CHARNET et al. , 2008)

Da mesma forma que aconteceu anteriormente o critério de informação Bayesiano (BIC), selecionou apenas as variáveis que eram significativas no modelo selecionado pelo critério AIC. Assim, ajustou-se o modelo reduzido, em que o $BIC = -925,04$, em que, a Tabela 9 é obtida sobre as variáveis selecionadas pelo critério BIC:

Tabela 9: Análise de variância para o modelo após transformação da variável resposta e seleção de variáveis, pelo critério BIC

Causas de Variação	GL	SQ	QM	F_{cal}	$Pr(> F)$
Insulina	1	7,05	7,05	222,62	0,0000
Teste	1	2,24	2,24	70,71	0,0000
Insulina^2	1	0,90	0,90	28,32	0,0000
Idade	1	037	0,37	11,53	0,0008
Resíduo	331	10,49	0,03	-	-
Total	335	21,05	10,59	-	-

Tabela 10: Estimativas dos parâmetros com respectivos erros padrão e estatística t

Efeitos	Estimativas	Erro Padrão	Valor de t	Pr(>|t|)
Intercepto	4,3731	0,0368	118,7500	0,0000
Insulina	0,0021	0,0002	8,9400	0,0000
Teste	0,1320	0,0238	5,5600	0,0000
Insulina^2	-0,0000	0,0000	-5,2800	0,0000
Idade	0,0035	0,0010	3,4000	0,0008

Desta forma, o modelo ajustado segundo o critério BIC, é representado por

$$\hat{y}(\lambda) = 4,3730 + 0,0020x_1 + 0,1320D_1 - 0,000001x_2^2 + 0,0034x_3 \tag{3.6}$$

Todas as variáveis que foram selecionadas para o modelo pelo critério BIC são significativas, pois o valor P foi menor que $\alpha = 0,05$, rejeita-se a hipótese H_0 ao nível de 5% de probabilidade, levando-se a concluir que os coeficientes das variáveis são estatísticamente diferentes de zero. Após o ajuste do modelo (3.6), verificou-se as suposições para validação do mesmo. Pode-se então observar por meio da Figura 13 os efeitos de cada variável selecionada pelo critério BIC:

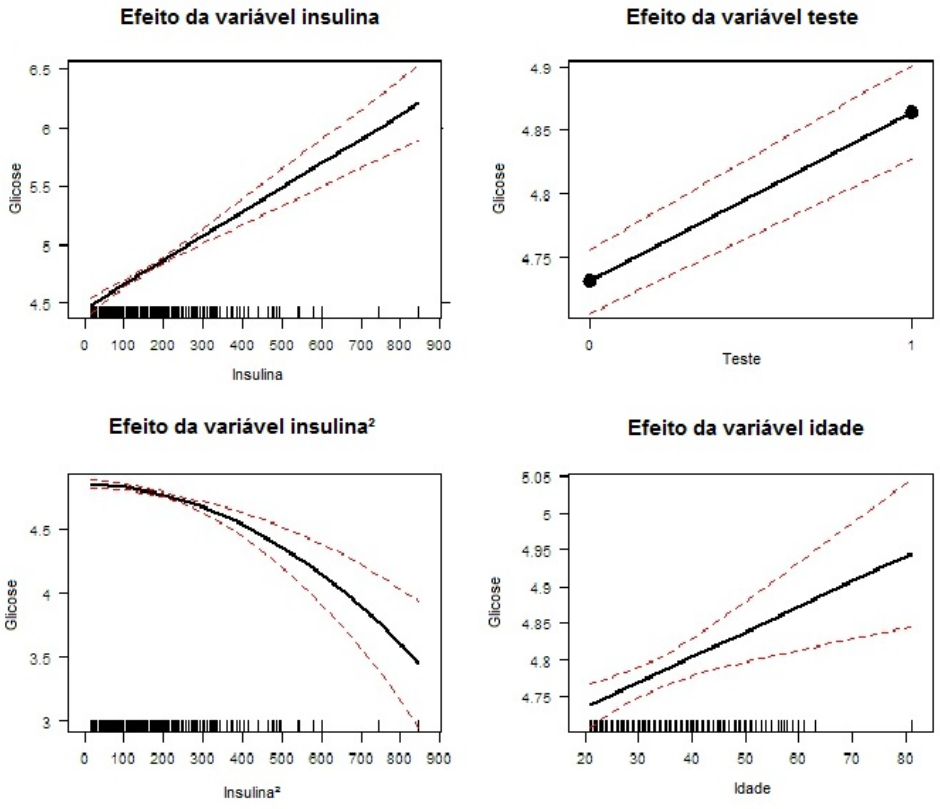

Figura 13: Gráfico de efeitos de cada variável selecionada pelo critério BIC

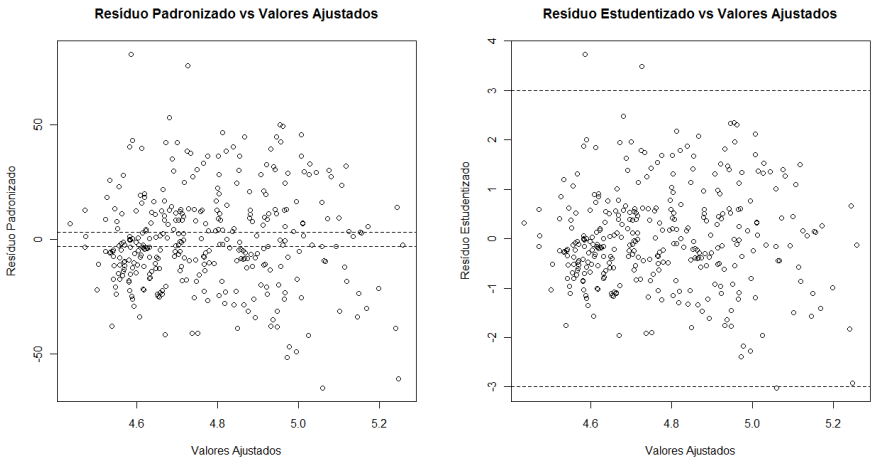

Figura 14: Gráfico de Resíduo *vs* Valores Ajustados

Na Figura 14 pode-se observar que os pontos estão aleatoriamente distribuídos em torno do 0, levando-nos a supor que a variância dos resíduos é homoscedástica. Suspeita-se novamente da presença de dois valores discrepantes que estão fora do intervalo $(-3, 3)$. Então, para que tenha-se a certeza de que os resíduos são homocedásticos, foi necessário realizar o teste de Goldfeld-Quandt em que $GQ = 1,001$, o $valor\ P = 0,4975 > \alpha = 0,05$, então não há indicios contra a hipótese de que as variâncias dos resíduos é homocedásticas ao nível de 5% de probabilidade.

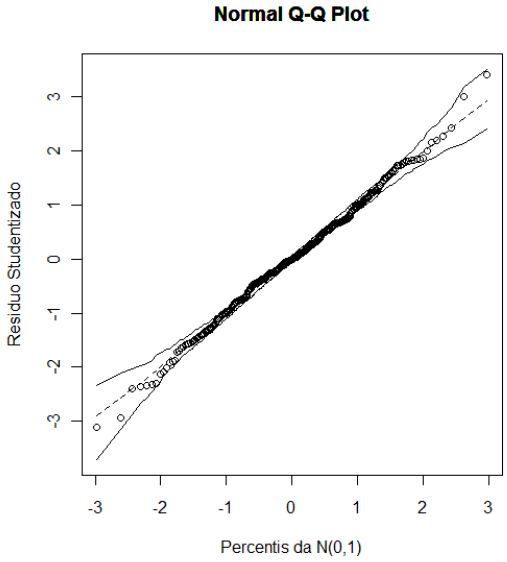

Figura 15: Gráfico de probabilidade normal envelopado

Após os dados transformados, tem-se que os resíduos são normais. Pois usando-se o gráfico de quantil normal envelopado, nota-se maior acomodação dos pontos entre as bandas de confiança, leva-nos a supor que não apresenta evidência contra a suposição

de normalidade nos resíduos. Por meio do teste de Shapiro-Wilk, pode-se verificar a existência de normalidade nos resíduos. Pois o valor da estatítica foi $W = 0.9965$, com valor $P = 0,6638 > \alpha = 0,05$, então concluiu-se ao nível de 5% de probabilidade que não há evidências para dizer que a distribuição dos resíduos não seja normal.

Após a escolha e validação do modelo, partiu-se para interpretação dos dados e na Figura 16 tem-se a dispersão dos dados após transformação, em que, tem-se as covariáveis no eixo do x e a variável glicose logaritmada no eixo do y, as cores indicam a diferença entre os níveis da variável teste.

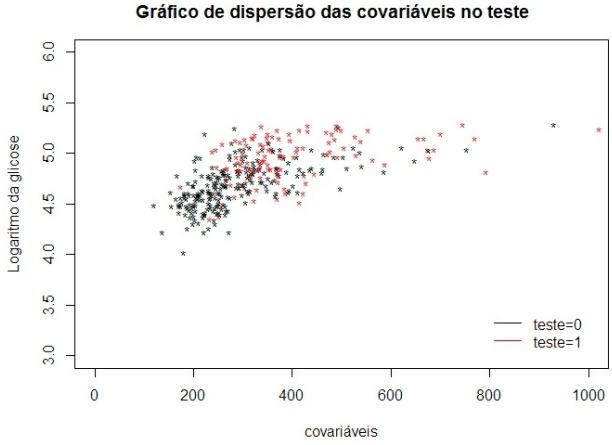

Figura 16: Gráfico de dispersão do logaritmo da glicose e as covariáveis

Percebe-se pela Figura 13 que é possível verificar dois comportamentos, uma para o teste = 0 quando índias são diagnósticadas como não diabéticas e a outra reta para o teste = 1 quando índias são diagnósticadas diabéticas. O que foi verificado no modelo ajustado 3.6, em que, a variável *dummy* foi significativa.

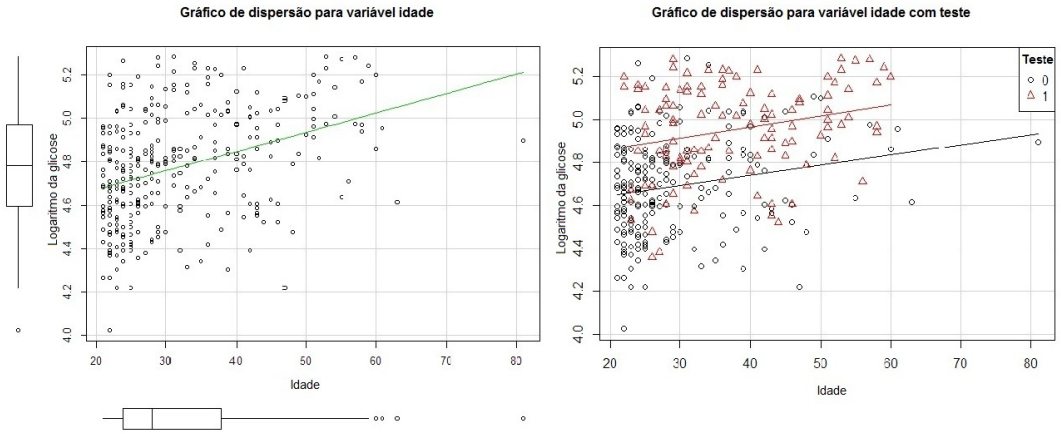

Figura 17: Gráfico de efeito da variável idade

Na Figura (17 a e b), têm-se os efeitos da variável idade no ajuste sem a variável *dummy* Figura (17 a) e com a variável *dummy* Figura (17 b), em que, percebe-se que não houve efeito de interação entre idade com o teste, ou seja, as retas de regressão são paralelas. Então os dois modelos parciais são descritos da seguinte forma na escala logaritimada:

Quando $D_1 = 0$, ou seja, quando o teste é negativo (não diabética)

$$\hat{y}_(\lambda) = 4,3730 + 0,0034x_4,$$

Quando $D_1 = 1$, quando o teste é positivo (diabética)

$$\hat{y}_(\lambda) = 4,5050 + 0,0034x_4.$$

ou seja, a medida que se passa um ano a concentração plasmática de glicose aumenta

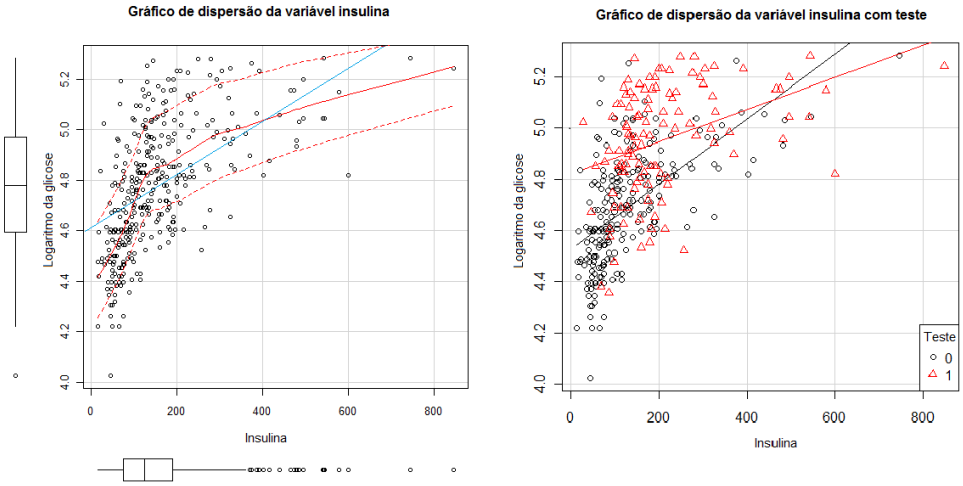

Figura 18: Gráfico de efeito da variável insulina

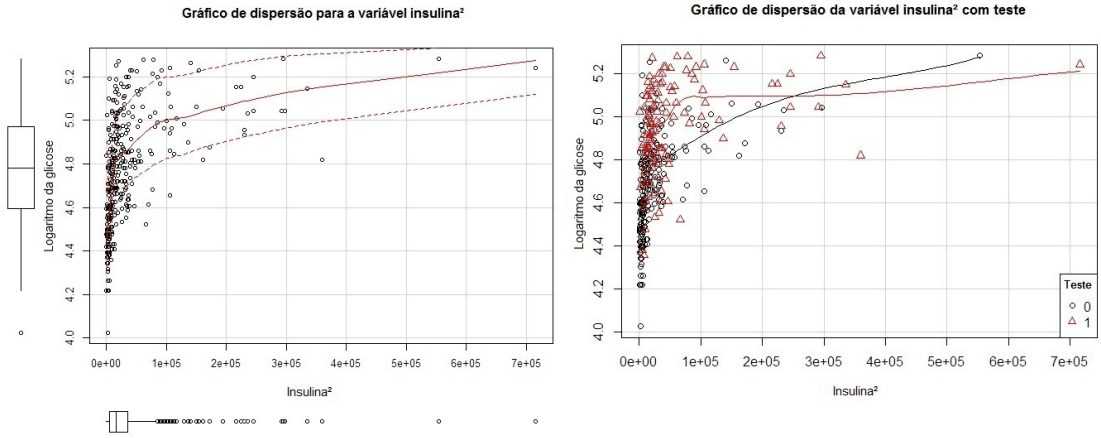

Figura 19: Gráfico de efeito da variável insulina^2

Já na Figura (18 a e b) e a Figura (19 a e b), percebe-se que insulina e teste apresentam efeito interativo, porém não se chegou a testar este efeito no modelo. Este resultado deve ser visto com certo cuidado pois não se verificou também a colinearidade entre as variáveis do modelo, pois pode causar impactos na estimativa dos parâmetros utilizando-se o VIF (*Variance Inflation Factor*).

Os dois modelos parciais para a variável insulina com teste são representados por:

Quando $D_1 = 0$, o teste é negativo (não diabética)

$$\hat{y}_(\lambda) = 4,3730 + 0,0020x_4,$$

Quando $D_1 = 1$, o teste é positivo (diabética)

$$\hat{y}_(\lambda) = 4,5050 + 0,0020x_4.$$

Em que, a cada duas horas de insulina no soro a concentração plasmática de glicose aumenta.

Já para a variável insulina^2, tem-se os dois modelos parciais da seguinte forma:

Quando $D_1 = 0$, o teste é negativo

$$\hat{y}_(\lambda) = 4,3730 - 0,000001x_4,$$

Quando $D_1 = 1$, o teste é positivo

$$\hat{y}_(\lambda) = 4,5050 - 0,000001x_4.$$

De modo que, a cada duas horas de insulina no soro a concentração plasmática de glicose diminui.

Do modelo proposto ajustado na escala logarítmica, faz-se o uso do estimador proposto por Miller (MILLER, 1984), e as estimativas dos valores esperados de glicose, na escala original foram obtidas a partir da Equação (3.7)

$$E[Y|x] = exp(4,3730 + 0,0020x_1 + 0,1320x_2 - 0,000001x_3 + 0,0034x_4)exp\left(\frac{0,03}{2}\right) \quad (3.7)$$

em que, os valores da variável glicose e suas estimativas dos valores esperados de glicose são apresentadas na Tabela 11:

Tabela 11: Valores das variáveis glicose e suas estimativas dos valores esperados de glicose

Glicose	E[Y|x]
89	103,3958
78	118,7067
197	242,5847
189	297,9376
166	150,3213
103	105,5633
⋮	⋮
121	110,10737

A variável *dummy* teste, utilizando-se o estimador de Miller propiciou $\hat{\beta}_0 = 80,23$ na ausência de diabétes e de 91,56 na presença de diabétes, indicando-se que a propecstão à diabetes implica em um acréscimo de 11,23 na concentração plasmática de glicose a 2 horas em um teste oral.

4 Conclusão

Neste trabalho ajustou-se um modelo de regressão multipla com uma variável *dummy* ao conjunto de dados do Instituto Nacional de Diabetes e Doenças Digestivas e Renais em 768 índias Pimas adultas que vivem perto de *Phoenix - Arizona*. As variáveis regressoras consideradas neste estudo foram: partos diastólica, tríceps, insulina, IMC, função da genealogia, idade e teste que categorizada em 0 se negativo (não diabético) e em 1 se positivo (diabético), já a variável resposta foi concentração plasmática de glicose no sangue.

Após o ajuste com todas as variáveis utilizou-se o critério de seleção de variável regressora *stepwise* sobre o critério de informação Baysiano (BIC), em que, outro modelo foi ajustado após seleção. Em seguida verificou-se as suposições para a validação do modelo, observou-se a necessidade de transformação da variável glicose pela violação da suposição de normalidade nos resíduos. Utilizando-se o logaritmo para transformação, pelo fato de o 0 está incluído no intervalo para o melhor λ do gráfico do perfil de verossimilhança para o modelo de transformação de Box-Cox. Após transformação e realização de todos os passos anteriores ajustou-se o seguinte modelo na escala transformada, $\hat{y}(\lambda) = 4,3730 + 0,0020x_1 + 0,1320D_1 - 0,000001x_2^2 + 0,0034x_3$ em que, as variáveis regressoras selecionadas foram insulina, teste, insulina^2 e idade respectivamente. De posse do novo modelo ajustado verificou-se as suposições e constatou-se por meio dos testes que o modelo era confiável e então partiu-se para interpretação dos resultados, o modelo ajustado utilizando-se o estimador de Miller foi $E[Y|x] = exp(4,3730 + 0,0020x_1 + 0,1320x_2 - 0,000001x_3 + 0,0034x_4)exp\left(\frac{0,03}{2}\right)$. A propecstão a diabetes implica em um acréscimo de 11,23 na concentração plasmática de glicose a 2 horas em um teste oral.

Referências

BARROS, E.A.C.; SIMÕES, P.A.; ACHCAR, J.A.; MARTINEZ, E.Z.; SHIMANO, A.C.; Métodos de estimação em regressão linear múltipla: aplicação a dados clínicos. *Revista Colombiana de Estadística*, Bogotá, Colombia, v.31, nº1, 111 - 129, 2008.

BOX, G.E.P.; COX, D.R.; An Analysis of Transformations. *Journal of the Royal Statistical Society*, London, v.26, nº2, 211 - 252, 1964.

CHARNET, R.; FREIRE, C.A.L.; CHARNET, E.M.R.; BONVINO, H. **ANÁLISE DE MODELOS DE REGRESSÃO LINEAR - Com aplicações**. 2º ed. Campinas, SP: UNICAMP, 2008. 368 p.

DRAPER, N.R.; SMITH, H.; **Applied regression analysis**. 3º ed. New York, New York: John Wiley & Sons, 1998. 706 p.

FERREIRA, D.F.; **Estatística Básica**. 2º ed. rev. Lavras, MG: UFLA, 2009. 664 p.

HOFFMAN, A. **ANÁLISE DE REGRESSÃO - Uma Introdução à Econometria**. 4º ed. São Paulo, SP: Hucitec, 2006. 378 p.

MILLER, D.M.; Reducing Transformation Bias in Curve Fitting. *The American Statistician*, v.38, nº2, 124 - 126, 1984.

MISSIO, F.; JACOBI, L.F.; Variáveis dummy: especificações de modelos com parâmetros variáveis. *Ciência e Natura*, Universidade Federal de Santa Maria - Santa Maria, RS, v. 29, nº1, 111 - 135, 2007.

MONTGOMERY, D.C.; PECK, E.A.; VINING, G.G.; **Introduction to Linear Regression Analysis**. 3º ed. New York, New York: John Wiley & Sons, 2003. 641 p.

OLINDA, R.A.; RIBEIRO JÚNIOR, P.J.; MOLIN, J.P; **Uso de técnicas geoestatísticas para determinar a dependência espacial do índice de cone da adequação de malhas amostrais**. 3º ed. Escola Superior de Agricultura Luiz de Queiroz, Piracicaba, 2010. 36 p. (apostila)

QUEIROZ, M.P.F.; *Testes de Hipóteses em Regressão Beta Baseados em Verossimilhança Perfilada Ajustada e em Bootstrap*. (Dissertação mestrado), Universidade Federal de Pernambuco, PE, 2011.

TAYLOR, J.M.G; *The Retransformed Mean after a Fitted Power Transformatio*, JAVA, 81, 114 - 118, 1986.

RODRIGUES, S.A.; DINIZ, C.A.R.; MODELO DE REGRESSÃO HETEROS-CEDÁSTICO. *Revista de Matemática e Estatística*, São Paulo, SP, v.24, nº2, 133 - 146, 2006.

VALLE, P.O.; REBELO, E.; ANÁLISE DE VARIÂNCIA E ANÁLISE DE REGRESSÃO COM VARIÁVEIS DUMMY: MAIS SEMELHANÇAS DO QUE DIFERENÇAS. *Revista de Estatística*, Instituto Nacional de Estatística, Lisboa, Portugal, v.1, 1º Quadrimestre, 31 - 67, 2002.

VALLE, P.O.; REBELO, E.; DUALIDADES ENTRE ANÁLISE DE COVARIÂNCIA E ANÁLISE DE REGRESSÃO COM VARIÁVEIS DUMMY. *Revista de Estatística*, Instituto Nacional de Estatística, Lisboa, Portugal, v.2, 2º Quadrimestre, 3 - 22, 2002.

VALLE, P.O.; REBELO, E.; O USO DE REGRESSORES DUMMY NA ESPECIFICAÇÃO DE MODELOS COM PARÂMETROS VARIÁVEIS. *Revista de Estatística*, Instituto Nacional de Estatística, Lisboa, Portugal, v.3, 3º Quadrimestre, 7 - 26, 2002.