UNIVERSIDADE ESTADUAL DA PARAÍBA

CENTRO DE CIÊNCIAS E TECNOLOGIA

DEPARTAMENTO DE ESTATÍSTICA

Filipe de Sousa Silva

Teoria dos grafos e regressão linear múltipla na análise de sistemas biológicos

Campina Grande - PB

Dezembro de 2017

Filipe de Sousa Silva

Teoria dos grafos e regressão linear múltipla na análise de sistemas biológicos

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

Orientador: Gustavo Henrique Esteves

Campina Grande - PB

Dezembro de 2017

S586t Silva, Filipe de Sousa.
Teoria dos grafos e regressão linear múltipla na análise de sistemas biológicos [manuscrito] : / Filipe de Sousa Silva. - 2017.
34 p. : il. colorido.

Digitado.
Trabalho de Conclusão de Curso (Graduação em Estatística) - Universidade Estadual da Paraíba, Centro de Ciências e Tecnologia, 2017.
"Orientação : Prof. Dr. Gustavo Henrique Esteves, Departamento de Estatística - CCT."

1. Redes de relevância. 2. Teoria dos grafos. 3. Regressão Linear Múltipla.

21. ed. CDD 519.5

Filipe de Sousa Silva

Teoria dos grafos e regressão linear múltipla na análise de sistemas biológicos

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.

Trabalho aprovado em 07 de dezembro de 2017.

BANCA EXAMINADORA

Gustavo Henrique Esteves
Universidade Estadual da Paraíba

Tiago Almeida de Oliveira
Universidade Estadual da Paraíba

Sílvio Fernando Alves Xavier Júnior
Universidade Estadual da Paraíba

A toda minha família.

Agradecimentos

Ao fim deste ciclo na minha vida, guardo na lembrança experiências incríveis, aprendizados que levarei para futuros projetos. Nesta caminhada conheci pessoas que contribuíram indiscutivelmente para a realização deste sonho.

No início, concluir o curso de estatística parecia impossível. Ouvi várias vezes comentários do tipo: “Esse curso é dificílimo”, “ninguém consegue pagar essa cadeira...” e muito mais destes conselhos adversos. Desistir nesta altura do campeonato parecia ser o caminho mais fácil. Inclusive, foi o que muitos optaram a fazer.

Mas diante de todos esses obstáculos, resolvi provar pra mim mesmo que era possível. E não tinha segredo, me restou sentar e estudar muito. E depois de tudo que passei, posso dizer que valeu muito a pena...

Primeiramente quero agradecer ao merecedor de toda honra e toda glória, nosso senhor Jesus Cristo. Quero agradecer também a uma pessoa muito importante na minha vida, dizer que não teria conseguido nada sem o apoio incondicional da minha noiva Andreza, companheira de todas as horas, que me deu força, me incentivou e me fez enfrentar essas barreiras sempre de cabeça erguida.

E todo aprendizado adquirido não seria possível sem a ajuda e dedicação dos professores do Departamento de Estatística, em especial, quero agradecer imensamente ao meu orientador, professor Gustavo Esteves, por toda atenção prestada e pela disponibilidade. Queria agradecer também aos meus colegas e amigos pela cumplicidade e companheirismo. Amigos que a universidade acabou me proporcionando já na reta final do curso: Shirlley, Ângela, Fátima, Vitória e Carlos.

Não poderia esquecer da minha família, minha base de sustentação, meus pais e irmãos, por toda a paciência depositada ao longo desses anos.

Enfim, a todos que contribuíram de alguma forma nesta empreitada, meu muito obrigado! E que venha o próximo desafio...

"Ora, a fé é a certeza daquilo que esperamos e a prova das coisas que não vemos."

(Hebreus 11:1)

Resumo

Várias técnicas experimentais vem sendo usadas na Biologia Molecular atual com a capacidade de medir os níveis de expressão de milhares de genes simultaneamente. Possibilitando, assim, justificar ou levantar novas hipóteses biológicas acerca de mecanismos de transformação molecular entre os diversos tipos de tecidos estudados. Tais hipóteses podem ser representadas na forma de grupos de moléculas com perfis específicos de interação entre si, frequentemente conhecidos como redes de regulação gênica. Neste sentido, uma rede de relevância é uma técnica de engenharia reversa usada para construir tais redes de regulação a partir de dados de expressão, onde representam interações entre entidades químicas complexas (como proteínas, substratos ou metabólitos) que ocorrem no nível molecular das células. Diante disso, existe a necessidade de representar e compreender o comportamento dessas vias biológicas através da construção de modelos matemáticos e gráficos. Assim, a teoria dos grafos é imprescindível para a compreensão das redes de regulação gênicas, em que os resultados deste tipo de análise são apresentados como grafos, no qual vértices e arestas representam componentes biológicos e interações entre eles, respectivamente. Este trabalho objetiva avaliar a possibilidade de se estimar interações entre genes de redes de regulação através da técnica de análise de regressão linear utilizando dados biológicos de expressão gênica.

Palavras-chaves: Redes de Relevância. Teoria dos grafos. Análise de regressão linear.

Abstract

Several experimental techniques have been used in Molecular Biology nowadays potentially measuring the expression levels of thousands of genes simultaneously. Thus, it is possible to justify or identify new biological hypotheses about mechanisms of molecular transformation among the different tissue types studied. Such hypotheses can be represented in the form of groups of molecules with specific profiles of interaction between themselves, often known as gene regulatory networks. In this way, a relevance network is a reverse engineering technique used to construct such regulation networks from gene expression data, where they represent interactions between complex chemical entities (such as proteins, substrates or metabolites) that occur in molecular level of the cells. Then, there is a need to represent and understand the behavior of these biological pathways through the construction of mathematical and graphical models. In this sense, the graph theory is essential for understanding the gene regulatory networks, in which the results of this type of analysis are presented as graphs, where vertices and edges represent biological components and the interactions between themselves, respectively.This work aims to evaluate the possibility of estimating interactions between genes of regulation networks through the linear regression analysis technique using biological data of gene expression.

Key-words: Relevance Networks. Graph theory. Regression analysis.

Lista de ilustrações

Lista de tabelas

Sumário

1 Introdução

A tecnologia de *microarray*, ou microarranjos de DNA permite mensurar os níveis de expressão de milhares de genes simultaneamente, possibilitando comparações entre amostras de tecidos pelos perfis de expressão. O avanço desta técnica vem propiciando estudos para o entendimento de problemas biológicos, como por exemplo, a comparação de células cancerígenas em relação as células sadias em um determinado tecido. Antes disso, a Biologia atravessou a era genômica, que foi mais focada na construção de sequenciadores de moléculas de DNA, onde se tentava compreender o código genético de qualquer ser vivo. Entretanto, o conhecimento de todo o alfabeto genético não bastava para explicar a complexidade biológica característica de um ser vivo.

Partindo disso, surgiu a era proteômica, que tentou viabilizar a explicação da complexidade biológica dos diversos organismos vivos através de interações entre moléculas de DNA e vários metabólitos do microambiente celular, chegando a ser conhecida como era interatômica (VIDAL, 2005). Em seguida, a Biologia Molecular conseguiu obter dados de expressão e transcrição gênica em larga escala, e acabou sendo beneficiada pela Bioinformática que, por sua vez, avançou a partir da necessidade crescente de ferramentas matemáticas, estatísticas e computacionais para armazenar e processar dados (LIM; VENKATESH, 2000; LOPES; CRUZ, 2011).

1.1 Ambiente computacional

A complexidade de analisar dados de *microarray* é notória, onde existe a necessidade de um ambiente computacional que integre diversos métodos de análise já disponíveis na literatura. Neste sentido, o `maigesPack`, ambiente computacional implementado no *software* de programação estatística `R`, que foi proposto por Esteves (2007) em sua tese de doutorado, é capaz de avaliar dados de expressão gênica em larga escala utilizando métodos de análise para buscar genes diferencialmente expressos, construir agrupamentos e classificadores.

Segundo Esteves (2007), o processo de análise dos dados divide-se em três etapas. A primeira etapa consiste em elaborar o grafo teórico de análise, onde os métodos a serem empregados devem ser identificados de acordo com a modelagem do conjunto de dados. Na segunda etapa, os métodos computacionais implementados devem ser identificados para cada análise teórica elaborada na primeira. Por fim, os parâmetros associados com os métodos computacionais a serem utilizados devem ser ajustados e os resultados analisados e interpretados.

Um bom exemplo de programa que pode ser usado para se analisar dados de *microarray* é o ambiente de programação estatística `R`[1] (R Development Core Team, 2011) que possui diversos pacotes para se trabalhar com as diversas etapas do processo de análise de dados. Uma de suas vantagens é o seu código livre, o que facilita que pessoas de qualquer parte do mundo desenvolvam e implementem pacotes adicionais, como é o caso do *Bioconductor*[2] que analisa dados genômicos. Porém, por ser uma linguagem interpretada, a desvantagem do R é a sua lentidão para análises computacionalmente intensivas.

1.2 Normalização dos dados

Como já foi dito, a tecnologia de *microarray* é capaz de mensurar os níveis de expressão de milhares de genes simultaneamente. No entanto, a obtenção de dados através desta técnica requer todo um processo sistemático que pode ser acompanhado de variações que influenciam diretamente nos resultados obtidos. A técnica de *microarray* resume-se na fixação do material genético em lâminas, ou seja, na deposição de sequências de cDNA conhecidas (oligonucleotídios) em posições específicas de um substrato, geralmente, em lâminas de vidro, onde é hibridizado contra cDNAs marcados, e logo depois são "tingidos" com os corantes fluorescentes *Cy3* e *Cy5*. Em seguida, os fluorocromos são excitados, emitindo sinais luminosos que são captados por um *scanner*.

Os dados de *microarrays* são originalmente imagens que representam os níveis de expressão dos genes fixados no substrato, analisados por um *software* específico que gera uma tabela de dados numéricos contendo os valores de intensidade de cada um dos fragmentos de DNA de interesse (ESTEVES, 2007). Todo esse processo experimental gera efeitos sistemáticos ou aleatórios requer atenção especial na análise desses dados. Dentre essas causas especiais estão:

- diferenças na eficiência da incorporação dos corantes;
- diferenças na quantidade de RNA inicial utilizado para marcação e hibridização;
- diferenças de ajuste de parâmetros do *scanner* de leitura das lâminas;
- falhas na impressão das sondas;
- imprecisão de equipamentos;
- procedimentos de localização e quantificação adotados pelos *softwares*.

[1] <http://www.r-project.org>
[2] <http://www.bioconductor.org>

Na normalização dos dados toda essa fonte de variabilidade sistemática deve ser identificada e removida. Para uma compreensão melhor, adotaremos alguns conceitos da linguagem de processamento de imagens de *microarray*, conforme descrito a seguir:

- *foreground* é a região ocupada pelo *spot*, que por sua vez é a região onde se encontra sinal para um dado cDNA;
- *background* é a imagem de fundo da lâmina (região onde não se encontram os *spots*);
- ruído é a falta de contribuição de sinal devido as moléculas que não se anelam com nenhuma molécula fluorescente;
- artefato são sinais inespecíficos decorrentes de sujeira na lâmina ou hibridização inespecífica que contaminam o *background*.

Um dos primeiros passos para a análise de dados é a correção do *background* que é feita subtraindo as intensidades do *background* das intensidades do *foreground*, a fim de minimizar os efeitos dos valores de ruído encontrados nos *spots*. Depois disso, os métodos usados para normalizar dados de *microarray* assumem que a maioria dos spots da lâmina não devem apresentar diferenças de expressão entre os tipos biológicos estudados. A normalização por energia total centraliza os dados ao redor de zero. Na normalização dependente da intensidade o objetivo é estimar curvas não lineares capazes de corrigir eventuais vieses sistemáticos observados nos dados. Também é possível ajustar os efeitos da localização geométrica dos *spots* ao longo da lâmina, onde os valores de intensidade de sinal podem variar sistematicamente de acordo com a posição que os seus respectivos *spots* ocupam no substrato utilizado, esta etapa é conhecida como normalização dependente de localização espacial.

Os dados de expressão gênica possibilitam a análise direcionada para grupos de genes específicos que tenham perfis de interação entre si, denominados redes de regulação gênica. Em um trabalho de iniciação científica anterior, desenvolvido pelo autor deste trabalho na cota 2014/2015 (projeto número 3902), foi estudada uma metodologia para a construção de redes de relevância, que é uma técnica de engenharia reversa para construir redes de regulação gênica a partir de dados de expressão (BUTTE; KOHANE, 1999). Esta técnica está mais detalhada na fundamentação teórica.

Este trabalho tem como objetivo principal avaliar a possibilidade de estimar interações entre genes de redes de regulação através da técnica de análise de regressão linear. Deste modo, o presente trabalho está organizado da seguinte maneira. O Capítulo 2, a seguir, apresenta a fundamentação teórica acerca de grafos, redes de relevância e regressão linear múltipla. No Capítulo 3 são apresentados os resultados associados a um modelo de regressão para um pequeno grupo de quatro genes, com as conclusões finais sobre o trabalho no Capítulo 4.

2 Fundamentação Teórica

Esta seção do trabalho aborda alguns aspectos importantes para a análise de dados de expressão gênica. Mais especificamente, a teoria dos grafos, os métodos de construção de redes de relevância e a análise de regressão linear.

2.1 Teoria dos grafos

O entendimento dos mecanismos geradores das diversas patologias que provocam transtornos à humanidade é um dos grandes objetivos da Biologia Molecular. Dentre os alvos principais dos estudos, está o fato de poder representar e compreender o comportamento de vias biológicas, que representam interações entre entidades químicas complexas (como proteínas, substratos ou metabólitos) que ocorrem no nível molecular das células, através da construção de modelos matemáticos e gráficos. A conectividade entre as vias biológicas pode ser analisada por meio de um grafo, no qual vértices e arestas representam componentes biológicos, e há uma notação gráfica associada com cada elemento. Uma aresta pode ser definida como (V, W) e será denotada por VW ou WV, sendo que a aresta VW incide em V e em W e assim são as pontas da aresta. Nesse caso, os vértices V e W são vizinhos ou adjacentes.

Uma das primeiras contribuições cuja solução envolveu conceitos do que viria ser teoria dos grafos foi aplicada pelo matemático suíço Leonhard Euler (1707-1783), em que o objetivo era encontrar um passeio que visitasse todas as pontes da cidade de Königsberg, passando uma única vez em cada ponte. Segundo Feofiloff, Kohayakawa e Wakabayashi (2011), um grafo não pode ter duas arestas diferentes com mesmo par de pontas (ou seja, não pode ter arestas "paralelas") e também não pode ter uma aresta com pontas coincidentes (ou seja, não pode ter "laços"). É importante sempre dar-se um nome ao grafo. Suponha que um grafo chamado G seja dado, então o conjunto dos seus vértices será denotado por $V(G)$ e o número de vértices por $n(G)$. Analogamente, o conjunto de suas arestas será $A(G)$ e o número de arestas denota-se por $m(G)$.

Logo, conclui-se que,

$$n(G) = \#V(G) \quad \text{e} \quad m(G) = \#A(G),$$

onde o símbolo # representa a cardinalidade do conjunto, ou seja, seu número de elementos.

Dado um grafo (V, A), seu complemento será o grafo $(V, V^{(2)} \setminus A)$ denotado por $\bar{G}$. Caso, $A(G) = V(G)^{(2)}$ um grafo G será completo, e vazio se $A(G) = \phi$.

Um grafo é um par (V, A) em que V é um conjunto arbitrário chamado de vértices, e A é um subconjunto de pares de V conhecido como arestas. A necessidade do uso do grafo é muito importante para compreender melhor a ligação entre pares de elementos através de uma espécie de representação gráfica e são extremamente úteis para modelar problemas em muitas áreas de aplicação. É possível citar exemplos em que o uso do grafo é empregado:

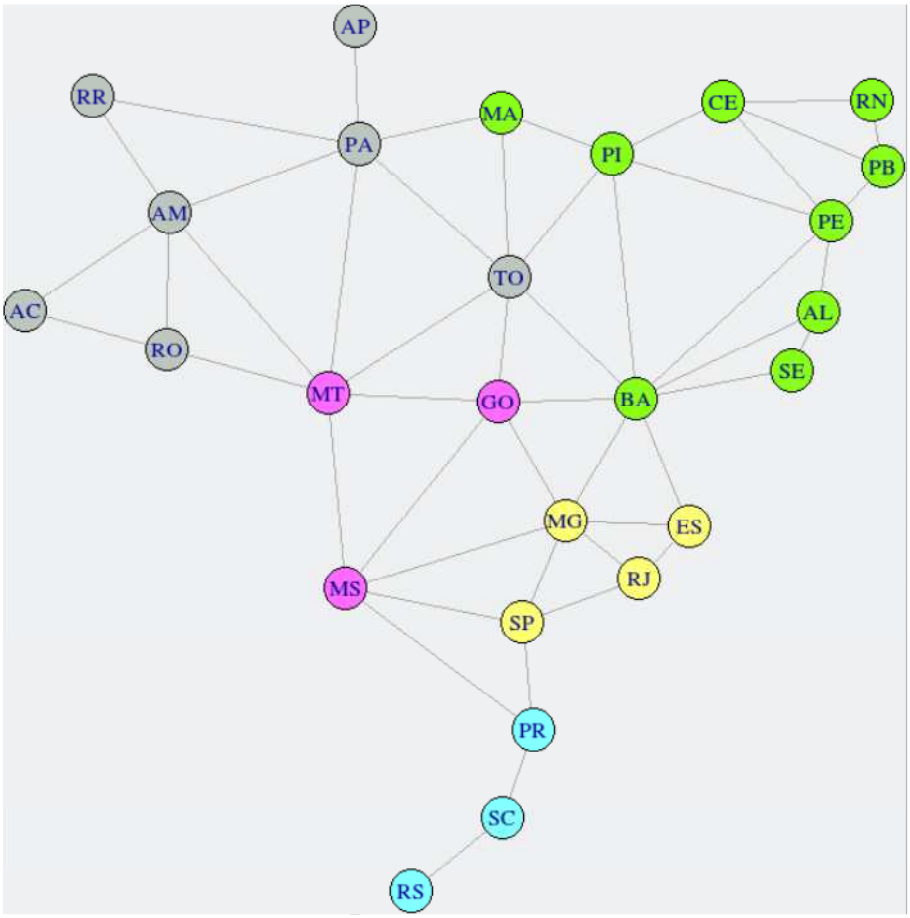
- a malha rodoviária de um Estado, em que as cidades são os vértices, e cada trecho de estrada entre cidades consecutivas é uma aresta;
- um circuito elétrico, onde os vértices são condutores metálicos e as arestas são resistores, capacitores, e outros componentes;
- uma molécula, onde os átomos são os vértices e as arestas são as ligações covalentes;
- uma treliça metálica que pode ser entendida como um grafo onde as arestas são as barras e os vértices são as juntas.

Ainda pode-se citar como exemplo, o grafo dos estados do Brasil, mostrado na Figura 1, elaborado no programa `R` utilizando o pacote `igraph`. Na ilustração da figura, os vértices são exatamente cada estado brasileiro e as arestas representam fronteiras comuns entre esses estados.

Quando as arestas apontam para uma direção associada, normalmente indicada por uma seta na representação gráfica, dizemos que é um grafo direcionado ou dígrafo. No caso em que temos um grafo com um único vértice e nenhuma aresta, chamamos de grafo trivial. Existe isomorfismo entre dois grafos G e H quando é possível alterar os nomes dos vértices de um deles de tal modo que os dois grafos permaneçam iguais. Veremos a seguir alguns tipos de grafos:

- grafo simples é um grafo não direcionado, sem laços em que existe no máximo uma aresta entre quaisquer dois vértices (sem arestas paralelas);
- grafo completo é o grafo simples em que, para cada vértice do grafo, existe uma aresta conectando este vértice a cada um dos demais;
- grafo regular é um grafo em que todos os vértices tem o mesmo grau;
- multigrafo é um grafo que permite múltiplas arestas ligando os mesmos vértices (arestas paralelas);
- árvore é um grafo simples acíclico e conexo;

Figura 1 – Grafo dos Estados do Brasil.



- subgrafo de um grafo G é um grafo cujo conjunto dos vértices é um subconjunto dos vértices de G, e cujo conjunto de arestas é um subconjunto das arestas G, e cuja função W é uma restrição da função G;

- grafo bipartido é o grafo cujos vértices podem ser divididos em dois conjuntos, nos quais não há arestas entre vértices de um mesmo conjunto. Modelar uma rede metabólica utilizando um grafo é equivalente, portanto, a associar entidades biológicas relacionadas às redes metabólicas aos conjuntos de vértices e arestas de um grafo.

O estudo da teoria dos grafos é imprescindível para a compreensão da seção a seguir, pois uma forma de se estudar redes de expressão gênica é através da avaliação das interações entre os níveis de expressão dos genes de interesse, o que nos dá um esboço dos perfis de interação das proteínas produzidas por estes genes. E estas redes são representadas na forma de grafos orientados, onde os vértices representam os genes (ou proteínas) da rede e cada interação entre pares de moléculas é dada por uma aresta orientada que representa uma relação indutora ou repressora.

2.2 Redes de relevância

Um dos objetivos deste trabalho é poder representar e compreender o comportamento de vias biológicas através da construção de modelos matemáticos e gráficos. Para isso o estudo de redes de relevância é essencial para alcançar esse propósito. Tal método nada mais é que uma tentativa de se reconstruir redes de expressão gênica interagentes a partir de dados de expressão gênica obtidos através da técnica de *microarray*.

Uma rede de relevância, introduzida por Butte e Kohane (1999), consiste em calcular o quadrado do coeficiente de correlação linear de Pearson, r^2, entre todos os pares de genes para cada tipo biológico, definindo um grafo completamente interligado. Posteriormente se define um ponto de corte $r^{2'}$ com o objetivo de subdividir o grafo completo em pequenos subgrafos de modo que se tenha $r^2 > r^{2'}$. Tais subgrafos é o que os autores do trabalho original designaram por redes de relevância. O ponto de corte a ser usado pode tanto ser especificado arbitrariamente como estimado, por exemplo, por *bootstrap*. Outras medidas de associação além do coeficiente de correlação linear de Pearson também podem ser usadas. A correlação linear de Pearson é a medida de associação mais utilizada para a estimação da interação entre pares de genes, na qual consegue mensurar o quão duas variáveis numéricas são correlacionadas. O coeficiente de linear de Pearson $r(x, y)$ é dado por,

$$r(x,y) = \sum_{i=1}^{n} \frac{(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{(x_i - \bar{x})^2 (y_i - \bar{y})^2}},$$

onde $\bar{x}$ e $\bar{y}$ são as médias amostrais das variáveis X e Y e x_i e y_i, $i = 1, 2 \cdots, n$, são as respectivas observações de ambas variáveis.

Tal coeficiente varia no intervalo $[-1, 1]$, em que quanto mais próximo de -1 indica forte associação negativa, e ao se aproximar de 1 indica forte associação positiva, ao passo que o valor 0 (zero) representa ausência total de associação linear. As vezes, a desvantagem de utilizar o coeficiente de correlação de Pearson é a influência que ele sofre pela presença de valores extremos (ou *outliers*, do inglês), seja muito baixos ou muito altos. Daí, a necessidade da utilização de medidas mais robustas que superem esses inconvenientes. Os resultados deste tipo de análise são apresentados como grafos, que representam as redes de relevância geradas.

Para a avaliação da significância dos valores de interação obtidos, independentemente da medida de associação utilizada, é possível utilizar estratégias de permutação dos dados, onde os valores observados para genes de interesse são permutados um número grande de vezes e os valores de interação são recalculados em cada repetição do processo. Assim, é possível contar o número de vezes em que se obtém valores maiores que o valor observado nos dados originais e definimos o nível descritivo do teste. Se o interesse for o teste bicaudal, devemos tomar os valores absolutos das estatísticas, se o teste for feito ‘a esquerda ou à direita devemos contar o número de vezes em que a estatística permutada é

menor ou maior que o valor originalmente observado (ESTEVES, 2007).

A seguir é apresentada uma breve revisão sobre o modelo de regressão linear, que foi utilizado na tentativa de se estimar interações entre genes em uma pequena rede de regulação gênica extraída a partir de uma rede de relevância obtida para dados reais de expressão gênica.

2.3 Regressão Linear Múltipla

O modelo de análise de regressão é uma das técnicas estatísticas mais utilizadas nas aplicações em diferentes áreas do conhecimento, tais como computação, administração, engenharias, biologia, agronomia, saúde, sociologia, entre outras. De forma geral, pode-se dizer que o modelo de Regressão Linear Múltipla é qualquer modelo de regressão linear com duas ou mais variáveis independentes com o objetivo de prever uma variável dependente (Y), ou seja, é uma equação linear (ou modelo matemático) que propicia ao pesquisador estimar respostas de uma variável, levando em conta os valores de duas ou mais variáveis explicativas (GAZOLA, 2002). De modo geral, o modelo pode ser representado da seguinte forma:

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

onde y_i e x_{ji} são os valores observados para as variáveis Y (dependente) e X_j (independentes), e β_0, β_j os parâmetros a serem estimados, com $j = 1, \cdots, k$ e $i = 1, \cdots, n$, sendo n o número de observações da amostra, e finalmente, ϵ_i são os erros associados ao modelo.

Torna-se necessário definir uma forma mais geral para obtenção dos estimadores dos parâmetros do modelo de regressão, neste caso, será utilizado o Método dos Mínimos Quadrados (MMQ), dado que é um dos mais empregados (BARRETO et al., 2016). Então, em situações como essa, é mais útil e prático fazer uso de operações matriciais. Assim, é necessário reescrever o modelo algébrico dado acima na forma matricial, para aplicar o MMQ:

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{2.1}$$

em que, na expressão (2.1) acima, tem-se que

$$\boldsymbol{Y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \quad \boldsymbol{X} = \begin{bmatrix} 1 & x_{11} & x_{21} & \cdots & x_{k1} \\ 1 & x_{12} & x_{22} & \cdots & x_{k2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{1n} & x_{2n} & \cdots & x_{kn} \end{bmatrix},$$

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix}, \quad \boldsymbol{\varepsilon} = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix},$$

onde $\boldsymbol{Y}$ representa uma matriz de valores observados para a variável resposta com dimensão $n \times 1$ (n linhas e 1 coluna); $\boldsymbol{X}$ representa uma matriz de variáveis independentes com dimensão $n \times (k+1)$, onde a primeira coluna é reservada para constante 1 e as demais colunas para as k variáveis independentes; $\boldsymbol{\beta}$ é uma matriz coluna com dimensão $(k+1) \times 1$, contendo como primeiro elemento da matriz o intercepto β_0 e $\boldsymbol{\varepsilon}$ constitui uma matriz coluna dos resíduos com dimensão $n \times 1$.

Segundo Hoffmann (2016), todo modelo probabilístico precisa da fixação de premissas nas quais é importante serem atendidas. No entanto, às vezes pode acontecer de tais premissas serem violadas, levando o pesquisador a realizar ajustes necessários. A seguir, segue as suposições sobre o modelo de regressão múltipla:

1. A variável dependente Y é função linear das variáveis explanatórias X_j, $j = 1, \cdots, k$;
2. Os valores das variáveis explanatórias são considerados fixos;
3. $E(\boldsymbol{\varepsilon}) = \mathbf{0}$, onde $\mathbf{0}$ representa um vetor de zeros;
4. Os erros são homocedásticos, isto é, $var(\epsilon_i) = E(\epsilon_i^2) = \sigma^2$;
5. Os erros são não-correlacionados entre si;
6. Os erros têm distribuição normal.

2.3.1 Estimação de Parâmetros

Usando o método dos mínimos quadrados, o qual objetiva a estimação de uma equação de reta que minimize a distância entre os pontos observados e ela, realizando-se em média, a soma dos desvios quadráticos ser igual a zero (ARAÚJO, 2012). Em outras palavras, pretende-se encontrar os valores do vetor $\boldsymbol{\beta}$ que minimizam a soma de quadrados de resíduos, onde estes resíduos configuram uma espécie de estimativa para os erros associados ao modelo.

Sejam $\hat{\boldsymbol{\beta}}$ e $\boldsymbol{e}$ os vetores das estimativas dos parâmetros e dos resíduos (ou desvios) do modelo, respectivamente, em que:

$$\hat{\boldsymbol{\beta}} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \vdots \\ \hat{\beta}_k \end{bmatrix}, \quad \boldsymbol{e} = \begin{bmatrix} \hat{\epsilon}_1 \\ \hat{\epsilon}_2 \\ \vdots \\ \hat{\epsilon}_n, \end{bmatrix},$$

pode-se estimar os valores da variável Y por $\boldsymbol{X}\hat{\boldsymbol{\beta}}$, e tem-se então que

$$\boldsymbol{Y} = \boldsymbol{X}\hat{\boldsymbol{\beta}} + \boldsymbol{e} = \hat{\boldsymbol{Y}} + \boldsymbol{e},$$

e então, é fácil observar que é possível escrever os resíduos do modelo em função dos vetores de valores observados, $\boldsymbol{Y}$, e do vetor das respectivas estimativas

$$\boldsymbol{e} = \boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{Y} - \hat{\boldsymbol{Y}},$$

onde

$$\hat{\boldsymbol{Y}} = \begin{bmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \vdots \\ \hat{y}_n \end{bmatrix}.$$

A partir destes resultados, a soma de quadrados dos resíduos é dada por

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

aplicando a propriedade da transposta, vemos que $\boldsymbol{Y}'\boldsymbol{X}\hat{\boldsymbol{\beta}}$ e $\hat{\boldsymbol{\beta}}'\boldsymbol{X}'\boldsymbol{Y}$ são equivalentes e portanto:

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

Derivando a expressão acima em relação à $\hat{\boldsymbol{\beta}}$ e igualando a zero, para o obtenção do ponto mínimo, temos que

$$\begin{aligned} \frac{\partial}{\partial\hat{\boldsymbol{\beta}}}\left(\boldsymbol{Y}'\boldsymbol{Y} - 2\hat{\boldsymbol{\beta}}'\boldsymbol{X}'\boldsymbol{Y} + \hat{\boldsymbol{\beta}}'\boldsymbol{X}'\boldsymbol{X}\hat{\boldsymbol{\beta}}\right) &= -2\boldsymbol{X}'\boldsymbol{Y} + 2\boldsymbol{X}'\boldsymbol{X}\hat{\boldsymbol{\beta}} \\ &\Rightarrow -2\boldsymbol{X}'\boldsymbol{Y} + 2\boldsymbol{X}'\boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{0} \\ &\Rightarrow \boldsymbol{X}'\boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{X}'\boldsymbol{Y}. \end{aligned}$$

Para isolar o que se deseja, isto é $\hat{\boldsymbol{\beta}}$, é necessário pré-multiplicar ambos os membros da expressão $\boldsymbol{X}'\boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{X}'\boldsymbol{Y}$ pela inversa $(\boldsymbol{X}'\boldsymbol{X})^{-1}$. Deste modo, chega-se a uma matriz identidade $\boldsymbol{I_n}$, que é uma matriz neutra na multiplicação de matrizes. Então

$$\begin{aligned} (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{X}\hat{\boldsymbol{\beta}} &= (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y} \\ \Rightarrow \hat{\boldsymbol{\beta}} &= (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}. \end{aligned} \tag{2.2}$$

Desta forma, o vetor $\hat{\boldsymbol{\beta}}$ dado pela expressão (2.2) acima é a solução de mínimos quadrados para $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}$, desde que exista inversa simples $(\boldsymbol{X}'\boldsymbol{X})^{-1}$.

2.3.2 Somas de quadrados

De forma geral, em Estatística o número de graus de liberdade está associado com a quantidade de termos livres independentes necessários para se obter uma estatística. Por

exemplo, para se estimar a média de uma população a partir de uma amostra de tamanho n, usa-se o estimador

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} x_i,$$

de modo que esta estatística tem n graus de liberdade porque são necessários todos os valores da amostra para se obter a média.

Por outro lado, quando se pensa na estatística usada para se estimar a variância de uma população, geralmente usa-se a seguinte expressão

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2,$$

porém, neste caso o conhecimento da média estimada, $\bar{X}$, é necessário para que se possa estimar a variância dos dados, de modo que são necessários $n-1$ termos livres (graus de liberdade) para se obter S^2.

Além desta interpretação, algumas distribuições de probabilidade teóricas apresentam parâmetros que também são chamados de graus de liberdade, são os casos das distribuições t-Student, qui-quadrado e F-Snedecor.

No caso dos modelos de análise de regressão linear é possível definir algumas estatísticas conhecidas como somas de quadrados que geralmente estão associadas com a variabilidade total e o quanto desta variabilidade pode ser explicada pelo modelo de regressão ou fica caracterizada apenas no resíduos. Em geral tais estatísticas são dadas por:

- SQ_{tot} é a soma de quadrados total, que representa a variabilidade total dos dados;
- SQ_{reg} é a soma de quadrados da regressão, que denota a variabilidade explicada pelo modelo de regressão;
- SQ_{res} é a soma de quadrados dos resíduos, representando a parcela da variabilidade representada nos resíduos.

É possível demonstrar, através da teoria de probabilidades que estas somas de quadrados têm distribuições conhecidas, mais especificamente todas as três estatísticas dadas acima seguem distribuição de qui-quadrado (denotada por χ^2), sendo que as somas de quadrados total, de regressão e de resíduos, apresentam números de graus de liberdade dados por $n-1$, $k-1$ e $n-k$, respectivamente, ou usando a notação mais classicamente

conhecida em teoria de probabilidade, tem-se que

$$SQ_{tot} \sim \chi^2_{(n-1)}.$$
$$SQ_{reg} \sim \chi^2_{(k-1)},$$
$$SQ_{res} \sim \chi^2_{(n-k)},$$

A partir destes resultados das distribuições de probabilidades associadas às somas de quadrados, também define-se os quadrados médios de regressão e de resíduos, que são dados como a razão cada soma de quadrados e seu respectivo número de graus de liberdade, ou seja

- $QM_{reg} = \frac{SQ_{reg}}{k-1}$ é o quadrado médio da regressão,
- e $QM_{res} = \frac{SQ_{res}}{n-k}$ é o quadrado médio dos resíduos.

Ainda de acordo com a teoria de probabilidades também é possível se demonstrar que a razão entre duas variáveis independentes com distribuição de qui-quadrado, cada uma dividida pelo seu respectivo número de graus de liberdade, resulta em uma nova variável com distribuição F-Snedecor, de modo que é fácil perceber que

$$F = \frac{QM_{reg}}{QM_{res}} \sim F_{(k-1,n-k)}. \tag{2.3}$$

A equação (2.3) pode ser usada para se verificar se existe algum tipo de regressão entre a variável dependente Y e pelo menos uma das variáveis explicativas usadas no modelo, em um teste conhecido como teste F da análise de variância do modelo de regressão. Neste teste se a estatística F calculada a partir da amostra for maior do que o quantil da distribuição $F_{(k-1,n-k)}$ para um nível de significância α, ou seja se $F_{calc} > F_{(k-1,n-k);\alpha}$, rejeita-se a hipótese nula de que nenhum parâmetro seja significativo, o que implica que pelo menos um dos parâmetros β_j, com $j = 1, 2, \cdots, n$, difere estatisticamente de zero ao nível de significância α.

Toda informação relativa as estatísticas de somas de quadrados, quadrados médios e estatística F pode ser sistematizada na forma de uma tabela, conforme o que está representado na Tabela 1, que é frequentemente denominada de tabela da análise de variância (*ANOVA*, do inglês *analysis of variance*). Na linha de títulos da tabela, os graus de liberdade estão abreviados como GL, as somas de quadrados como SQ, os quadrados médios como QM e a estatística F como F.

Tabela 1 – Análise de variância para a regressão linear múltipla.

Causas de variação	GL	SQ	QM	F
Regressão	$k-1$	$SQreg$	$\frac{SQreg}{k-1}$	$\frac{QMreg}{QMres}$
Resíduo	$n-k$	$SQres$	$\frac{SQres}{n-k}$	-
Total	$n-1$	$SQtot$	-	-

2.3.3 Coeficiente de Determinação R^2

A partir de um conjunto de dados observado, uma vez que o modelo foi ajustado e verificou-se a existência de regressão significativa é importante verificar o quanto da variabilidade da variável dependente Y pode ser explicada pelo modelo ajustado.

Após o ajuste do modelo, se a soma de quadrados de resíduos ficar próxima de zero, ou seja se $SQ_{res} \approx 0$, então a maior parte da variabilidade de Y é explicada pelo modelo, e consequentemente a soma de quadrados de regressão deve ficar próxima da soma de quadrados total, isto é $SQ_{reg} \approx SQ_{tot}$. Neste caso uma medida simples para o quanto da variabilidade total é explicada pela regressão é o coeficiente de determinação, dado por

$$R^2 = \frac{SQ_{reg}}{SQ_{tot}}, \quad 0 \leq R^2 \leq 1,$$

que, naturalmente, quanto mais próximo de 1 estiver, melhor para a qualidade do ajuste do modelo.

Porém, o R^2 é inflacionado pelo número de variáveis independentes incluído no modelo. Então, para evitar esse problema recomenda-se utilizar o coeficiente de determinação ajustado, definido por

$$R^2{}_a = 1 - (1 - R^2)\left(\frac{n-1}{n-P-1}\right)$$

em que,

- R^2 é o coeficiente de determinação;
- P é o número de parâmetros;
- n é o numero de observações.

2.3.4 Testes de Hipóteses

Após a realização da análise de variância, é possível testar a hipótese nula de que todos os parâmetros do modelo de regressão linear são nulos. Contudo, se a hipótese nula for rejeitada pelo teste F da $ANOVA$, isso significa dizer que pelo menos um dos parâmetros é diferente de zero. Mas não é possível afirmar qual ou quais deles especificamente apresentam esta característica.

Então, para se identificar quais parâmetros em particular diferem ou não de zero, é necessário se formular hipóteses para cada um dos parâmetros do modelo. Assim, como se sabe da inferência estatística, para se fazer testes de hipóteses ou construir intervalos de confiança, é necessário conhecer a distribuição amostral do estimador do(s) parâmetro(s) de interesse, para que seja possível encontrarmos o erro padrão dos estimadores. Neste caso, como a variância real da população não é conhecida, se faz necessária a utilização do teste t de Student, que se baseia na distribuição de mesmo nome, para cálculo das estatísticas de teste para cada parâmetro individual.

2.3.5 Intervalos de Confiança de Regressão para $\boldsymbol{\beta_j}$

A estimação pontual dos parâmetros do modelo não fornece a ideia de margem de erro, que é fornecida ao se estimar um determinado parâmetro. No entanto, a estimação por intervalo procura corrigir essa lacuna a partir da criação de um intervalo que garanta uma alta probabilidade de conter o verdadeiro valor desconhecido do parâmetro de interesse. Deste modo, o intervalo de confiança para se estimar os coeficientes β_j do modelo de regressão é definido por

$$IC(\beta_j) = \beta_j \pm \sigma(\beta_j),$$

onde $\sigma(\beta_j) = \sqrt{\sigma^2 C_{jj}}$, sendo C_{jj} o elemento da diagonal principal da matriz $(\boldsymbol{X'X})^{-1}$ referente ao parâmetro β_j estimado e σ^2 representa o quadrado médio de resíduos (QM_{res}) da *ANOVA*.

2.3.6 Análise de Resíduos

Uma vez determinado o modelo de regressão, deve-se verificar se o mesmo é adequado para a explicação da relação estatística entre as variáveis. A análise de resíduos é uma peça fundamental na análise de modelos lineares e deve ser realizada sempre na análise de regressão.

É notória a importância da análise de resíduos, pois serve para verificar todas as pressuposições do modelo (como normalidade, independência, homocedasticidade), sob as quais foram considerados toda a inferência. Pode-se também determinar outras relações funcionais para os dados, como por exemplo uma relação não linear. E ainda fornece ferramentas para a identificação da presença de *outliers* (pontos atípicos) ou pontos influentes no modelo ajustado.

Existem basicamente três tipos de resíduos que são utilizados na análise de regressão, são eles: resíduo ordinário, resíduo padronizado e resíduo estudentizado. Tais resíduos foram propostos com o objetivo de solucionar problemas específicos em relação à forma tradicionalmente utilizada. Para o resíduo bruto temos a relação

$$e_i = y_i - \hat{y}_i, \quad i = 1, 2, \cdots, n.$$

Os demais resíduos necessitam do conhecimento de uma matriz denominada $\boldsymbol{H}$. Sabe-se que $\hat{\boldsymbol{Y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}}$, em que

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

Assim $\hat{\boldsymbol{Y}} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}$, o que implica que $\hat{\boldsymbol{Y}} = \boldsymbol{H}\boldsymbol{Y}$, com

$$\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'.$$

A matriz $\boldsymbol{H}$ é denominada de matriz chapéu (ou *hat*, do inglês). Os elementos de $\boldsymbol{H}$ são denotados por h_{ij}, sendo que ela quadrada, de ordem $n \times n$, e simétrica ($h_{ij} = h_{ji}$), sendo representada por

$$H = \begin{bmatrix} h_{11} & h_{12} & \cdots & h_{1n} \\ h_{21} & h_{22} & \cdots & h_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ h_{n1} & h_{n2} & \cdots & h_{nn} \end{bmatrix}.$$

Na notação matricial, o vetor de resíduos é expresso por

$$\boldsymbol{e} = \boldsymbol{Y} - \hat{\boldsymbol{Y}} = \boldsymbol{Y} - \boldsymbol{H}\boldsymbol{Y} = (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{Y},$$

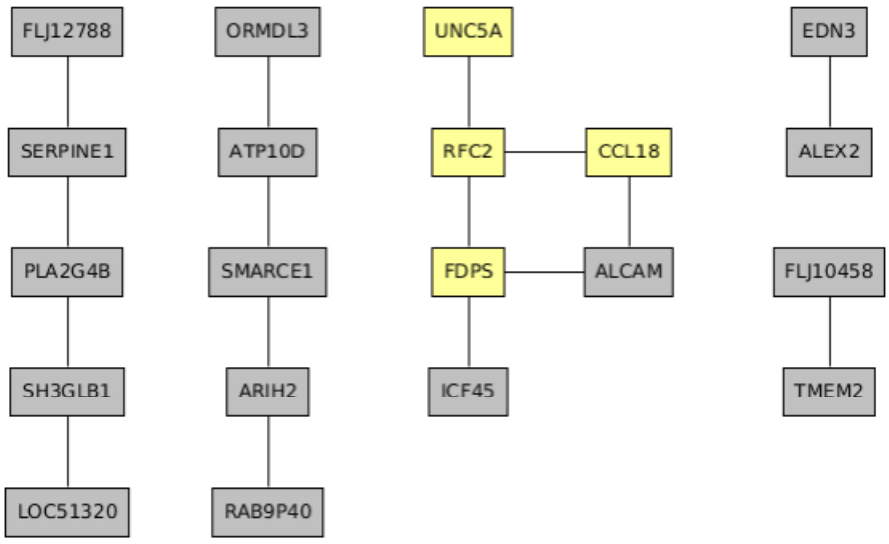
e obtendo-se a esperança e variância de $\boldsymbol{e}$, sabe-se que sob a suposição $\epsilon_i \sim N(0, \sigma^2)$, então, $\boldsymbol{Y} \sim N(\boldsymbol{X}\boldsymbol{\beta}; \sigma^2\boldsymbol{I})$ e $\hat{\boldsymbol{Y}} \sim (\boldsymbol{X}\boldsymbol{\beta}; \sigma^2\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}')$.

3 Aplicação

Esta seção apresenta alguns resultados extraídos a partir de um banco de dados real de expressão gênica obtido através da técnica de *microarray*, em um trabalho publicado previamente por Gomes et al. (2005). Este banco de dados contém informações acerca dos níveis de expressão de cerca de 4.500 genes para amostras de RNA extraídas de tecidos gastro-esofágicos de 71 pacientes, sendo 39 observações de esôfago e da junção gastro-esofágica (9 normais, 6 com esofagite e 10 com mucosa de Barrett), e 32 observações de tecidos de estômago (11 normais, 9 metaplasia e 12 adenocarcinomas). Maiores detallhes sobre o banco de dados podem ser obtidos diretamente do artigo.

Os dados foram obtidos já pré-processados e normalizados (conforme descrito no artigo) e foram usados previamente, em um outro trabalho onde foi obtida uma rede de relevância a partir de alguns genes obtidos aleatoriamente, que está representada na Figura 2. Na figura observa-se um grupo de genes representado na forma de um grafo não orientado onde vértices representam os genes da rede e as arestas representam associações hipotéticas entre os genes.

Figura 2 – Rede de relevância obtida a partir de alguns genes escolhidos ao acaso do banco de dados.



Neste caso, para simplificar o problema neste ponto, foi usado apenas o subgrafo representado pelos vértices destacados em amarelo na Figura 2, tentando explicar a expressão do gene `RFC2`, em função da expressão dos genes `UNC5A`, `FDPS` e `CCL18`.

Deste modo, pretendeu-se aqui averiguar de que forma os níveis de expressão dos genes `UNC5A`, `FDPS` e `CCL18` influenciam no nível de expressão do gene `RFC2`, a partir de

um modelo de regressão linear múltipla, dado pela seguinte expressão

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \epsilon_i, \quad i = 1, 2, \cdots, 71,$$

onde tem-se que:

- ✓ y_i é o nível de expressão do gene `RFC2` na i-ésima observação (variável dependente);
- ✓ x_{1i} é o nível de expressão do gene `UNC5A` na i-ésima observação;
- ✓ x_{2i} é o nível de expressão do gene `FDPS` na i-ésima observação;
- ✓ x_{3i} é o nível de expressão do gene `CCL18` na i-ésima observação.

Neste sentido foi utilizado o programa estatístico R para se fazer o ajuste do modelo proposto acima. Assim, foi realizada a estimação dos parâmetros através do método dos mínimos quadrados, cujo resultado da análise de variância está na Tabela 2.

Tabela 2 – Análise de variância para o modelo proposto.

Causas de variação	GL	SQ	QM	F	valor p
`UNC5A`	1	$0,0026$	$0,0026$	$0,1023$	$0,7501$
`FDPS`	1	$0,4038$	$0,4038$	$15,8888$	$< 0,001$
`CCL18`	1	$0,0216$	$0,0216$	$0,8482$	$0,3604$
Resíduo	67	$1,7025$	$0,0254$	-	-
Total	70	$2,1303$	-	-	-

De acordo com a Tabela 2, pode-se verificar que a variável `FDPS` foi a única significativa para explicar o nível de expressão do gene `RFC2`. Esta conclusão pode ser confirmada observando o p-valor menor que 0,05 onde rejeita-se a hipótese de que o parâmetro β_2 é igual a zero.

O coeficiente de determinação calculado para o modelo ajustado ficou dado por

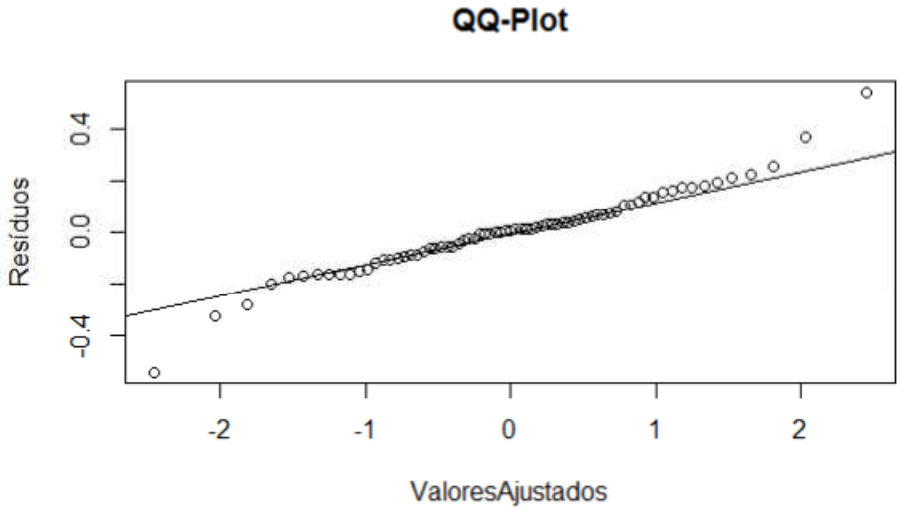
$$R^2 = \frac{0,4279}{2,1303} = 0,2009;$$

isto é, cerca de $20,09\%$ da soma de quadrados total é "explicada" pela regressão linear ajustada. Já o coeficiente de determinação ajustado foi de $16,51\%$. Então, sem levar em conta a significância dos parâmetros e a baixa explicação da variabilidade dos dados, o modelo ajustado tem como equação

$$Y = -0,0173 - 0,0621x_1 + 0,1606x_2 - 0,0099x_3.$$

Com o intuito de verificar se o modelo acima é adequado, em seguida, é feita uma análise dos pressupostos da regressão linear múltipla. Observando a Figura 3, parecem

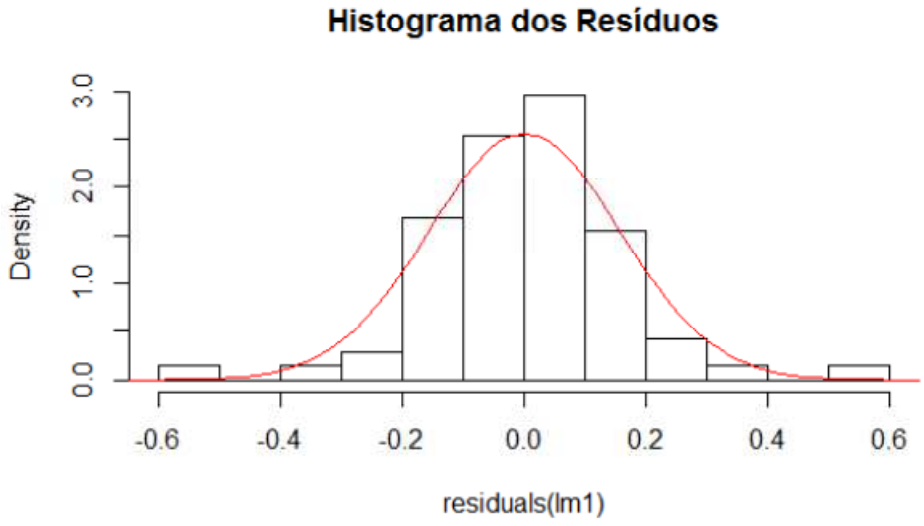
Figura 3 – Normal qq-plot da regressão dos resíduos estandardizados.



existir alguns pontos, principalmente nas caudas, que se afastam da diagonal principal, não sendo conclusivos quanto à normalidade dos resíduos.

Ainda para se avaliar a normalidade dos resíduos, um outro tipo de análise gráfica interessante é o histograma, apresentado na Figura 4. No gráfico, embora o histograma pareça mais ou menos ajustado à curva da distribuição normal, alguns problemas também sugerem uma baixa aderência.

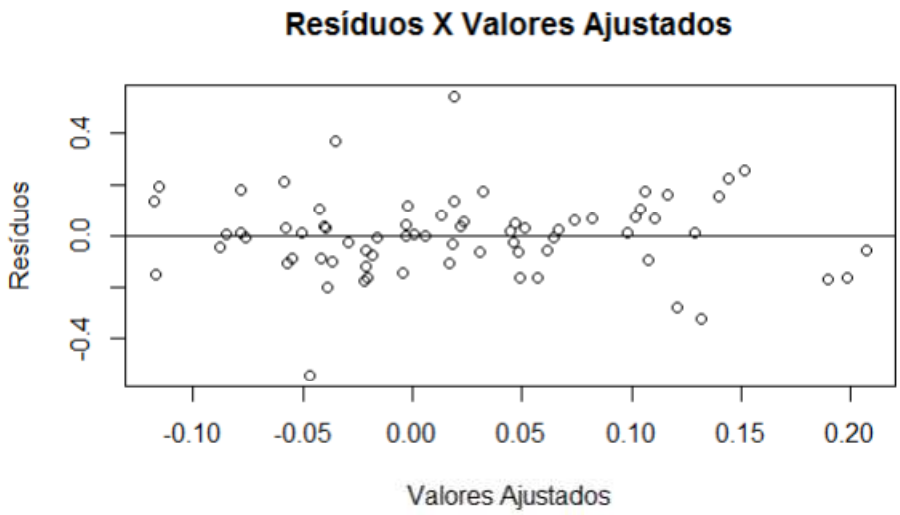
Figura 4 – Histograma dos resíduos estandardizados.



Uma forma de confirmar a normalidade dos resíduos é a execução do teste de Shapiro-Wilk, que resultou em um valor de significância obtido de 0,0213 que aponta para a não normalidade dos resíduos, devido a rejeição da hipótese nula.

A partir da análise gráfica dos resíduos estandardizados mostrado na Figura 5, é possível se observar que os resíduos se distribuem aleatoriamente em torno de zero. Com isso, conclui-se que os resíduos parecem ser homoscedásticos, conforme a suposição do modelo de regressão.

Figura 5 – Gráfico dos Resíduos *versus* Valores Ajustados.



Como se trata de uma análise de regressão linear múltipla, um dos pressupostos que precisa ser verificado é a existência de colinearidade entre as variáveis independentes. Para isso, foi realizado o teste de Fator de Inflação da Variância *(VIF)*, o qual mede o quanto da variância do coeficiente é inflacionada por sua colinearidade.

Segundo (HAIR; ANDERSON; TATHAM, 2005) um *VIF* máximo acima de 5 indica que a multicolinearidade pode estar influenciando as estimativas de mínimos quadrados. A partir da Tabela 3, notamos que as três variáveis apresentam valores de *VIF* menores que 5 e assim é possível se concluir que não parece existir problemas de colinearidade.

Tabela 3 – Valores do Fator de Inflação da Variância para cada variável.

Variáveis	Valores < 5
UNC5A	$1,168$
FDPS	$1,038$
CCL18	$1,131$

Se não fosse o problema de significância dos parâmetros e um leve indício de violação da normalidade dos resíduos, o modelo poderia ser considerado adequado. Enfim, levando-se em consideração a significância estatística para o modelo ajustado, foram retiradas as variáveis X_1 e X_3, relativas aos níveis de expressão dos genes `UNC5A` e `CCL18`, respectivamente, que não foram significativas. Logo em seguida o modelo foi ajustado novamente apenas com a inclusão da variável X_2, sendo que a equação encontrada foi

$$Y = -0,0293 + 0,1575 \times X_2.$$

Neste caso, a interpretação do modelo é a seguinte:

- quando a variável X_2 (associada ao nível de expressão do gene `FDPS`) for igual a zero, logo, a variável resposta, ou seja, o nível de expressão do gene `RFC2` será a própria média, isto é
$$Y = -0,0293$$
- no entanto, quando a variável X_2 variar, então, a variável resposta, ou seja, o nível de expressão do gene `RFC2` variará em 0,1575 unidades, em média.

Finalmente, como o objetivo deste trabalho não foi o de apresentar um modelo de regressão múltipla significativo, mas sim verificar a possibilidade de se usar a abordagem de regressão linear para esse tipo de dados. Assim, pode-se dizer que o objetivo foi atingido.

4 Conclusão

Inicialmente, focou-se no estudo teórico do problema abordado, que era construir redes de relevância, técnica de engenharia reversa para construir redes de regulação gênica a partir de dados de expressão. Dentre os alvos principais dos estudos, estava o fato de poder representar e compreender o comportamento de vias biológicas, que representam interações entre entidades químicas complexas (como proteínas, substratos ou metabólitos) que ocorrem no nível molecular das células, através da construção de modelos matemáticos e gráficos. Neste sentido, foi feito um aprofundamento na teoria dos grafos, imprescindível para a compreensão das redes de relevância geradas, em que os resultados deste tipo de análise são apresentados como grafos.

Assim, foi avaliada a possibilidade de estimar interações entre genes de redes de regulação utilizando a técnica de análise de regressão linear a partir de dados de expressão. A ideia inicial era averiguar de que forma os genes `UNC5A`, `FDPS` e `CCL18` influenciavam no nível de expressão do gene `RFC2`. Para isso, o modelo foi estimado e os resultados foram apresentados como se todas as variáveis fossem significativas, apesar de que a única significativa para explicar o nível de expressão do gene `RFC2`, foi a variável `FDPS`, reduzindo o modelo para uma regressão linear simples.

Mas como o objetivo principal deste trabalho não era apresentar um modelo de regressão linear múltipla significante, e sim mostrar que era possível estimar interações entre genes de redes de regulação gênica. Conclui-se, então, que o propósito estabelecido foi cumprido.

Este trabalho contribuiu também para incentivar cada vez mais a pesquisa na área da Biologia Molecular, que ainda tem muito a colaborar com a sociedade, como por exemplo, encontrar mecanismos eficazes que ajudem na cura do câncer. Um outro ponto a ser destacado no trabalho é a respeito do olhar científico e de toda a experiência adquirida.

Referências

ARAÚJO, F. M. de. *Uso da técnica bootstrap em modelos de regressão não linear*. 2012. Trabalho de Conclusão de Curso. Universidade Estadual da Paraíba. Citado na página 19.

BARRETO, V. C. S. et al. Regressão linear múltipla aplicada ao preço do leite. *Revista Eletrônica Paulista de Matemaítica*, v. 7, p. 109–118, 2016. Citado na página 18.

BUTTE, A. J.; KOHANE, I. S. Unsupervised Knowledge discovery in medical databases using relevance networks. In: *Proc. AMIA Symp.* [S.l.: s.n.], 1999. p. 711–715. Citado 2 vezes nas páginas 13 e 17.

ESTEVES, G. H. *Métodos estatísticos para a análise de dados de cDNA microarray em um ambiente computacional integrado*. Tese (Doutorado) — Universidade de São Paulo, 2007. Disponível em: <http://www.teses.usp.br/teses/disponiveis/95/95131/tde-03062007-210232/>. Citado 3 vezes nas páginas 11, 12 e 18.

FEOFILOFF, P.; KOHAYAKAWA, Y.; WAKABAYASHI, Y. *Uma introdução suscinta à teoria dos grafos*. São Paulo, 2011. Disponível em: <http://www.ime.usp.br/~pf/teoriadosgrafos/>. Citado na página 14.

GAZOLA, S. *Construção de um modelo de regressão para avaliação de imóveis*. Dissertação (Mestrado) — Universidade Federal de Santa Catarina, Florianópolis, 2002. Citado na página 18.

GOMES, L. I. et al. Expression profile of malignant and nonmalignant lesions of esophagus and stomach: differential activity of functional modules related to inflammation and lipid metabolism. *Cancer Res.*, v. 65, n. 16, p. 7127–36, aug 2005. ISSN 0008-5472. Citado na página 26.

HAIR, J.; ANDERSON, R.; TATHAM, R. *Analise Multivariada de Dados*. Bookman, 2005. ISBN 9788536304823. Disponível em: <https://books.google.com.br/books?id=LxFb5JzXdbUC>. Citado na página 29.

HOFFMANN, R. *Análise de Regressão: Uma Introdução à Econometria*. 5. ed. Piracicaba: Portal de Livros Abertos da USP, 2016. Citado na página 19.

LIM, H. A.; VENKATESH, T. V. Bioinformatics in the pre- and post-genomic eras. *Trends Biotechnol*, v. 18, p. 133–135, 2000. Citado na página 11.

LOPES, H. S.; CRUZ, L. M. (Ed.). *Computational biology and applied bioinformatics*. [S.l.]: InTech, 2011. ISBN 978-953-307-629-4. Citado na página 11.

R Development Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: [s.n.], 2011. Disponível em: <http://www.r-project.org>. Citado na página 12.

VIDAL, M. Interactome modeling. *FEBS letters*, v. 579, n. 8, p. 1834–8, mar. 2005. ISSN 0014-5793. Disponível em: <http://www.ncbi.nlm.nih.gov/pubmed/15763560>. Citado na página 11.

APÊNDICE A – Rotina utilizada na análise

Todo o trabalho apresentado foi desenvolvido com o uso do *software* de programação estatística R. Este apêndice é dedicado a apresentar a rotina dos comandos utilizados nas análises.

```
####Aplicação de regressão múltipla######

## carregando os dados
dados = load("C:/Users/Usuario/Dropbox/modelo_TCC/aplicacao/dados2.RData")

load("dados2.RData")
tabela <- as.data.frame(tabela);tabela
attach(tabela)

###nome de cada varável
colnames(tabela)

###Sumário de variáveis
summary(tabela)

#####Estimando o modelo de regressão linear múltipla.

lm1 <- lm(RFC2 ~ UNC5A + FDPS + CCL18, data = tabela);lm1

##############################################################################

####Testando a significância do modelo de regressão.
summary(lm1)

###Tabela da ANOVA
anova(lm1)

#####Verificando os pressupostos do modelo

### Resíduos
res = rstandard( lm(RFC2 ~ UNC5A + FDPS + CCL18, data = tabela));res

## Histograma com sobreposição da curva da normal
```

```
hist(residuals(lm1),main = "Histograma dos Resíduos", prob=TRUE)
curve(dnorm(x, 0, sd(residuals(lm1))), -1, 1, col=2, add=TRUE)

## Teste de Normalidade(Shapiro-Wilk)
shapiro.test(res)

####Q-Q PLOT
qqnorm(residuals(lm1), main ="QQ-Plot", ylab="Resíduos",
xlab="ValoresAjustados")
qqline(residuals(lm1))

### Análise dos Resíduos

## Teste de homogeneidade das variâncias(Breush-Pagan)
library(car)
ncvTest( lm(RFC2 ~ UNC5A + FDPS + CCL18, data = tabela))

### Valores Ajustados x Resíduos

plot(fitted(lm1), residuals(lm1), xlab="Valores Ajustados", ylab="Resíduos",
main = "Resíduos X Valores Ajustados")
abline(h=0)

###teste de independência das variáveis (multicolinearidade)
library(car)
vif(lm1)

###Determinando os ICs a 95% para os parâmetros do modelo.
confint(lm1)

### Reestimando o modelo de regressão já sem as variáveis UNC5A e CCL18,
### as quais não foram significativas.

lm2 <- lm(RFC2 ~ FDPS, data = tabela);lm2

###Descrição do modelo
summary(lm2)

###Tabela da ANOVA
anova(lm2)
```