NORMALIZAÇÃO DE BOX-COX EM DADOS DE EXPERIMENTOS COM EXPRESSÃO GÊNICA

NATÁLIA FARAJ MURAD[1], ROSIANA RODRIGUES ALVES[2]

RESUMO

Este trabalho foi realizado com o objetivo de fazer um estudo sobre a normalidade das sondas de um experimento com expressão gênica cujos dados foram gerados a partir de microarrays de humanos. As sondas foram normalizadas através da transformação de Box & Cox para que se ajustassem às pressuposições de normalidade. A transformação de Box-Cox não foi eficiente em todas as sondas, mas o teste de Shapiro-Wilk permitiu selecionar aquelas que obedeciam aos critérios de normalidade propiciando assim, estimadores válidos.

Palavras-chaves: Box & Cox, expressão gênica, normalização

INTRODUÇÃO

Experimentos com microarrays têm sido bastante usadosna genômica funcional para o estudo de padrões de expressão gênica de genes ligados a características de interesse. Em um experimento com microarray, um fragmento de DNA conhecido é fixado a uma superfície sólida, usualmente uma membrana de náilon ou lâmina de vidro. É aplicada uma mistura de ácidos nucléicos marcada com marcadores radioativos ou fluorescentes. As sequências complementares irão se hibridizar e emitir uma fluorescência que é detectada por um scanner automatizado. Então, é feita a quantificação do material expresso nas lâminas e a análise estatística que consiste na interpretação dos dados (Pierce, 2004).

Neste tipo de experimento, a variável que se deseja analisar é a razão da hibridização entre duas amostras de cDNA que competem por um mesmo sítio, o que é obtido pela intensidade da fluorescência emitida por cada uma das amostras (Pereira, 2006).

O primeiro passo da análise estatística consiste na normalização dos dados. Essa transformação é feita após os dados já terem sofrido a transformação padrão logaritmo da intensidade luminosa e tem como objetivo ajustar os efeitos decorrentes da variação na matriz de micro tecnologia, para que eles não sobreponham às diferenças biológicas entre as amostras de RNA, ou entre as sondas impressas (Smyth et al, 2003).

Segundo Broche (2003), dentre as razões para que a transformação seja feita estão: colocação de quantidades diferentes de mRNA inicial, diferenças de eficiência de detecção do marcador utilizado, erros sistemáticos ao medir os erros de expressão. Dessa forma, os dados não obedecem às pressuposições da análise de variância (distribuição normal, variância constante e independência de sua média) e o modelo análise de variância comum não se ajusta de maneira ideal.

Box & Cox (1964) propuseram um tipo de transformação para dados que não obedeciam **às** pressuposições de normalidade e essa transformação tem sido amplamente utilizada para normalização de dados de diversos tipos de experimentos. Pode ser aplicada a regressões, a uma combinação delas ou a variáveis dependentes numa regressão fazendo com que os resíduos da regressão sejam mais homocedasticos ou mais próximos de uma distribuição normal. A transformação é baseada em uma função de verossimilhança que faz o ajuste dos dados através da expressão:

$$Y^{\lambda} = \begin{cases} \log(Y) \text{ se } \lambda = 0 \\ \dfrac{Y^{\lambda} - 1}{\lambda} \text{ se } \lambda \neq 0 \end{cases}$$

Onde $\boldsymbol{\lambda}$ é o valor que se procura estimar, pois a partir dele os dados transformados são encontrados e então é possível fazer a análise de variância. **Y** é a variável resposta a ser transformada e deve ser maior que zero.

[1] Graduanda em Ciências Biológicas, DBI/UFLA, fnataliam@gmail.com
[2] Mestranda em Estatística e experimentação agropecuária, DEX/UFLA, rosiestrela@gmail.com

MATERIAL E MÉTODOS

Os dados utilizados foram retirados do 15º Genetic Analysis Workshop que ocorreu em novembro de 2006. Os microarrays foram obtidos a partir de células linfoblastóides de 14 famílias, utilizando-se Affymetrix Human Focus Array que contém sondas para 8500 transcritos. Para 3554 das 8500 sondas testadas, Morley et al (2004) encontraram maior variação dentre os indivíduos do que em segmentos replicados num mesmo indivíduo. Essas 3554 expressões fenotípicas foram escolhidas pelo GAW para análises.

Das 14 famílias, duas possuem 13 e doze, 14 indivíduos. O experimento possui 3554 sondas que indicam os níveis de expressão gênica de cada indivíduo, num total de 194. O arquivo de dados encontra-se organizado em colunas, sendo a primeira referente à família, a segunda ao sexo do indivíduo, representado por 1 (masculino) ou 2 (feminino), a terceira ao número do indivíduo na família e as restantes às sondas.

O conjunto de dados sofreu a transformação de Box & Cox para que se ajustassem às pressuposições de normalidade. A transformação foi feita através da função boxcox( ) do pacote MASS no software R v2.10.0 (R Development Core Team, 2010). Essa função estima por máxima verossimilhança os parâmetros de uma transformação de Box-Cox para cada variável. Essa estimativa é feita a partir do gráfico de normalidade de Box-Cox, que é um gráfico dos coeficientes de correlação entre os eixos vertical e horizontal do gráfico de probabilidade para vários valores do parâmetro λ. O valor de λ correspondente a máxima correlação no gráfico é escolhido (Figura 1).

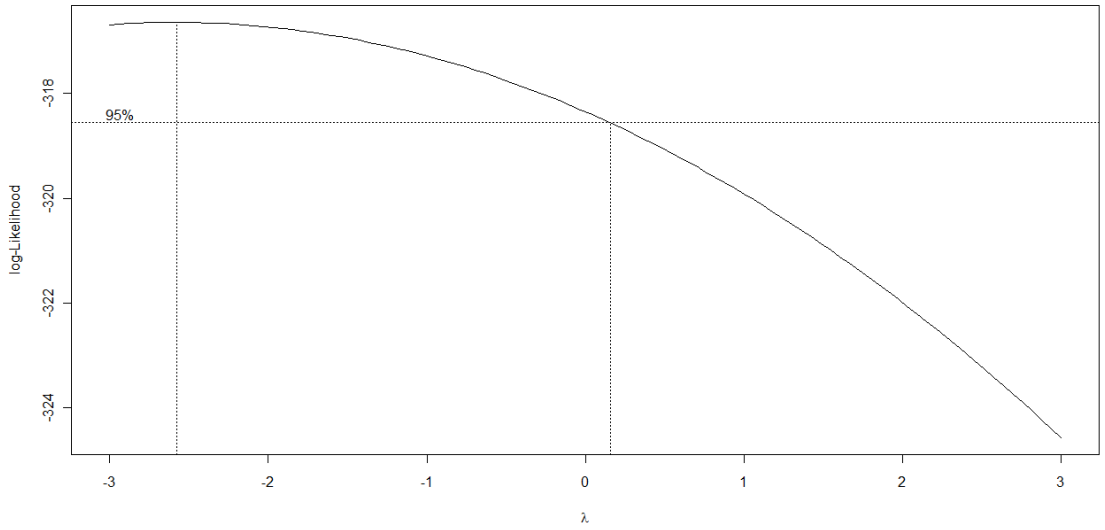

Figura 1: Exemplo do gráfico plotado para o cálculo do **λ** para a sonda 1294_at.

O modelo estatístico aplicado para cada sonda foi:

$$Y_{ij} = \mu + \tau_i + \varepsilon$$

em que:
Y_{ij} representa as intensidades de expressão em escala logarítmica.
μ é o efeito constante (média geral);
$\boldsymbol{\tau_i}$ é o efeito do i-ésimo tratamento (efeito de sexo e família);
$\boldsymbol{\varepsilon_{ij}}$ é o erro associado ao i-ésimo tratamento na j-ésima unidade experimental ou parcela.

Para cada uma das sondas foi obtida uma análise de variância. A avaliação da eficiência da transformação foi feita pelo teste de Shapiro-Wilk.

RESULTADOS E DISCUSSÃO

Segundo Johnson e Wichern (1998), a transformação obtida tende a melhorar a aproximação à normalidade, porém não é garantido que mesmo a melhor escolha de λ produza um conjunto de dados que seja adequado a suposição de normalidade. Através do teste de Shapiro Wilk foi verificado em quais sondas a transformação obteve sucesso.

Antes da normalização, 1914 sondas, isto é, aproximadamente 53% das sondas, apresentavam normalidade confirmada pelo teste de Shapiro-Wilk a 5%. Após a normalização, esse número sofreu um aumento de 802 sondas, ou seja, a porcentagem de sondas normais foi aumentada para 77%. Através da elaboração do gráfico de Normal Q-Q Plot é possível observar o quanto do quantil de probabilidade observado é próximo do esperado. O esperado é o que aconteceria se a função fosse normal em função dos resíduos. Abaixo segue o exemplo de duas sondas, uma que apresentou normalidade e outra que não obedeceu os pressupostos de normalidade. Na Figura 1a, pode-se observar que os dados se aproximam de uma reta, indicando que os dados apresentam distribuição normal. Já na figura 2b, é possível observar que os pontos não tiveram comportamento similar a uma reta, indicando a não-normalidade dos dados.

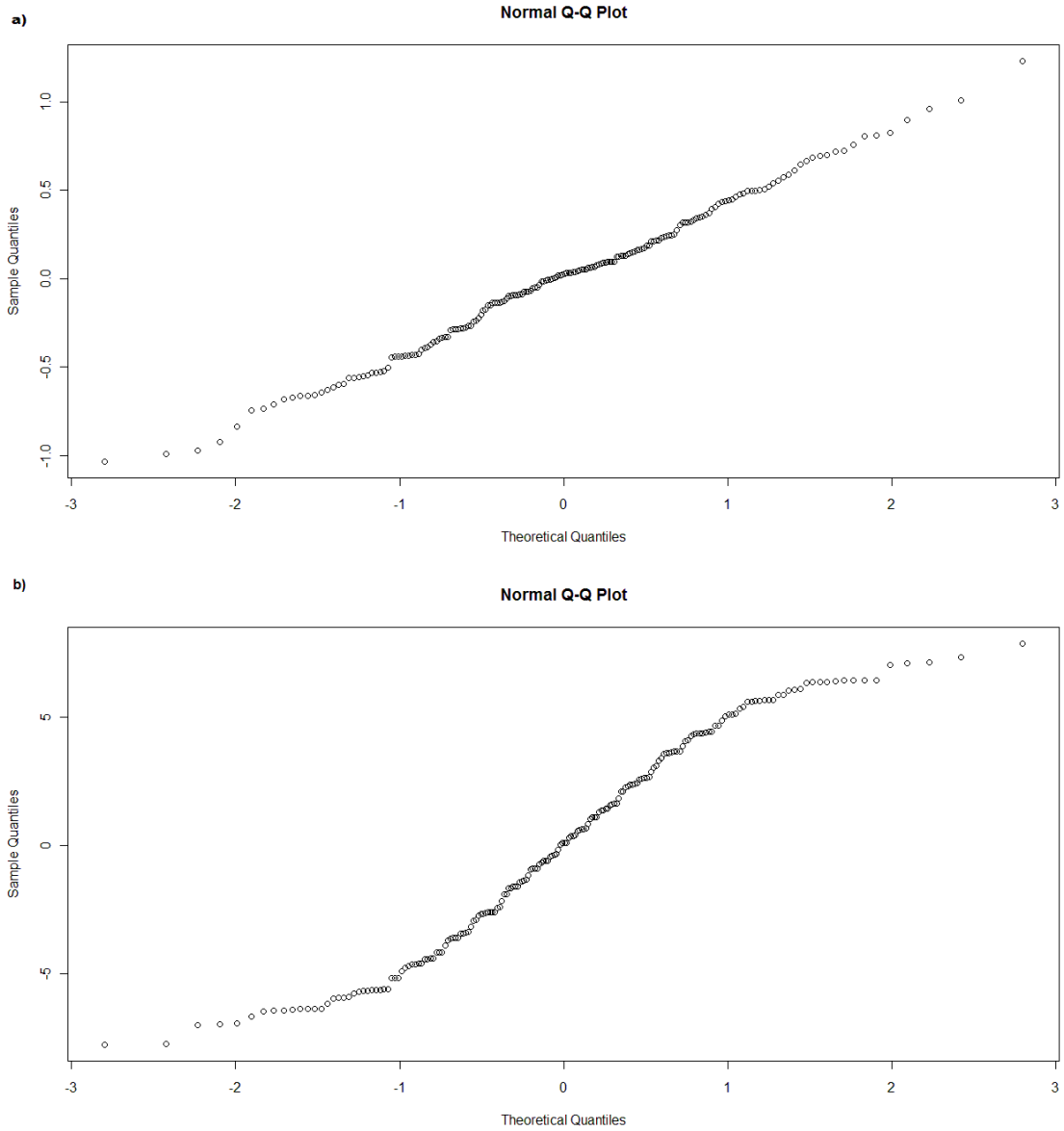

Figura 2: a) Normal Q-Q Plot para as sondas 1487_at (normal) b) 121_at (não-normal). Variável-resposta após ter sofrido normalização de Box-Cox.

Outra maneira de observar se os dados apresentam normalidade é através do gráfico de distribuição do erro. Se os pontos do gráfico se distribuírem de forma aleatória em torno da reta que corresponde ao resíduo zero formando uma área de largura uniforme indicam que os erros são independentes, de média nula e variância constante. Um exemplo disso pode ser observado na Figura 3a, que representa a distribuição dos resíduos para uma sonda que apresentou normalidade. Já na sonda 121_at, não houve normalidade, os resíduos apresentam um comportamento padronizado e, portanto não há independência.

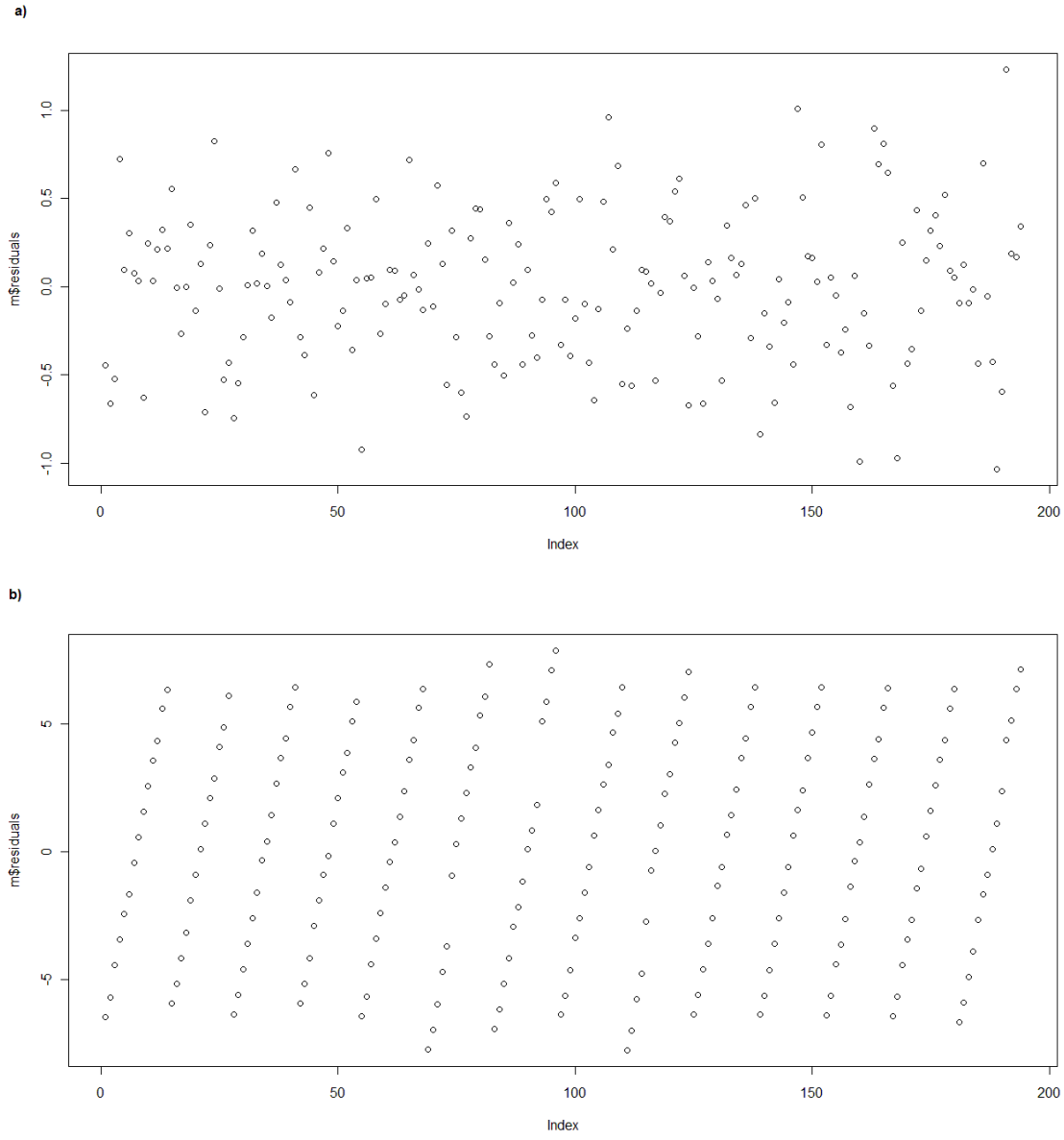

Figura 3: a) Distribuição dos resíduos para a sonda 1487_at (normal). b) Distribuição dos resíduos para a sonda 121_at.

CONCLUSÃO

A transformação de Box-Cox não foi eficiente em todas as sondas, mas o teste de Shapiro-Wilk permitiu selecionar aquelas que obedeciam aos critérios de normalidade propiciando assim, estimadores válidos.

REFERENCIAL BIBLIOGRÁFICO

BOX, G. E. P.; COX, D. R. An analysis of transformations. **Journal of the Royal Statistical Society.** Series B (Methodological), Vol. 26, No. 2., p. 211-252. 1964.

BROCHE, E.C. Métodos Estatísticos na Análise de Experimentos de Microarray. USP-IME. **Dissertação de Mestrado.** 108p. 2003.

JOHNSON, R. A.; WICHERN, D. W. **Applied multivariate statical analysis.** New Jersey: Prentice Hall, 816p. 1998.

MORLEY, M.; MOLONY, C. M.; WEBER, T. M.; DEVLLN, J. L.; EWENS,K. G; PLELMAN, R. S.; CHUNG, V. G. Genetic analysis of genome-wide variation inhuman gene expression. **Nature**, London,v.430, n. 7001, p.743-747, Aug. 2004.

PEREIRA, R. N. Controle do erro tipo I em um experimento de microarrays com eucalipto. 57p. **Tese (Mestrado em Estatística e Experimentação Agropecuária)** – Universidade Feral de Lavras, Minas Gerais, Lavras. 2008.

PIERCE, Benjamin A. **Genética: um enfoque conceitual.** Rio de Janeiro: Guanabara Koogan, p. 423, 546-548, 339. 2004.

R DEVELOPMENT CORE TEAM. R: A Language and Environment for Statistical Computing. Vienna, Austria. R Foundation for Statistical Computing. Disponível em: **http://www.r-project.org. 2010**

SMYTH, G.K.; YANG, Y.H.; SPEED, T. Statistical issues in cDNA microarray data analysis. **Methods Mol. Biol.** 224, 111–136. 2003.