

Determinação de um horizonte preditivo para estimativas de índices de qualidade de ajustamento em modelos de equações estruturais com amostras de tamanho reduzido

Clodoaldo José Figueredo¹

UFPR, Curitiba, PR

Jair Mendes Marques²

UTP, Curitiba, PR

Resumo. Um dos pressupostos para uso de modelos de equações estruturais (MEE) é um número suficientemente grande de observações para obtenção das estimativas dos parâmetros, que possibilitem aferir a qualidade do modelo e as relações entre as variáveis. Essa pesquisa busca desenvolver uma técnica para determinação da qualidade de ajustamento de um modelo aos dados, através do cálculo preditivo de índices de qualidade, obtidos por modelos de regressão não linear, ajustados sobre parâmetros obtidos com o número reduzido de observações. Para alcançar os objetivos propostos, primeiramente foi estipulado um número mínimo de observações, usando para isso, a taxa de convergência do algoritmo iterativo que minimiza a variação dos parâmetros. A estabilidade dos parâmetros em relação ao número de observações, possibilitou a aplicação da técnica de regressão não linear sobre as médias das estimativas dos índices de qualidade. Foram preditos os valores dos índices de qualidade de ajustamento para valores maiores de observações e, os limites dos intervalos de predição, calculados pelo método Delta, indicaram baixa variância nos resultados obtidos sobre amostras maiores. Os resultados mostraram que a aplicação de regressão não linear sobre as estimativas e, a extrapolação do modelo para um número de observações maiores, determina um horizonte preditivo com bom grau de confiabilidade e mínimo erro, possibilitando determinar o ajustamento do modelo de equações estruturais aos dados, mesmo sendo esses de número reduzido.

Palavras-chave. Modelos de Equações Estruturais, Regressão Não Linear, Índices de Qualidade de Ajustamento

1 Introdução

No uso da modelagem de equações estruturais (MEE), nem sempre é possível ter acesso a um número suficientemente 'grande' de observações que possibilite o cálculo dos parâmetros, pela falta de convergência do algoritmo iterativo, ou a determinação da qualidade de ajustamento do modelo aos dados. Avaliar o ajustamento é uma das etapas do processo de aplicação de MEE e, determina quão bem o modelo teórico é capaz de reproduzir a estrutura correlacional das variáveis manifestas na amostra sob estudo [5]. Com a aplicação de modelos de regressão não linear, ajustados sobre as estimativas calculadas sobre uma amostra reduzida de observações e, a extrapolação dos resultados para além dos dados com cálculo das chamadas 'bandas' de confiança, usando o método Delta, busca-se determinar um horizonte preditivo, que possibilitará uma avaliação do comportamento dos índices de ajuste do modelo para um número maior de observações, possibilitando a validação do modelo aos dados.

¹clodoaldo.figueredo@gmail.com

²jair.marques@utp.br

2 Fundamentação Teórica

2.1 Modelagem de Equações Estruturais

A MEE é capaz de testar relações causais entre variáveis. Essas relações são expressas por parâmetros que, conforme Chou e Bentler [3] são coeficientes de regressão, variâncias e covariâncias das variáveis independentes e indicam a magnitude do efeito que as variáveis apresentam umas sobre as outras. A resolução de modelos de equações estruturais possui etapas, divididas em grau de complexidade pré-estabelecidos, que devem ser obedecidas de forma criteriosa. Uma dessas etapas está no levantamento dos dados ou tamanho da amostra, e um dos pressupostos sobre os dados que deve ser validado é a presença de normalidade multivariada. Uma forma de verificação é feita mediante o uso de testes baseados nos coeficientes de assimetria e curtose, como o teste de Mardia [4], ou a aplicação do Teste Monte Carlo de Normalidade Multivariada baseado em Distâncias (TMCNMD) [2] que não possui restrições quanto ao tamanho da amostra.

A presença ou não de normalidade multivariada nos dados do modelo interfere diretamente na escolha do método de estimação dos parâmetros do modelo. Os métodos de estimação clássicos normalmente empregados em MEE para realizar os ajustes são o Mínimos Quadrados Generalizados (GLS – *Generalized Least Square*) e Máxima Verossimilhança (ML – *Maximum Likelihood*).

Se o pressuposto da normalidade multivariada não for verificado, os métodos de estimação que não exigem normalidade, como o método GLS, devem ser usados com parcimônia e com cuidado a certos fatores, como tamanho da amostra e curtose excessiva. Após a estimação dos parâmetros, outra etapa fundamental da análise de um modelo de MEE é a avaliação da qualidade do ajuste do modelo. Essa etapa tem como objetivo determinar o quão bem o modelo teórico é capaz de reproduzir a estrutura correlacional das variáveis manifestas na amostra sob estudo. A avaliação do ajuste do modelo em MEE foi inicialmente enquadrada no processo decisório dicotômico do teste de hipóteses. A decisão de aceitar ou rejeitar a hipótese de ajuste foi baseada na probabilidade (*p-value*) associada ao valor do teste χ^2 , que avalia a discrepância entre a matriz de covariância da amostra original e a matriz de covariância reproduzida com base no modelo [1].

Alguns índices de qualidade de ajustamento foram criados para substituir o índice qui-quadrado, destacando-se:

- 1) Os índices absolutos, relativos à proporção de variância e covariância explicadas após o processo de estimação, com destaque para o índice GFI (*Goodness of Fit Index*), definido por:

$$GFI = 1 - (\vartheta_{residual}/\vartheta_{total}) \quad (1)$$

onde: $\vartheta_{residual}$ é a variância residual na matriz de covariância e ϑ_{total} é a variância total na matriz de covariância.

- 2) Os índices relativos, que trabalham com a lógica de comparação do χ^2 de modelos independentes com o χ^2 do modelo em teste, com destaque para os índices NFI (*Normed Fit Index*), que avalia a porcentagem de incremento da qualidade do ajustamento do modelo hipotético ($\chi_{hypmodel}^2$) relativamente ao modelo de independência total ou modelo de base, e CFI (*Comparative Fit Index*), que compara o ajustamento do modelo em estudo (χ^2) com graus de liberdade (*gl*) com o ajustamento do modelo de base (χ_b^2) com graus de liberdade (*gl_b*), definidos respectivamente por:

$$NFI = 1 - (\chi_{hypmodel}^2/\chi_b^2) \quad (2)$$

$$CFI = 1 - (max(\chi^2 - gl, 0)/max(\chi_b^2 - gl_b, 0)) \quad (3)$$

- 3) Os índices de discrepância populacional que comparam o ajustamento do modelo obtido com os momentos amostrais (médias e variâncias amostrais) relativamente ao ajustamento do modelo

que se obteria com os momentos populacionais (médias e variâncias populacionais) [5]. Um índice amplamente utilizado é o RMSEA (*Root Mean Square Error of Approximation*), definido da forma:

$$RMSEA = \sqrt{F_0/gl} \quad (4)$$

onde $F_0 = \max(\chi^2 - gl/(n - 1), 0)$.

2.2 Regressão Não Linear

A análise de regressão é uma metodologia estatística que utiliza a relação entre duas ou mais variáveis quantitativas para que uma variável possa ser prevista a partir da outra ou de outras, servindo a três propósitos principais: descrição, controle e predição.

Modelos de regressão são usados para descrever o comportamento de uma variável aleatória (v.a.) Y como uma função de condicionantes x chamadas de variáveis explicativas, covariáveis ou estímulos, buscando encontrar e descrever padrões de homogeneidade dentre a heterogeneidade observada em um conjunto de observações de Y [10].

Entre os diversos exemplos de modelos não lineares, pode-se citar os modelos de Michaelis-Menten e Exponencial descritos abaixo respectivamente.

$$Y = (\beta_0 x / (\beta_1 + x)) + \varepsilon \quad (5)$$

$$Y = \beta_0(1 - e^{-\beta_1 x}) + \varepsilon \quad (6)$$

Uma das maneiras de se escolher as funções é a partir da observação do gráfico dos pontos tabelados (gráfico de dispersão) ou baseando-se em fundamentação teórica do experimento. Existem duas classes de métodos para aproximação de dados, e a distinção entre elas está em considerar, ou não, a existência de erros nos dados. Quando se leva em consideração possíveis erros introduzidos na obtenção dos dados, o método dos mínimos quadrados, que tem por objetivo minimizar uma função definida a partir dos resíduos, tem sido o mais usado para ajustamento dos dados ao modelo de regressão. Uma das formas de avaliar a qualidade do ajuste do modelo em uma regressão é através do coeficiente de determinação R^2 . Este coeficiente indica quanto o modelo foi capaz de explicar os dados coletados. Em previsões obtidas com regressão deve-se estar atento a duas fontes de erros de variância. Por causa desses possíveis erros, no ajuste por regressão não linear, pode-se determinar intervalos de confiança para os parâmetros determinados e, com isso também calcular intervalos de confiança e predição para os valores preditos pelo modelo.

Esses intervalos, chamados de “bandas de confiança” ou “faixas de confiança”, retratam os limites de confiança superior e inferior para todos os pontos em uma linha ajustada dentro do intervalo de dados, além de fornecer uma ideia sobre a variabilidade global do estimador. Confiar em estimativas pontuais pode ser perigoso, portanto se geram essas faixas de confiança para afastar-se dessas estimativas que podem ser duvidosas. Vale a pena ressaltar que existem diferenças entre um intervalo de confiança e um intervalo de previsão. Um intervalo de previsão se refere à incerteza em torno de um único valor e devem considerar tanto a incerteza em saber o valor da média da população, quanto a dispersão de dados, enquanto um intervalo de confiança é um intervalo estimado de um parâmetro populacional e são usados para indicar a confiabilidade de uma estimativa.

O método Delta é uma abordagem geral para calcular intervalos de confiança para funções de estimativas de máxima verossimilhança. O método usa uma função que é muito complexa para calcular analiticamente a variância, então cria uma aproximação linear dessa função e, em seguida, calcula a variância da função linear mais simples que pode ser usada para inferência de amostra grande [8].

3 Material e Metodologia

3.1 Material

Para o desenvolvimento dos trabalhos propostos, obtenção dos resultados, contemplando os objetivos propostos foram utilizados cinco modelos de equações estruturais, obtidos a partir dos trabalhos de Maroco [5] e Prado [6]

Modelo 1 - Modelo de análise fatorial de 1a. ordem com 17 variáveis observáveis, 3 fatores e 116 graus de liberdade.

Modelo 2 - Modelo de análise fatorial de 2a. ordem com 13 variáveis observáveis, 3 fatores de primeira ordem, 1 fator latente de segunda ordem e 51 graus de liberdade.

Modelo 3 - Modelo causal com variável latente com 9 variáveis observáveis, 3 fatores e 23 graus de liberdade.

Modelo 4 - Modelo de regressão linear múltipla com 13 variáveis observáveis, 4 fatores e 59 graus de liberdade.

Modelo 5 - Modelo causal com variável latente com 23 variáveis observáveis, 8 fatores e 222 graus de liberdade.

Os cálculos necessários e gráficos foram desenvolvidos no ambiente do software R (R Core Team, 2019) versão 3.4.4 utilizando algoritmos escritos, funções e “pacotes” próprios para a linguagem.

3.2 Metodologia

A metodologia aplicada para obtenção dos resultados esperados, contemplando os objetivos propostos para esse trabalho, foi delineada nas seguintes etapas:

- 1) Análise dos dados disponibilizados em relação à normalidade multivariada, com utilização dos testes de Mardia e teste TMCNMD. A presença, ou não, de normalidade multivariada determina o método de estimação dos parâmetros.
- 2) Cálculo do número mínimo de observações a ser considerada, através da taxa de convergência dos modelos utilizando-se os métodos de estimação GLS ou ML. Entende-se como problema de convergência quando o algoritmo não consegue chegar a uma solução que minimize o limite mínimo para a variação dos parâmetros. A menor quantidade de observações que apresentar taxa de convergência acima de 50 %, para todos os modelos, será considerada como quantidade mínima para cálculo das médias das estimativas.
- 3) Cálculo da média das estimativas dos índices de qualidade de ajustamento (RMSEA, GFI, NFI e CFI) obtida a partir de 100 simulações calculadas sobre amostras, de tamanho determinado na etapa anterior com incremento de 10 observações até o limite de 150 (valor escolhido conforme aplicação da fórmula de Westland [9], escolhida randomicamente entre as observações disponíveis. Essas médias serão utilizadas para o ajustamento de um modelo de regressão não linear.
- 4) Ajustamento de um modelo de regressão não linear sobre as médias das estimativas dos índices de qualidade de ajustamento antes calculados. Serão utilizados dois modelos, o modelo de Michaelis-Menten e o modelo exponencial de Stanford-Smith reparametrizado proposto por Zeviani *et. al.* [10] $\hat{y} = \hat{A} \cdot (1 - e^{(-x \cdot \ln 2 / \hat{B})}) + \hat{C}$ para que seu comportamento se compare ao modelo de Michaelis-Menten [10]. O ajuste que apresentar o melhor coeficiente de determinação R^2 será considerado para calcular as estimativas dos índices para um número maior de observações (Até o limite das observações originais do modelo). A extrapolação do modelo ajustado também será utilizada para cálculo da banda de confiança dos valores preditos e análise dos resultados, criando um horizonte preditivo, onde será comparado as estimativas previstas com as calculadas sobre os reais valores. Os limites para o intervalo de predição, determinado pelo método Delta, será usado como referência aos valores de corte de cada índice, permitindo avaliar o ajuste do modelo em um cenário com um maior número de observações.

4 Resultados e Discussões

Em relação a normalidade multivariada, observa-se que todos os dados dos cinco modelos analisados não apresentam normalidade multivariada, sendo que dois modelos apresentaram acentuada assimetria e curtose. A não normalidade dos dados não impede de se aplicar a técnica de estimação por Máxima Verossimilhança (ML), mas como alguns modelos apresentam anormalidade multivariada acentuada, a aplicação da técnica de estimação por mínimos quadrados generalizados (GLS) é mais apropriada. Os modelos estimados usando o método GLS apresentaram uma taxa de convergência, maior ou igual a 50 %, para valores a partir de 40 observações, como se pode observar na tabela 1. Desta forma adota-se o valor de 40 observações como mínimo, bem como o método GLS como padrão, nas estimativas dos parâmetros.

Tabela 1: Taxa de convergência dos modelos usando o método de estimação GLS (em porcentagem).

| | Tamanho da amostra | | | | | | | | | | | |
|----------|--------------------|------|------|------|------|------|------|------|------|------|------|------|
| | 20 | 30 | 40 | 50 | 60 | 70 | 80 | 90 | 100 | 110 | 120 | 130 |
| Modelo 1 | 41,4 | 47,6 | 50,2 | 59,5 | 71,4 | 81,8 | 90,2 | 94,5 | 96,4 | 98,8 | 99,8 | 99,7 |
| Modelo 2 | 48,5 | 52,8 | 63,8 | 72,8 | 89,4 | 91,4 | 95,5 | 95,3 | 97,3 | 98,8 | 98,3 | 99,4 |
| Modelo 3 | 86,1 | 96,8 | 99,6 | 99,8 | 99,8 | 100 | 100 | 100 | 100 | 100 | 100 | 100 |
| Modelo 4 | 51 | 64,8 | 76,1 | 78,4 | 84,4 | 85,8 | 86,2 | 87,9 | 90,6 | 92,1 | 91,7 | 94,1 |
| Modelo 5 | 0 | 57,8 | 58,1 | 59,2 | 64,8 | 70,4 | 74,3 | 77,8 | 82,7 | 84,3 | 87,6 | 90,6 |

Na avaliação da estabilidade dos índices de qualidade de ajustamento em relação ao número de observações, usando o método de estimação GLS, observou-se que as estimativas dos índices de qualidade apresentaram grande diferença de valores para pequenas quantidades de observações, diminuindo a medida que o número de observações é incrementado. Esse comportamento pode ser observado nas estimativas representadas nos gráficos da figura 1. Nos modelos com baixo grau de liberdade, as estimativas apresentaram grande oscilação em relação ao incremento da quantidade de observações. Os índices NFI e CFI apresentaram valores próximos de 1 e isso pode ocorrer quando o valor do qui-quadrado se aproxima do valor dos graus de liberdade, ou porque o modelo é saturado, ou indicar baixa potência (baixo poder) nesses tamanhos de amostra para detectar qualquer erro de especificação. Estudos de simulação tem mostrado que, principalmente os índices GFI e NFI, são sensíveis ao tamanho da amostra e que o índice GFI tem maior poder estatístico [7].

Na sequência ajustou-se um modelo de regressão não linear sobre as médias dos valores dos parâmetros obtidos nas 100 simulações realizadas para cada quantidade de observações. Todos os ajustes apresentaram um coeficiente de determinação acima de 0,8 o que pode ser considerado satisfatório. A extrapolação do modelo de regressão, feita para um número de observações maiores, com a garantia de que não haverá fortes mudanças no comportamento das estimativas, apresentou valor muito próximos dos parâmetros obtidos calculados com os dados originais, sendo a diferença dos valores menor que 0,1 o que não compromete as considerações a respeito do ajustamento do modelo de equações estruturais. O intervalo de confiança também apresentou limites muito próximos, o que nos faz concluir que a variância dos parâmetros, calculada para um número de amostras maiores é mínima. Os ajustamentos dos modelos de regressão e respectivas bandas de confiança para os índices considerados podem ser observados na Figura 1 e os valores extrapolados para o índice RMSEA, a diferença calculada para os parâmetros calculados sobre as observações originais e os limites para o intervalo de confiança podem ser observados na Tabela 2. O mesmo comportamento observado para o índice RMSEA também foi observado para os demais índices considerados (GFI, NFI e CFI).

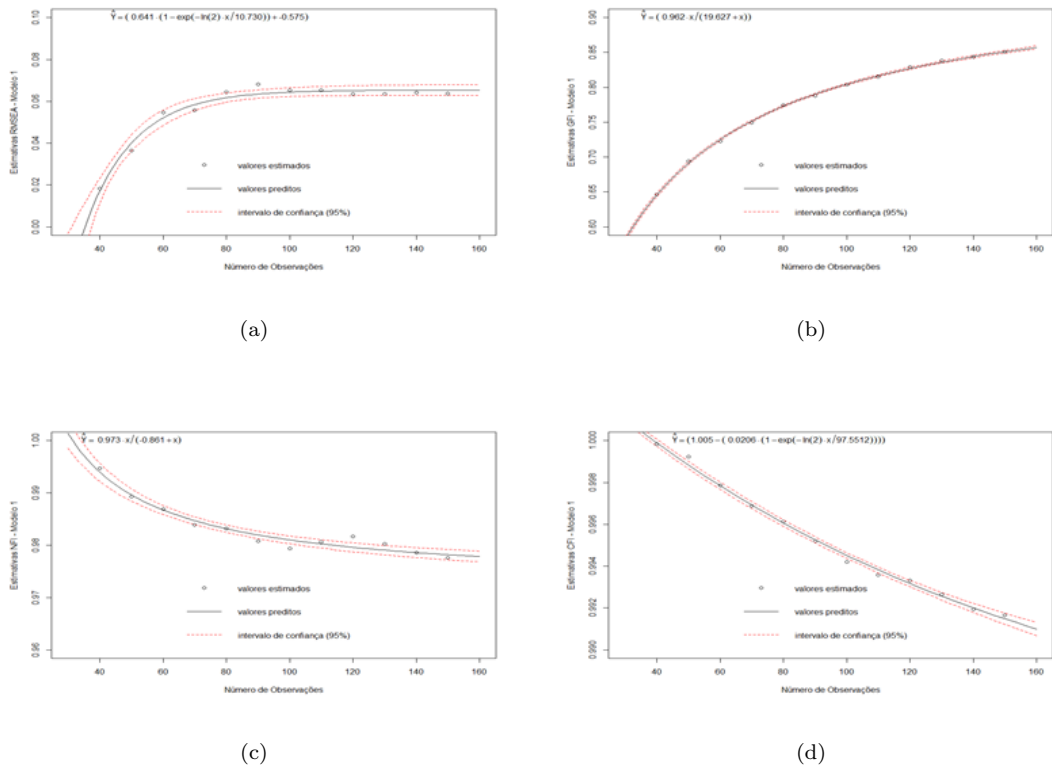


Figura 1: Ajustamento sobre as médias das estimativas dos índices de qualidade de ajustamento
 (a) Índice RMSEA - $\hat{y} = 0,641 \cdot (1 - \exp(-\ln(2) \cdot x/10,73)) - 0,575$ com $R^2 = 0,9721$
 (b) Índice GFI - $\hat{y} = 0,962 \cdot x/(19,627 + x)$ com $R^2 = 0,9991$
 (c) Índice NFI - $\hat{y} = 0,973 \cdot x/(-0,861 + x)$ com $R^2 = 0,9582$
 (d) Índice CFI - $\hat{y} = 1,005 - (0,0206 \cdot (1 - \exp(-\ln(2) \cdot x/97,5512)))$ com $R^2 = 0,9163$

Tabela 2: Estimativas para o Índice RMSEA, Valores Preditos e Intervalos de Predição

| Observações | Vlr. Est. | Vlr. Predito | Diferença | Lim. Inf. | Lim. Sup. |
|-------------|-----------|--------------|-----------|-----------|-----------|
| 260 | 0,062372 | 0,066 | -0,00363 | 0,058684 | 0,072132 |
| 270 | 0,062174 | 0,066 | -0,00383 | 0,058687 | 0,072143 |
| 280 | 0,062152 | 0,066 | -0,00385 | 0,058692 | 0,072151 |
| 290 | 0,063172 | 0,066 | -0,00283 | 0,058694 | 0,072159 |
| 300 | 0,063519 | 0,066 | -0,00248 | 0,058698 | 0,072167 |

Legenda Vlr. Est: Valor das estimativas em relação aos dados originais
 Vlr. Predito: Valor predito pelo modelo ajustado
 Diferença: Vlr. Est - Vlr. Predito
 Lim. Inf.: Limite inferior do intervalo de predição
 Lim. Sup.: Limite superior do intervalo de predição

5 Conclusões

O ajuste dos modelos de regressão não linear, sobre as estimativas de índices de qualidade de ajustamento, apresentou um bom coeficiente de determinação. O modelo exponencial reparametrizado apresentou melhor ajuste para todos os índices. Considerando um cenário com um número maior de observações, podemos concluir que, os valores extrapolados ficaram num nível acima do satisfatório para o ajustamento do modelo de equações estruturais aos dados. Também observou-se que a diferença dos valores extrapolados em relação as estimativas calculadas sobre os dados originais foi bem pequena, o que representa pequena margem de erro, bem como os limites do intervalo de predição bem próximos representa uma variância entre as estimativas próxima de zero. Podemos concluir que a aplicação de regressão não linear, sobre as estimativas dos índices de qualidade de ajustamento, em MEE apresenta-se com uma boa alternativa para validação do ajuste do modelo aos dados. As limitações da pesquisa realizada, no que diz respeito aos modelos utilizados e dados disponibilizados, deixam espaço para futuros trabalhos com outros tipos de modelos de equações estruturais, modelos de regressão não linear e análises sobre os erros de predição.

Referências

- [1] Bentler, P. M. *The interdependence of theory, methodology, and empirical data: Causal modeling as an approach to construct validation*, In Kandel D. B. (Ed.), *Longitudinal research on drug abuse: Empirical findings and methodological issues* Wiley: New York, 1978.
- [2] Biase, A. G., Ferreira, D. F. *Teste Computacional Intensivo Baseado na Distância de Mahalanobis para Normalidade Multivariada*, Rev. Bras. de Biometria, São Paulo, v. 30, n.1, 2012.
- [3] Chou, C.-P., Bentler, P. M. *Estimates and tests in structural equation modeling*, in Hoyle R. H. (Ed.) *Structural equation modeling: Concepts, issues, and applications*, Sage Publications: Thousand Oaks, 1995. pages 37-55. ISBN: 0-8039-5317-8
- [4] Hair Jr, J. F., Black, W. C., Babin, B. J., Anderson, R. E., Tatham, R. L. *Análise Multivariada de Dados, 6a edição* Bookman: Porto Alegre, 2009. ISBN 978-85-7780-402-3
- [5] Marôco, J. *Análise de Equações Estruturais: Fundamentos Teóricos, Software e Aplicações, 2a edição* Report Number: Lisboa, 2014. ISBN 978-989-96763-3-6
- [6] Prado, P. H. M. *A Avaliação do Relacionamento sob A Ótica do Cliente: Um estudo em Bancos de Varejo* 497 f. Tese EAESP – FGV, São Paulo, 2004
- [7] Tanguma J. Effects of Sample Size on the Power of Selected Fit Indices: A Graphical Approach, *Annual Meeting of the Southwest Educational Research Association*, New Orleans, 2001.
- [8] Xu, J., Long, J. S. *Using the Delta Method to Construct Confidence Intervals for Predicted Probabilities, Rates, and Discrete Changes*, 2005. Disponível em http://indiana.edu/~jlsoc/stata/ci_computations/spost_deltaci.pdf, acesso em 20 Jan. 2019.
- [9] Westland, J. C. Lower Bounds on Sample Size in Structural Equation Modeling, *Electronic Commerce Research and Applications*, v.9, n.6, pages 476-487, 2010. DOI: 10.1016/j.elerap.2010.07.003
- [10] Zeviani, P. J., Ribeiro, P. J., Bonat W. H. *Modelos de Regressão não Linear*, Laboratório de Estatística e Geoinformação, Departamento de Estatística, UFPR, 2013. Disponível em <http://leg.ufpr.br/~walmes/cursoR/mrn12013/master.pdf>, acesso em 10 Jul. 2019.