

USO DO MODELO DE EQUAÇÕES DE ESTIMAÇÕES GENERALIZADAS NA ANÁLISE DE DADOS LONGITUDINAIS

USE OF THE GENERALIZED ESTIMATING EQUATION MODEL IN LONGITUDINAL DATA ANALYSIS

Luciano Santos Pinto Guimarães, Vânia Naomi Hirakata

RESUMO

Este artigo tem como objetivo apresentar o método de Equações de Estimações Generalizadas (GEE) para a análise de dados longitudinais, considerada, em muitas ocasiões da área médica, melhor que a análise de variâncias para medidas repetidas. Para um melhor entendimento, apresentamos um exemplo no programa SPSS, versão 18.0.0, mostrando como esta pode ser realizada e interpretada.

Palavras-chave: Equações de Estimações Generalizadas; dados longitudinais; SPSS

ABSTRACT

The objective of the present manuscript was to describe the method of Generalized Estimating Equations (GEE) for longitudinal data analysis, which is considered to be better than the analysis of variance for repeated measures in many medical situations. For a better understanding, we present an example using the SPSS, version 18.0.0, and showing how GEE can be performed and interpreted.

Keywords: Generalized Estimating Equations; longitudinal data; SPSS

Na área da saúde são comuns os estudos em que se têm diversas medidas de uma mesma variável em um mesmo sujeito, seja em estudos com delineamento longitudinal, apenas observacional, seja ele com alguma intervenção entre as medidas. Quando desejamos avaliar a alteração de uma variável ao longo do tempo em apenas dois momentos, a opção de análise mais apropriada para variáveis com distribuição normal seria o teste t para amostras emparelhadas. No entanto, quando a mensuração for feita em mais de dois momentos, ou quando se deseja comparar o comportamento da variável ao longo dos diferentes

momentos entre dois grupos, ou ainda, quando existir alguma variável cujo efeito se quer ajustar a esta comparação, seria necessário que se realizasse uma análise de variâncias para medidas repetidas (ANOVA-MR), considerada um GLM (modelo linear generalizado).

Contudo, por uma série de pré-requisitos desta análise, muitas vezes é inviável que ela possa ser aplicada na prática. Uma das exigências do modelo de ANOVA-MR é a de que a variável em estudo seja quantitativa e tenha distribuição normal multivariada. Ou seja, a variável dependente em cada subgrupo, para cada momento, necessita ter distribuição

Revista HCPA. 2012;32(4):503-511

Unidade de Bioestatística, Grupo de Pesquisa e Pós-graduação, Hospital de Clínicas de Porto Alegre.

Contato:

Luciano Santos Pinto Guimarães
lsguimaraes@hcpa.ufrgs.br
Porto Alegre, RS, Brasil.

normal, além de que as variâncias sejam iguais em todos os momentos e a correlação seja constante entre quaisquer dois momentos (esfericidade) (1). Essa situação é praticamente impossível de ser alcançada na área médica, à medida que os pacientes alteram sua situação clínica ao longo do tratamento(2). Também é necessário que os dados sejam balanceados, ou seja, não exista nenhuma falha na coleta de dados de nenhum indivíduo. Isso quer dizer que, se em um estudo em que seriam necessárias cinco medidas de seguimento, mesmo que um sujeito não tenha sido avaliado apenas uma vez, todas as demais avaliações que ele tenha feito seriam desconsideradas da análise, a não ser que seja feita alguma forma de imputação dos dados faltantes (3,4). Além disso, são possíveis apenas fatores e covariáveis que não possam variar ao longo do tempo, e se for importante considerar a variação desta no comportamento da variável em estudo, na ANOVA-MR não é possível fazê-la na sua forma original. O fato dos intervalos de tempo entre as diferentes mensurações serem considerados constantes também é uma importante limitação, uma vez que a variável tempo é sempre considerada categórica nessa análise.

Devido a estas limitações da ANOVA-MR e ao surgimento de novas técnicas ao longo dos últimos anos, juntamente com a implementação delas nos *softwares* estatísticos, têm se visto cada vez mais estudos mostrando as vantagens de métodos como o de Modelos Lineares Mistos (*Mixed Linear Models* - MLM) (2,5-7) e das Equações de Estimativas Generalizadas (*Generalized Estimating Equations* - GEE) (2,7,8) sobre a ANOVA-MR em estudos longitudinais. Além disso, os MLM e GEE também têm sido preferidos quando se trata de ensaios clínicos randomizados (6,9,10).

O objetivo deste artigo é apresentar a análise e sua execução no programa SPSS, por meio da utilização de uma variável dependente contínua. São apresentadas interpretações dos resultados e uma possível tabela de exposição dos dados.

ANÁLISE DE GEE

A análise de GEE baseia-se na metodologia dos Modelos Lineares Generalizados (11), e inicialmente foi proposta por Liang e Zeger (12), para o contexto de dados longitudinais. Usar essa análise requer especificar: a) função de ligação, b) distribuição da variável dependente e c) a estrutura

de correlação (13). Os parâmetros de regressão (β 's) são estimados pelo estimador de máxima verossimilhança. Os cálculos das variâncias são realizados usando uma função de ligação, que transforma a variável dependente em uma equação de estimativas de parâmetros na forma de um modelo aditivo ($y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots$). Os resultados a partir dessas equações dão início a um processo iterativo usando procedimentos de mínimos quadrados que envolvem a minimização extensa da alteração do parâmetro estimado a partir de um modelo de regressão perfeitamente ajustado. À medida que o tamanho destas alterações tende a zero as estimativas dos parâmetros (β 's e erros-padrão) estabilizam. As especificações corretas tornam as estimativas mais eficientes.

No software SPSS, versão 18.0.0, a realização da análise pode ser encontrada nos procedimentos: Analyze >> Generalized Linear Models >> Generalized Estimating Equations. A análise de GEE é relativamente simples de ser realizada. Algumas escolhas anteriormente devem ser feitas, como por exemplo:

*A escolha da variável intrassujeito: para a estruturação do banco, apresentado a seguir, é necessário definir a variável que identifica os sujeitos, já que o mesmo sujeito possui mais de uma observação. É possível que haja mais de uma variável interferindo na medição, por exemplo, em um modelo hierárquico, onde pacientes de estados (UF) distintos podem ser acompanhados em tratamentos diferentes. Assim, a estrutura a ser considerada é o sujeito seguido do estado (UF) que ele pertence.

*A distribuição da variável dependente: o programa SPSS tem como default o tipo de resposta linear que utiliza uma função de ligação identidade (assumindo desta forma que a distribuição da variável seja normal). Já para uma resposta com distribuição gamma, a função de ligação padrão é a logarítmica e para uma resposta binária a função é a logito (13).

*A escolha do estimador da matriz de covariância: o SPSS possibilita a escolha de dois estimadores, o baseado no modelo e o robusto, sendo o último default do programa.

*A escolha da matriz de correlação trabalho: representa a dependência intraindividual. Uma melhor estrutura deve ser buscada usando como critério o menor *Quasi Likelihood Under Independence Model Criterion* (QIC).

Exemplo

O banco para o exemplo foi mascarado sendo adaptado de um projeto real. A ansiedade é um transtorno bastante estudado. Uma pesquisa foi realizada, sendo aplicado um questionário que media um escore de ansiedade em pacientes pré-diagnosticados. Dois tratamentos distintos foram usados, sendo estudado também o efeito combinado desses tratamentos para testar a redução desse transtorno. O estudo teve seguimento de um ano sendo realizadas quatro medições: T1: basal, T2: 2º mês, T3: 4º mês e T4: 12º mês. Assim foram testados os efeitos principais das variáveis (grupos e tempo) e o efeito da interação (grupo*tempo). A figura 1 mostra as médias e os erros-padrão da média de cada grupo em cada tempo.

MONTAGEM DO BANCO

Diferente da construção dos bancos de dados utilizados em testes mais simples, a análise de GEE exige um formato chamado de dados empilhados. O banco nesse exemplo foi composto de quatro colunas, sendo uma para a identificação dos sujeitos e outra especificando a qual grupo este pertence. O resultado é digitado em uma única variável sendo essa auxiliada por outra indicando o tempo. Cada sujeito terá no máximo quatro linhas, já que a perda de alguma informação não acarretará na perda do sujeito, como acontece na análise usando ANOVA-MR. Uma imagem de construção do banco de dados pode ser visto na Figura 2.

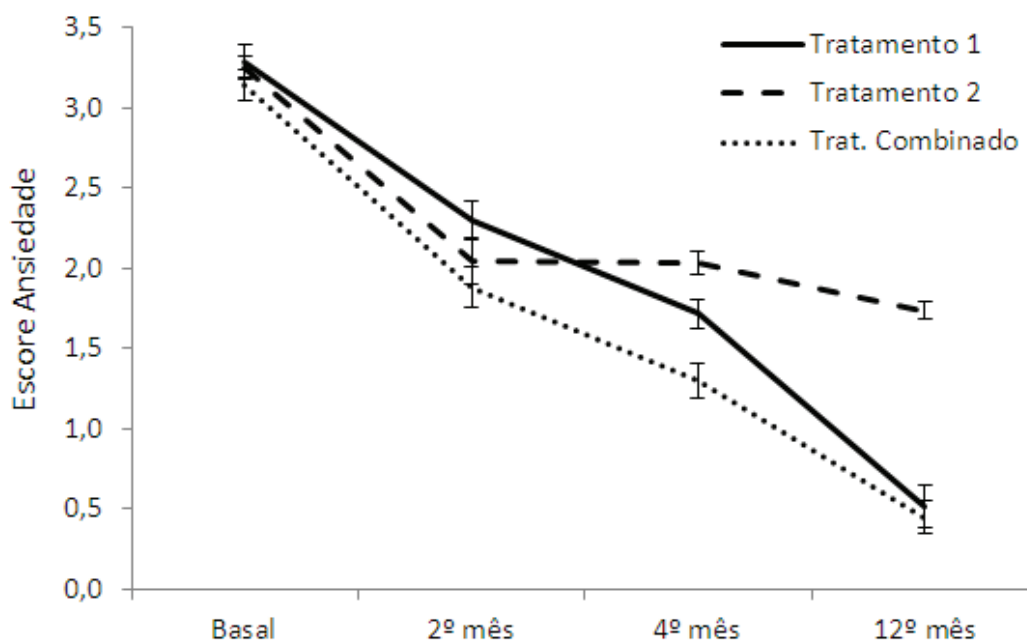


Figura 1: Gráfico de linha das médias do escore de ansiedade de cada grupo discriminado pelos tempos avaliados.

Numero	Sujeito	Grupo	Tempo	Ansiedade
1,00	1	Tratamento1	Basal	2,37
2,00	2	Tratamento1	Basal	2,62
3,00	2	Tratamento1	2 mes	1,87
4,00	2	Tratamento1	4 mes	1,62
5,00	3	Tratamento1	Basal	3,49
6,00	3	Tratamento1	2 mes	2,24
7,00	3	Tratamento1	4 mes	1,75
8,00	3	Tratamento1	12 mes	,62
...				
428,00	92	Tratamento2	Basal	3,24
429,00	92	Tratamento2	2 mes	1,99
430,00	92	Tratamento2	4 mes	1,87
431,00	92	Tratamento2	12 mes	1,75
433,00	93	Tratamento2	Basal	2,86
434,00	93	Tratamento2	2 mes	1,62
...				
850,00	183	Trat Combinado	Basal	3,24
851,00	183	Trat Combinado	2 mes	1,99
852,00	183	Trat Combinado	4 mes	1,12
854,00	183	Trat Combinado	12 mes	,88
855,00	184	Trat Combinado	Basal	2,86
856,00	184	Trat Combinado	2 mes	1,62
...				

Obs.: O banco foi construído a partir de códigos, o que se vê, na figura 2, são os rótulos. Para a variável Grupo usou-se: 1 – Tratamento 1; 2 – Tratamento 2; 3 – Trat Combinado. Para a variável Tempo usou-se: 1 – Basal; 2 – 2º mês; 3 – 4º mês; 4 – 12º mês.

Figura 2: Modelo de banco de dados utilizado para a análise de GEE no software SPSS v.18.0.0.

No material suplementar (<http://seer.ufrgs.br/hcpa/article/view/36971>) pode-se visualizar a construção da análise no *software* SPSS. A análise pode ser realizada pela sintaxe encontrada no Quadro 1. Uma introdução para o uso dessa ferramenta pode ser encontrada no artigo de Lindenau & Guimarães (2012) (14). Um complemento na programação deve ser feito na parte de comparação das médias

(*EMMEANS TABLES*), quando a interação for significativa, já que a disponibilizada pelo SPSS é de difícil compreensão. Esse complemento aparece em destaque no Quadro 1. Foi usado para esse exemplo o estimador robusto, uma matriz de correlação trabalho não estruturada (*unstructured*) e uma distribuição normal com função de ligação identidade.

Quadro 1: Sintaxe da análise de GEE, com modificação, no software SPSS v.18.0.0.

```
* Generalized Estimating Equations.
GENLIN Ansiedade BY Grupo Tempo (ORDER=ASCENDING)
/MODEL Grupo Tempo Grupo*Tempo INTERCEPT=YES
DISTRIBUTION=NORMAL LINK=IDENTITY /CRITERIA SCALE=MLE PCONVERGE=1E-006(ABSOLUTE) SINGULAR=1E-012
ANALYSISTYPE=3(WALD) CILEVEL=95 LIKELIHOOD=FULL
/EMMEANS TABLES=Grupo SCALE=ORIGINAL COMPARE=Grupo CONTRAST=PAIRWISE PADJUST=BONFERRONI
/EMMEANS TABLES=Tempo SCALE=ORIGINAL COMPARE=Tempo CONTRAST=PAIRWISE PADJUST=BONFERRONI
/EMMEANS TABLES=Grupo*Tempo SCALE=ORIGINAL COMPARE=Grupo CONTRAST=PAIRWISE PADJUST=BONFERRONI
/EMMEANS TABLES=Grupo*Tempo SCALE=ORIGINAL COMPARE=Tempo CONTRAST=PAIRWISE PADJUST=BONFERRONI
/REPEATED SUBJECT=sujeito WITHINSUBJECT=Tempo SORT=YES CORRTYPE=UNSTRUCTURED ADJUSTCORR=YES COVB=ROBUST
MAXITERATIONS=100 PCONVERGE=1e-006(ABSOLUTE) UPDATECORR=1 /MISSING CLASSMISSING=EXCLUDE
/PRINT CPS DESCRIPTIVES MODELINFO FIT SUMMARY SOLUTION.
```

RESULTADOS DO EXEMPLO

As primeiras tabelas do output do SPSS mostram as informações utilizadas na análise: a variável dependente, a distribuição e a função de ligação, a matriz de correlação, o número total de sujeitos e suas discriminações pelas variáveis independentes (fatores). A tabela nomeada de *Test of Model Effects* mostra os resultados dos efeitos principais e da interação (maiores explicações sobre efeitos principais e efeitos de interação podem ser encontrados no Anexo 1).

Pelo teste qui-quadrado de Wald é obtido o valor p (nomeado pelo SPSS de Sig). Nesse exemplo a interação (grupo*tempo) foi significativa ($p < 0,001$) conforme pode ser visualizado na tabela trazida do SPSS (tabela 1). Esse resultado mostra que ao longo do tempo há uma mudança no comportamento médio do escore de ansiedade entre pelo menos dois grupos. Cabe agora, através de um teste complementar, avaliar onde ocorre essa mudança. O teste *post-hoc* realizado foi o de Bonferroni. Já que a interação foi significativa, os efeitos principais não devem ser analisados.

Tabela 1: Resultados do teste dos efeitos principais e da interação pelo qui-quadrado de Wald dado pelo SPSS.

Source	Type III		
	Wald Chi-Square	df	Sig.
(Intercept)	6258,948	1	0,000
Grupo	92,652	2	0,000
Tempo	6966,089	3	0,000
Grupo * Tempo	490,846	6	0,000

Dependent Variable: Ansiedade

Model: (Intercept), Grupo, Tempo, Grupo * Tempo

Os resultados das médias do escore de ansiedade, erros-padrão e intervalos de confiança de cada grupo em cada tempo são encontrados na tabela *Estimated Marginal Means* na parte da interação (tabela 2).

Tabela 2: Estimativas da média de ansiedade discriminadas por grupo e tempo.

Grupo	Tempo	Mean	Std. Error	95% Wald Confidence Interval	
				Lower	Upper
Tratamento1	Basal	3,29	0,05	3,19	3,39
	2º mês	2,30	0,06	2,18	2,42
	4º mês	1,72	0,04	1,63	1,80
	12º mês	0,51	0,07	0,38	0,65
Tratamento2	Basal	3,25	0,04	3,18	3,32
	2º mês	2,05	0,07	1,91	2,19
	4º mês	2,03	0,04	1,96	2,10
	12º mês	1,73	0,03	1,68	1,79
Trat Combinado	Basal	3,14	0,05	3,05	3,24
	2º mês	1,88	0,07	1,75	2,01
	4º mês	1,30	0,06	1,19	1,41
	12º mês	0,45	0,05	0,34	0,55

Com a modificação realizada na sintaxe antes de executar o comando altera-se a maneira de apresentação dos dados na comparação par a par do efeito da interação. Na Tabela 3 destacam-se as informações mais relevantes do output do SPSS para essas comparações, já que há muitas informações disponíveis nos resultados originais. Nessa forma de apresentação os resultados são organizados de maneira mais direta, facilitando a interpretação do pesquisador.

Duas tabelas foram criadas. A Tabela 3.a fixa o tempo e compara os grupos de tratamentos, diferente da Tabela 3.b, que fixa os grupos e compara os tempos. Interpretando a Tabela 3.a: os três grupos de tratamentos não possuem diferenças em seus escores de ansiedade no tempo basal ($p > 0,05$), podendo confirmar esses valores pela Tabela 2. No segundo mês o tratamento 1 é menos eficaz que o tratamento 2 ($p = 0,019$) e o

tratamento combinado ($p < 0,001$); no quarto mês há uma diferença entre todos os escores médios, sendo que o tratamento combinado ($p < 0,001$) é mais eficaz, seguido do tratamento 1 ($p < 0,001$). Após um ano de seguimento o tratamento 2 é diferente dos demais, sendo menos eficaz (tratamento 1: $p < 0,001$; tratamento combinado: $p < 0,001$), ou seja, produz médias de escores de ansiedade mais altos.

A Tabela 3.b pode ser interpretada do seguinte modo: os quatro tempos possuem escores médios de ansiedade diferentes, tanto no tratamento 1 quanto no tratamento combinado, sempre com escores médios menores na decorrência do tempo ($p < 0,001$); já no tratamento 2, o segundo e o quarto mês não diferem em médias no escore de ansiedade ($p > 0,999$), diminuindo significativamente esse escore no décimo segundo mês ($p < 0,001$).

Tabela 3: Resultados parciais das comparações múltiplas do efeito da interação (grupo*tempo) na variável ansiedade fixando um dos fatores.

3a. Pairwise Comparisons – fixando Tempo

Tempo	(I) Grupo	(J) Grupo	Bonferroni Sig.
Basal	Trat. 1	Trat. 2	>0,999
		Trat. Comb.	0,115
	Trat. 2	Trat. Comb.	0,271
2º mês	Trat. 1	Trat. 2	0,019
		Trat. Comb.	<0,001
	Trat. 2	Trat. Comb.	0,271
4º mês	Trat. 1	Trat. 2	<0,001
		Trat. Comb.	<0,001
	Trat. 2	Trat. Comb.	<0,001
12º mês	Trat. 1	Trat. 2	<0,001
		Trat. Comb.	>0,999
	Trat. 2	Trat. Comb.	<0,001

3b. Pairwise Comparisons – fixando Grupo

Grupo	(I) Tempo	(J) Tempo	Bonferroni Sig.
Trat. 1	Basal	2º mês	<0,001
		4º mês	<0,001
		12º mês	<0,001
	2º mês	4º mês	<0,001
		12º mês	<0,001
		4º mês	12º mês
Trat. 2	Basal	2º mês	<0,001
		4º mês	<0,001
		12º mês	<0,001
	2º mês	4º mês	0,999
		12º mês	<0,001
		4º mês	12º mês
Trat. Comb.	Basal	2º mês	<0,001
		4º mês	<0,001
		12º mês	<0,001
	2º mês	4º mês	<0,001
		12º mês	<0,001
		4º mês	12º mês

Uma representação dessas tabelas, para publicação, pode ser feita usando o sistema de letras. Letras minúsculas distintas mostram que as médias dos escores de ansiedade em cada tratamento

diferem dentro de um determinado tempo. Letras maiúsculas distintas mostram a evolução de um determinado grupo conforme o passar do tempo. Esse exemplo é encontrado na Tabela 4.

Tabela 4: Representação da tabela com os resultados finais e a representação do sistema de letras das comparações múltiplas.

Grupo	Tempo			
	Basal	2º mês	4º mês	12º mês
Tratamento1	3,29 (0,05)aA	2,3 (0,06)aB	1,72 (0,04)aD	0,51 (0,07)aD
Tratamento2	3,25 (0,04)aA	2,05 (0,07)bB	2,03 (0,04)bB	1,73 (0,03)bC
Trat. Combinado	3,14 (0,05)aA	1,88 (0,07)bB	1,30 (0,06)cC	0,45 (0,05)aD

Podemos concluir nesse exemplo que pela análise de GEE os tratamentos com menores escores médios foram encontrados, no acompanhamento de 12 meses, naqueles com a composição do tratamento 1 (tratamento 1 e o tratamento combinado). O tratamento 2, a partir do quarto mês, produz as maiores médias, sendo diferente significativamente dos demais tratamentos, não sendo recomendado o seu uso.

CONCLUSÃO

O método de análise de dados longitudinais pelo GEE possibilita a análise de desfechos contínuos, mesmo quando a variável não apresenta

distribuição normal ou esfericidade (12). Além disso, mesmo quando há perda de informações de algum indivíduo da amostra, é possível a inclusão de todos os indivíduos, situação que pode evitar algum tipo de viés de seleção (4). Outra vantagem do GEE, demonstrada por Ma et al.(2), em relação à ANOVA-MR é a necessidade de um tamanho menor de amostra para conseguir evidenciar o mesmo tamanho de efeito, com um poder de 80%. A possibilidade de análise de dados categóricos ou ordinais também deve ser considerada, embora nosso estudo não tenha tido este enfoque (15). Por esses motivos, recomenda-se o método das equações de estimativas generalizadas (GEE) para a análise de dados longitudinais.

REFERÊNCIAS

- Garson GD. Univariate GLM, ANOVA, and ANCOVA: Statistical Associates; 2012.
- Ma Y, Mazumdar M, Memtsoudis SG. Beyond Repeated-Measures Analysis of Variance: Advanced Statistical Methods for the Analysis of Longitudinal Data in Anesthesia Research. *Regional Anesthesia and Pain Medicine*. 2012;37(1):99-105. 10.1097/AAP.0b013e31823ebc74.
- Grace-martin K. Approaches to Repeated Measures Data: Repeated Measures ANOVA, Marginal, and Mixed Models. Available from: <http://www.theanalysisfactor.com/repeated-measures-approaches/>.
- Liu M, Wei L, Zhang J. Review of guidelines and literature for handling missing data in longitudinal clinical trials with a case study. *Pharmaceutical Statistics*. 2006;5(1):7-18.
- Krueger C, Tian L. A Comparison of the General Linear Mixed Model and Repeated Measures ANOVA Using a Dataset with Multiple Missing Data Points. *Biological Research For Nursing*. 2004;6(2):151-7.
- Gueorguieva R KJH. Move over anova: Progress in analyzing repeated-measures data and its reflection in papers published in the archives of general psychiatry. *Archives of General Psychiatry*. 2004;61(3):310-7.
- Edwards LJ. Modern statistical techniques for the analysis of longitudinal data in biomedical research. *Pediatr Pulmonol*. 2000;30(4):330-44.
- Stiger TR, Kosinski AS, Barnhart HX, Kleinbaum DG. Anova for repeated ordinal data with small sample size? a comparison of anova, manova, wls and gee methods by simulation. *Communications in Statistics - Simulation and Computation*. 1998;27(2):357-75.
- Albert PS. Longitudinal data analysis

(repeated measures) in clinical trials. Stat Med. 1999;18(13):1707-32.

10. Twisk JR, Vente W. The analysis of randomised controlled trial data with more than one follow-up measurement. A comparison between different approaches. European Journal of Epidemiology. 2008;23(10):655-60.

11. McCullagh P, Nelder JA. Generalized Linear Models: Chapman & Hall; 1983.

12. Liang K-Y, Zeger SL. Longitudinal data analysis using generalized linear models. Biometrika. 1986;73(1):13-22.

13. Ballinger GA. Using Generalized Estimating Equations for Longitudinal Data Analysis 2013. Available from: <http://orm.sagepub.com/content/7/2/127.refs>.

14. Lindenau JD-R, Guimarães LSP. Calculando o tamanho de efeito no SPSS. Revista HCPA. 2012;32(3):363-81

15. Zeger SL, Liang KY. Longitudinal data analysis for discrete and continuous outcomes. Biometrics. 1986;42(1):121-30.

Recebido: 27/12/2012
Aceito: 10/01/2013

ANEXO 1

Efeito Principal x Efeito de Interação

Um "efeito principal" é o efeito direto de uma variável independente sobre a variável dependente.

Um "efeito de interação" é o efeito conjunto de duas ou mais variáveis independentes sobre a variável dependente.

Um exemplo didático pode ser realizado usando as variáveis, grupo e tempo, para explicar o comportamento da variável resposta ansiedade.

Interpretação

Na Figura A1 não vemos efeito dos fatores, grupo e tempo, na variável escore da ansiedade. Já em A2 há um efeito do grupo, mostrando que

o Tempo 2 possui um escore maior que os demais tempos e separadamente certificamos que os grupos possuem médias diferentes, sendo que o Grupo A é sempre maior que o Grupo B. No Gráfico A3 temos o efeito da variável grupo não havendo efeito no tempo, diferente do A4 que há mostra efeito do tempo e não do grupo. Nos dois últimos gráficos observamos a dependência de uma variável em relação à outra. No Gráfico A5 há divergência dos grupos em relação ao terceiro tempo e no A6 há um cruzamento entre elas observando que no Tempo 1 o Grupo B possui médias de escores de ansiedade maiores, se iguala em média no tempo 2 e reduz os valores no Tempo 3.

Anexo 1 : Análise gráfica com dois fatores para explicação de efeito principal e efeito de interação

