# ANÁLISE E MODELAGEM DE DADOS LONGITUDINAIS NO R

**Leandro Vitral Andraos**[1,2], **Marcel de Toledo Vieira**[1,2]

**Resumo:** *Neste trabalho serão estudadas técnicas de modelagem para dados longitudinais no software R. O interesse do estudo pode ser na trajetória de vida dos indivíduos incluídos na amostra, o que permite um estudo das relações entre as variáveis observadas. Como as observações para o mesmo indivíduo não são, em geral, independentes, temos de aplicar técnicas estatísticas especiais que levem em consideração o fato das observações repetidas serem correlacionadas. Primeiramente faremos uma breve explicação dos modelos e técnicas a serem utilizadas. Com funções desenvolvidas em nosso estudo e outras já existentes será possível fazermos uma análise exploratória e ajustes do modelo de efeitos aleatórios e do modelo marginal aos dados longitudinais. Para tais aplicações será utilizado um banco de dados real e os resultados dos modelos serão analisados.*

**Palavras-chave:** *dados longitudinais; modelos marginais, modelos de efeitos aleatórios; BHPS.*

## 1 Introdução

Existem diversas abordagens para a modelagem de dados provenientes de estudos com medidas repetidas. É fundamental levar em conta a estrutura de correlação resultante da estrutura longitudinal. Uma das formas de se trabalhar com esta característica é através de modelos lineares mistos, também chamados de modelos efeitos aleatórios. O presente trabalho tem como principal objetivo a compreensão e descrição do uso do software R para a modelagem de dados longitudinais. Sendo assim, na seção 2 será apresentada uma breve descrição de métodos de análise exploratória e de modelagem para dados longitudinais. Na seção 3 serão descritas funções disponíveis no R e outras desenvolvidas pelos autores, enquanto que na seção 4 serão apresentados resultados de uma aplicação com dados reais. Na seção 5, será feita uma discussão final.

## 2 Análise de Dados Longitudinais

Estudos longitudinais são levantamentos que visam analisar as variações nas características dos mesmos elementos ao longo de um período de tempo. Ou seja, os dados estudados são coletados em pelo menos dois pontos no tempo. Assim, não há, em geral, independência entre as observações do mesmo indivíduo. Nesta seção discutiremos de forma breve tanto técnicas exploratórias (Sub-seção 2.1), quanto de modelagem (Sub-seções 2.2 e 2.3).

[1]ICE- Universidade Federal de Juiz de Fora,
andraos@ice.ufjf.br, marcel.vieira@ice.ufjf.br

[2]Os autores agradecem pelo apoio financeiro recebido da FAPEMIG ao projeto e a bolsa de Iniciação Científica BIC/UFJF.

### 2.1 Análise Exploratória de Dados (AED) Longitudinais

A AED tem o objetivo básico de sintetizar uma série de valores, que podem ser de mesma natureza, permitindo dessa forma que se tenha uma visão global da variação desses valores. Sendo assim, visa-se maximizar a extração de informações na sua estrutura, organizando e descrevendo os dados. No contexto longitudinal, a AED inclui a produção de medidas descritivas e de gráficos, tais como gráficos de perfis de respostas individuais ao longo do tempo e de matrizes de diagramas de dispersão.

### 2.2 Modelo Marginal

Os modelos marginais são adotados com freqüência para a análise de dados longitudinais, onde a variável dependente é modelada em função tanto de variáveis cujos valores podem mudar com o tempo quanto daquelas cujos valores são mantidos fixos. Neste tipo de modelo a média marginal é modelada de forma semelhante ao que acontece com os modelos de regressão para dados transversais, porém levando em consideração a correlação nas medidas repedidas realizadas ao longo do tempo e modelando separadamente a média e a covariância (Diggle et al., 2002; Vieira, 2009). A estimação dos coeficientes tem como base um processo interativo através de equações de estimação generalizadas (Liang e Zeger, 1986).

### 2.3 Modelo de Efeitos Aleatórios

Uma abordagem alternativa aos modelos marginais são os modelos de efeitos aleatórios (ou modelos em múltiplos níveis), que trabalham com o pressuposto de que a correlação entre as medidas repetidas é causada pelo fato de que os coeficientes do modelo são aleatórios e, por isso, variam entre indivíduos. Esta classe de modelos é apropriada especialmente para situações em que o interesse maior é a produção de inferências sobre indivíduos ao invés de inferências agregadas populacionais. Maiores informações sobre os modelos de efeitos aleatórios podem ser encontradas em Goldstein (1995), Hand e Crowder (1996) e Vieira e Skinner (2008), por exemplo.

## 3 Software R

O R é um programa estatístico computacional disponível gratuitamente através da internet sob a General Public License. Muitos pacotes são disponíveis através da família CRAN de sítios na Internet cobrindo uma ampla variedade de estatísticas modernas. (Torgo, 2006).

### 3.1 Análise Exploratória de Dados (AED)

Na análise exploratória dos dados utilizou-se funções fornecidas pelo R e outras desenvolvidas ao longo da condução deste trabalho. A seguir apresentamos uma rotina (já abordando a aplicação motivadora da seção 4), que tem como objetivo analisar a evolução longitudinal dos valores de uma variável de interesse para um número qualquer de casos. Nesta aplicação, a variável de interesse é um escore de atitudes em relação ao papel do gênero, e gênero, escolaridade, faixa etária e atividade econômica são as covariáveis.

```
n=" Numero de individuos selecionados"
a<-data[,1]; aa<-sample(a,n)
worse<-matrix(rep(0,4*length(aa)),4,n)
for(j in 1:n){k=9
for(i in 1:4) {k=k+ 1
worse[i,j]<-(data[data$pid==aa[j],k])}}
worse<-as.numeric(worse)
Ano<- rep(c(91,93,95,97),length(aa))
plot(Ano,worse,ylim=c(0,30),pch=19,main=" Gráfico de Escore para todos os indivíduos nas 4 ocasiões")
k=0
for(i in 1:length(aa)){lines(Ano[(i+ k):(i+ 3+ k)],worse[(i+ k):(i+ 3+ k)],col="blue3")k=3*(i)}
```

Figura 1: Rotina para a produção de gráficos de perfis de resposta.

## 3.2 Modelo Marginal

O pacote Geepack, através da função geeglm, implementa o modelo marginal para dados longitudinais. Este pacote permite a escolha de diferentes matrizes de trabalho para a estrutura de correlação temporal intra-indivíduo, incluindo "independence", "exchangeable", "ar1", "unstructured" e "userdefined". A seguir apresentamos um exemplo de aplicação desta função no contexto de nossa aplicação motivadora. Outras informações em Højsgaard, Halekoh e Yan (2005).

```
m<-geeglm(score~asex+ aagecat+ aeduc+ amumwk+ ecact+ time,data=data2,id=pid,family=gaussian,corstr="exchangeable")
```

Figura 2: Rotina com exemplo de aplicação da função geeglm.

## 3.3 Modelo de Efeitos Aleatórios

Para a aplicação dos modelos de efeitos aleatórios é necessária a utilização das funções: groupedData e lme. Desta forma, a seguir apresentamos o exemplo de uma rotina, que utiliza a função groupedData e que tem como objetivo modificar a estrutura do banco de dados de forma a permitir o ajuste deste modelo.

```
groupedData( score~  time | pid,data = as.data.frame( data2 ),  FUN = mean, outer = ~ asex)
```

Figura 3: Rotina com exemplo de aplicação da função groupedData.

Depois de modificada a estrutura do banco de dados o modelo de efeitos aleatórios pode ser ajustado através da função lme, conforme o exemplo abaixo.

```
model1<- lme(score~asex + aagecat + aeduc + amumwk + ecact+ time, data=Orth, random= ~ time, method="ML")
```

Figura 4: Rotina com exemplo de aplicação da função lme.

# 4 Aplicação Motivadora

Para aplicação dos métodos estudados e a investigação de métodos de análise de dados longitudinais, buscando compreender as potencialidades dos mesmos, utilizamos um banco de dados obtidos pela *British Household Panel Survey* (BHPS).

## 4.1 BHPS

A BHPS é uma pesquisa longitudinal que teve uma amostra de indivíduos selecionada em 1991 por amostragem complexa em dois estágios, com a áreas geográficas sendo consideradas conglomerados. Os dados são recolhidos em ocasiões anuais. Nossas análises são baseadas em uma amostra de mulheres com idade entre 16-39 anos. A variável dependente é um escore de atitudes em relação ao papel do gênero (Vieira, 2009). As covariáveis incluem faixa etária, atividade econômica e qualificações educacionais. Os dados são analisados apenas para quatro ocasiões da pesquisa, 1991, 1993, 1995 e 1997.

## 4.2 Resultados

Utilizando nosso comando da seção 3.1 chegamos ao gráfico de perfil de pontuações de atitude do gênero. Serão utilizados 50 indivíduos. Com o gráfico poderemos perceber como foi a variação da pontuação para cada indivíduo.

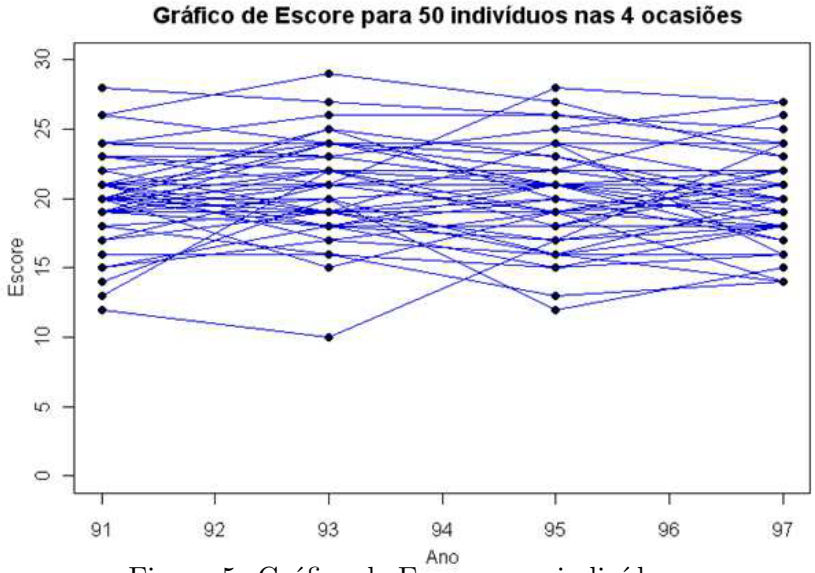


Figura 5: Gráfico de Escore para indivíduos.

Com o ajuste marginal temos:

```
geeglm(score~asex+ aagecat+ aeduc+ amumwk+ ecact+ time,family=gaussian,data=data2,id=pid,corstr="exchangeable")
             Estimate   Std.err    Wald Pr(>|W|)
(Intercept)  19.4937    0.1904   10479.087  < 2e-16 ***
asex2         1.2541    0.1516    68.439  < 2e-16 ***
aagecat2     -0.5794    0.1764    10.788 0.001022 **
aagecat3     -0.7364    0.2140    11.840 0.000580 ***
aagecat4     -1.6530    0.2874    33.082 8.83e-09 ***
aeduc1        0.5436    0.1509    12.977 0.000315 ***
amumwk1       1.0453     0.1569    44.412 2.66e-11 ***
ecact4       -1.8832    0.2225    71.627  < 2e-16 ***
ecact3       -0.3032    0.1036     8.565 0.003427 **
ecact2       -0.5167    0.1518    11.579 0.000667 ***
time          -0.1011    0.0147    47.273 6.17e-12 ***
Estimated Correlation Parameters:
      Estimate Std.err
alpha    0.583   0.015
```

Figura 6: Ajuste do Modelo Marginal.

Interpretamos os coeficientes a cima em relação a sua categoria de referência. Por exemplo, em termos da variável asex2, a referência se encontra em asex1, sexo masculino. O coeficiente 1.2541 nos mostra então, que a média populacional feminina é cerca de 1,25 unidades a mais no score de atitude do papel do gênero do que o sexo masculino. Para ecact4 (variável relacionado à situação econômica), dizemos que a média populacional para mulheres do lar tem cerca de 1,8 unidades a menos na atitude do score do que sua categoria de referência (trabalho em período integral).

Com o ajuste do modelo de efeitos aleatórios temos:

```
Linear mixed-effects model fit by REML
Random effects:
 Formula: ~time | pid
            StdDev  Corr
(Intercept) 2.7563 (Intr)
time        0.3318 -0.329
Residual    2.0117

Fixed effects: score ~ asex + aagecat + aeduc + amumwk + ecact
             Value    Std.Error   DF   t-value p-value
(Intercept) 19.151   0.1922    4284     99.66   0.0000
asex2       1.271    0.1513    1422      8.40   0.0000
aagecat2   -0.559    0.1836    1422     -3.04   0.0024
aagecat3   -0.711    0.2097    1422     -3.39   0.0007
aagecat4   -1.636    0.2925    1422     -5.59   0.0000
aeduc1      0.554    0.1527    1422      3.63   0.0003
amumwk1    1.054    0.1547    1422      6.81   0.0000
ecact4     -1.904    0.1928    4284     -9.88   0.0000
ecact3     -0.194    0.1090    4284     -1.78   0.0753
ecact2     -0.532    0.1527    4284     -3.49   0.0005
```

Figura 7: Ajuste do Modelo de efeitos aleatórios.

# 5 Conclusões

No trabalho utilizamos uma enorme quantidade de rotinas e funções para descrever os dados longitudinais. Pela figura 5 visualizamos pequenas mudanças de escores para alguns indivíduos e a partir da modelagem marginal percebemos que as mulheres possuem em média um escore um pouco maior do que os homens. Em relação às atividades econômicas percebemos uma dimi-

nuição no escore de atitude do gênero ao compararmos mulheres do lar e mulheres que trabalham em tempo integral. No modelo de efeitos aleatórios obtivemos as mesmas interpretações.

## Referências


[1] Hojsgaard, S., Halekoh, U., Yan J., The R Package geepack for Generalized Estimating Equations, *Journal of Statistical Software*, **15(2)**, 1-11, 2005.

[2] Liang, Zeger, Longitudinal Data Analysis Using Generalised Linear Models, *Biometrika*, **73(1)**, 13-22, 1986.

[3] Twisk, J. W. R., Applied Longitudinal Data Analysis for Epidemiology: A Practical Guide, 2003.

[4] Vieira, M. D. T., Analysis of Longitudinal Survey Data, 2009.