

Dados longitudinais: introdução a *generalized estimating equations* (GEE)

Pedro S.R. Martins

Lavis 2021



Roteiro

- Dados longitudinais
- Comparações com medidas repetidas
- Revisão brevíssima sobre modelos de regressão
- Anova de medidas repetidas VS. GEE
- Pressupostos GEE
- Introdução brevíssima sobre diferentes distribuições
- Matrizes de correlação
- Ajuste do modelo
- Tutorial SPSS

Primeiras coisas primeiro



Produção e Divulgação Científica
www.cientifica.com
http://naruhodo.b9.com.br

Científica & Podcast Naruhodo


HOME VIDEOS PLAYLISTS COMMUNITY CHANNELS ABOUT

BEM VINDO(A) AO CANAL CIENTÍFICA

Apresentação Canal Científica e Podcast Naruhodo no YT
Científica & Podcast Naruhodo • 3.7K views • 1 year ago
Canal Científica Produção Científica com base em vídeos de Estatística Aplicada, Metodologia e Epistemologia www.cientifica.com Podcast Naruhodo! Naruhodo! é o podcast pra quem tem...

Lives Estatística Nível I Psicobio UNIFESP 2021 ▶ PLAY ALL

| Thumbnail | Video Title | Views | Time |
|-----------|----------------------------------------------------------|------------|---------|
| | Estatística Psicobio I 2021 #01 - Teoria da Medida... | 3.1K views | 1:53:20 |
| | Estatística Psicobio I 2021 #02 - Tipos de Variável e... | 1.4K views | 2:35:21 |
| | Estatística Psicobio I 2021 #03 - Medidas Descritivas... | 1.1K views | 2:29:37 |
| | Estatística Psicobio I 2021 #04 - TCL e Intervalos de... | 988 views | 2:31:33 |
| | Estatística Psicobio I 2021 #05 - Intervalos de... | 790 views | 2:29:33 |



Produção e Divulgação Científica
www.cientifica.com
http://naruhodo.b9.com.br

Científica & Podcast Naruhodo

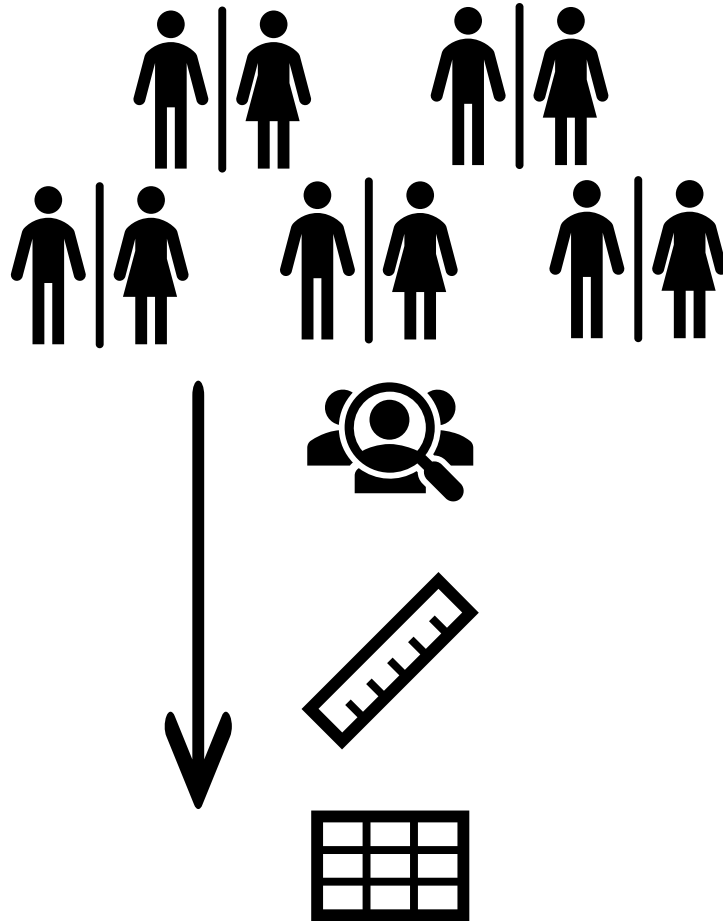
PLAY ALL

Lives - Disciplina Estatística aplicada a Psicobiologia Nível II - UNIFESP
28 videos • 5,713 views • Last updated on Dec 7, 2020

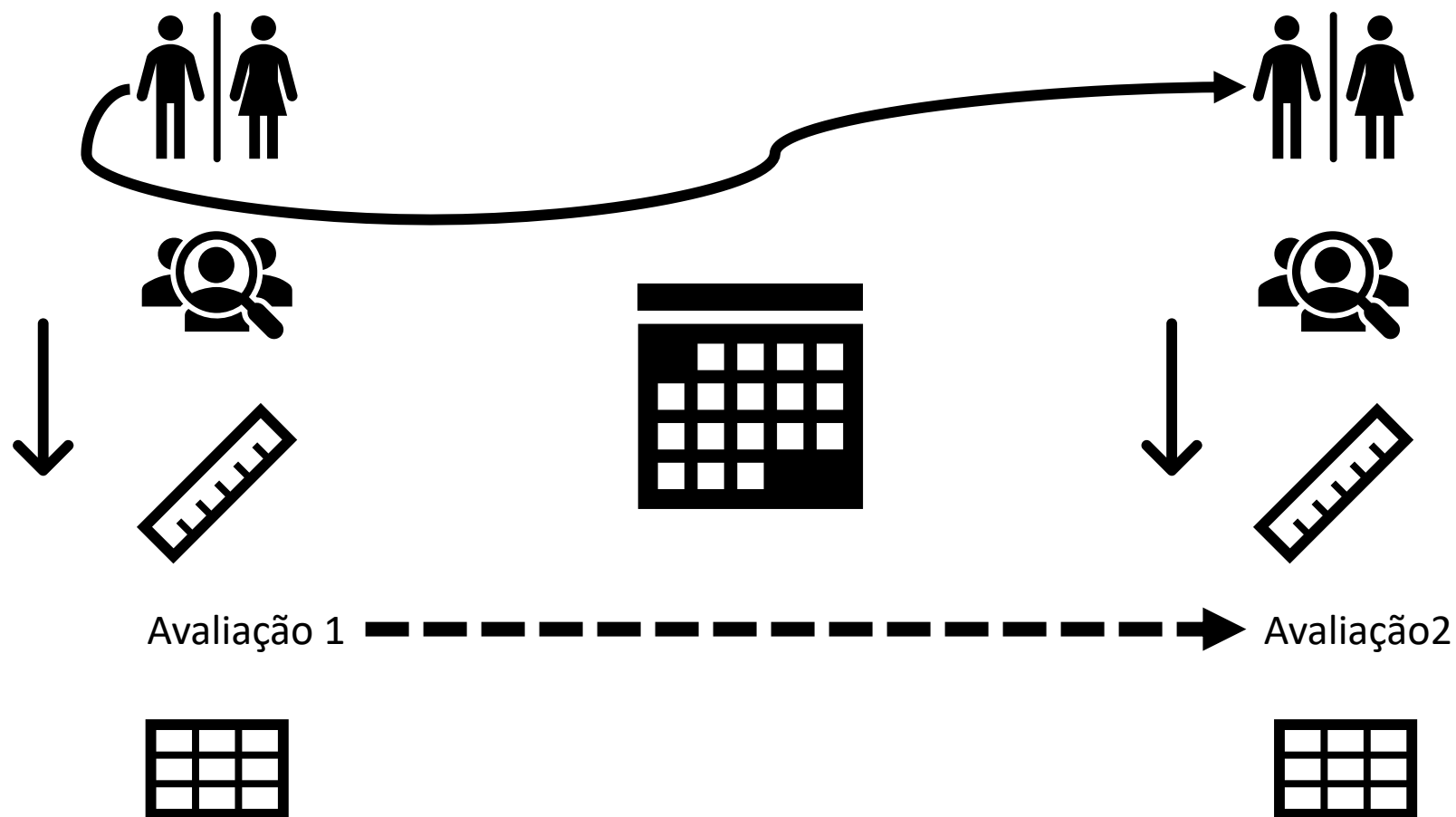
| Thumbnail | Video Title | Views | Time |
|-----------|--------------------------------------------------------------------------------------------------|-------|---------|
| | Estatística Psicobio II - Revisão Regressão e o início do pensamento científico | | 2:19:30 |
| | Estatística Psicobio II - Escrevendo modelos - Distribuições para variáveis discretas Parte I | | 2:40:56 |
| | Estatística Psicobio II - Distribuições para variáveis discretas Parte II - Bin, BinNeg, Poisson | | 2:29:25 |
| | Estatística Psicobio II - Distribuições para variáveis contínuas | | 2:32:12 |
| | Estatística Psicobio II - GEE - Generalized Estimated Equations - Parte I | | 2:35:51 |
| | Estatística Psicobio II - GEE - Generalized Estimated Equations - Parte II | | 2:24:38 |



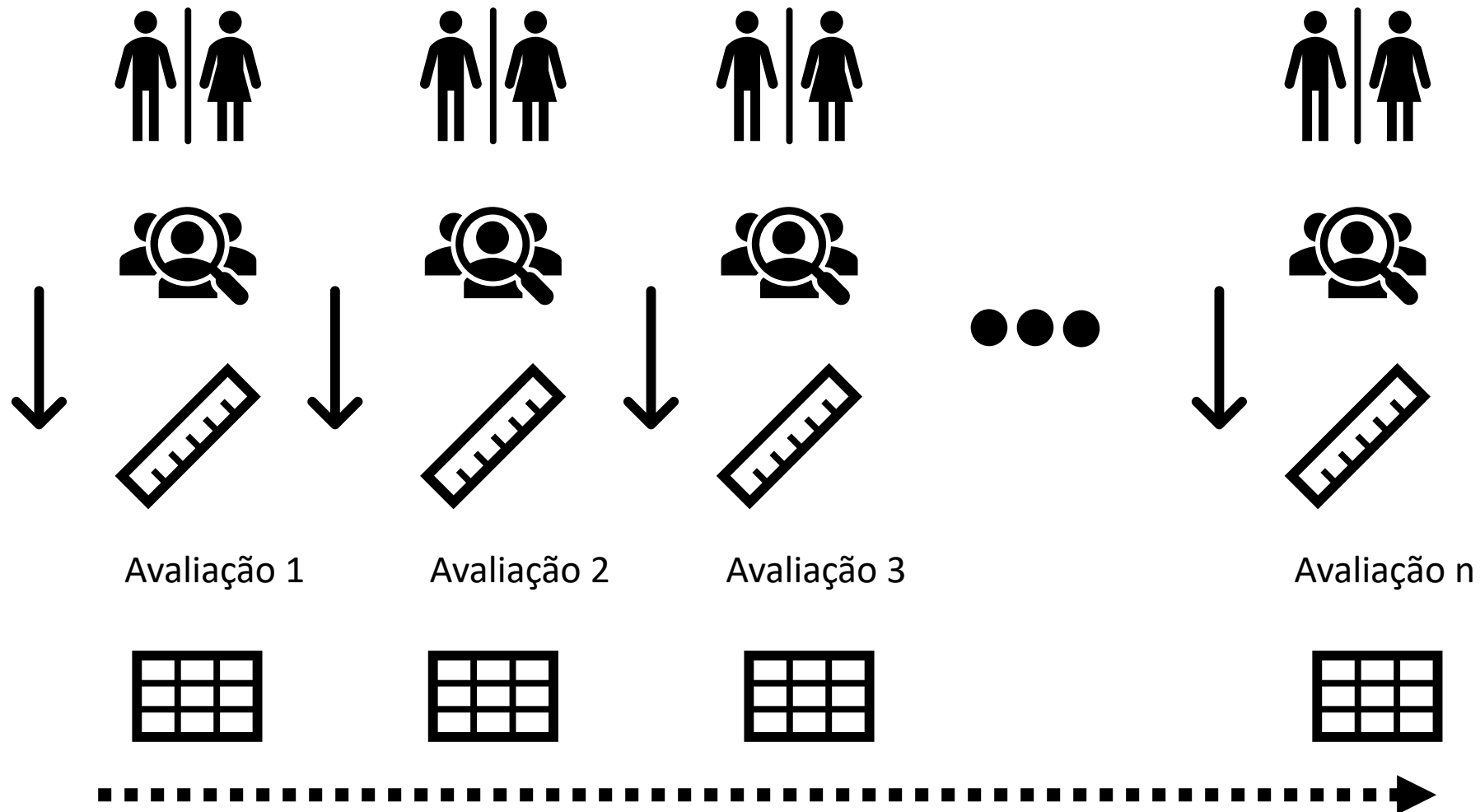
Dados Transversais



Dados longitudinais



Dados longitudinais

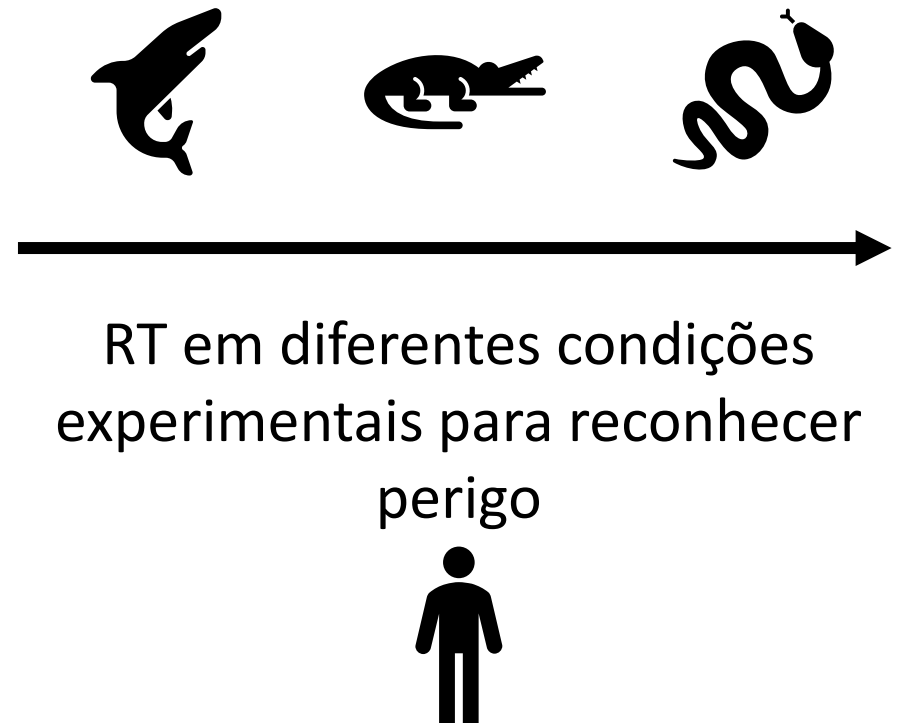


Dados longitudinais

- Dados longitudinais são aqueles em que as mesmas pessoas são avaliadas mais de uma vez
- **Dados em painel** (*panel data*), termo mais usado em econometria e faz referência a dados longitudinais
- **Cuidado:** “medidas repetidas” não é sinônimo de dado longitudinal



A pessoa pode ser avaliada uma vez só em diferentes condições experimentais



Dados longitudinais

Vantagens

- Pode necessitar de menos indivíduos para achar um efeito
- Cada sujeito serve como seu próprio controle, *within subject design*
- Oferece informações sobre padrões de mudança
- Possibilidades de avaliar a confiabilidade das medidas (e.g., McCrae e Möttus, 2019, McCrae, 2015)*

Desvantagens

- Diferenças entre efeitos marginais e condicionais
- Dados desbalanceados (e.g., diferenças entre N nos tempos)
- Dificuldades para analisar e interpretar dados longitudinais
- \$

Confiabilidade em dados longitudinais

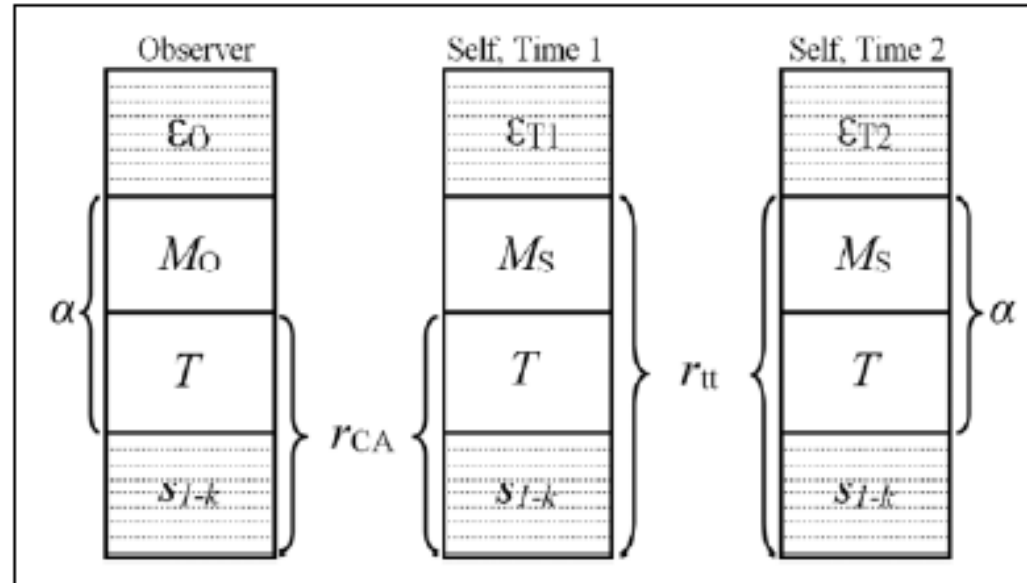


Figure 1. Schematic representation of the proportions of variance attributable to error (ϵ), method (M), common trait (T), and item specifics (s_{I-k}) in scale scores from an observer and from a self-report test and short-term retest. Dashed lines indicate distinct contributions from different items. Internal consistency (α) is attributable to method and trait variance shared by items. Cross-observer agreement (r_{CA}) and retest reliability (r_{tt}) reflect shared components of variance across observers and occasions, respectively.

Estrutura do banco de dados

Dados em formato *wide* (largo)

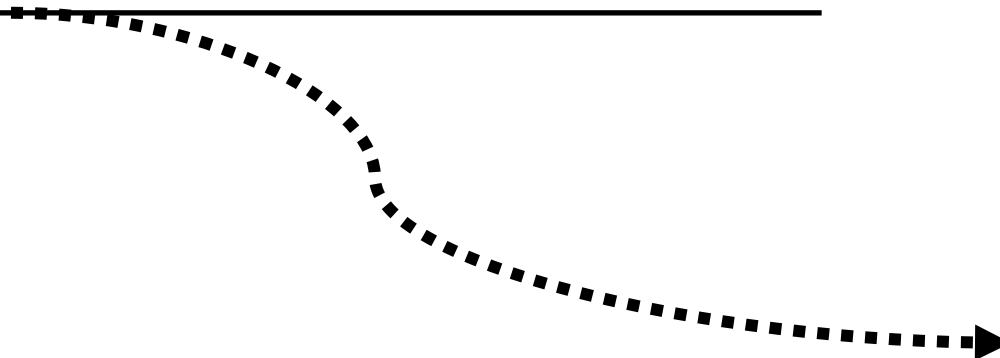
| Id | Medida_av1 | Medida_av2 | Medida_av3 | Medida_avn |
|----|------------|------------|------------|------------|
| 1 | 2 | 3 | 6 | 20 |
| 2 | 5 | 6 | 3 | 1 |
| 3 | 7 | 5 | 6 | 2 |
| 4 | 10 | 5 | 6 | 5 |

Estrutura do banco de dados

Dados em formato *long* (longo)

| Id | Medida_av1 | Medida_av2 | Medida_av3 | Medida_avn |
|----|------------|------------|------------|------------|
| 1 | 2 | 3 | 6 | 20 |
| 2 | 5 | 6 | 3 | 1 |
| 3 | 7 | 5 | 6 | 2 |
| 4 | 10 | 5 | 6 | 5 |

| Id | Tempo | Medida |
|----|-------|--------|
| 1 | av1 | 2 |
| 1 | av2 | 3 |
| 1 | av3 | 6 |
| 1 | avn | 20 |
| 2 | av1 | 5 |
| 2 | av2 | 6 |
| 2 | av3 | 3 |
| 2 | avn | 1 |
| 3 | av1 | 7 |
| 3 | av2 | 5 |
| 3 | av3 | 6 |
| 3 | avn | 2 |
| 4 | av1 | 10 |
| 4 | av2 | 5 |
| 4 | av3 | 6 |
| 4 | avn | 5 |



Estrutura do banco de dados

- O formato do banco pode variar de acordo com o programa e com a análise que será realizada. **Não existe um formato melhor do que outro.**
- O mais importante é saber que os dois existem e entender como cada programa trabalha os dados

| Id | Medida_av1 | Medida_av2 | Medida_av3 | Medida_avn |
|----|------------|------------|------------|------------|
| 1 | 2 | 3 | 6 | 20 |
| 2 | 5 | 6 | 3 | 1 |
| 3 | 7 | 5 | 6 | 2 |
| 4 | 10 | 5 | 6 | 5 |

| Id | Tempo | Medida |
|----|-------|--------|
| 1 | av1 | 2 |
| 1 | av2 | 3 |
| 1 | av3 | 6 |
| 1 | avn | 20 |
| 2 | av1 | 5 |
| 2 | av2 | 6 |
| 2 | av3 | 3 |
| 2 | avn | 1 |
| 3 | av1 | 7 |
| 3 | av2 | 5 |
| 3 | av3 | 6 |
| 3 | avn | 2 |
| 4 | av1 | 10 |
| 4 | av2 | 5 |
| 4 | av3 | 6 |
| 4 | avn | 5 |

Comparações dados longitudinais

Dados transversais

2 grupos

Distribuição normal – teste t

Não paramétrico – Mann-Whitney

3 +
grupos

Distribuição normal – ANOVA

Não paramétrico – Kruskal-Wallis

Dados longitudinais

2 tempos

Distribuição normal – teste t
pareado

Não paramétrico – Wilcoxon

3 +
tempos

Distribuição normal – ANOVA de
medidas repetidas

Não paramétrico – ANOVA de
Friedman

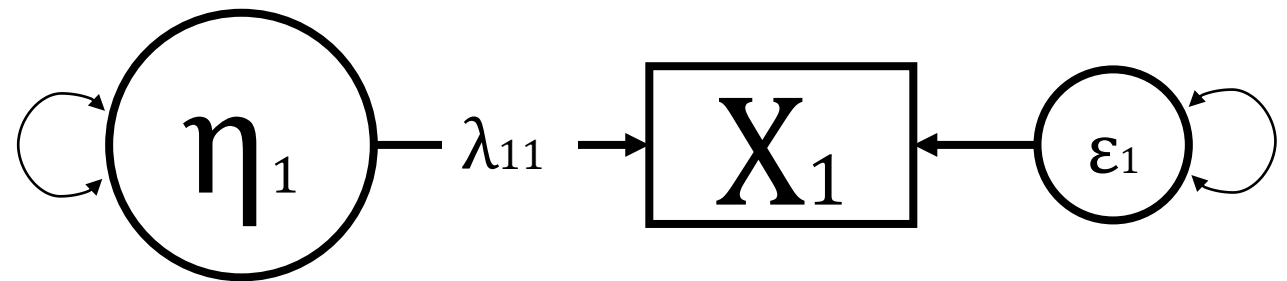
Modelos de regressão

Todas as análises mencionadas anteriormente são extrapolações dos modelos de regressão

- $y = a \cdot x + b$

- $Y = \beta_0 + \beta_{x1} + \varepsilon$

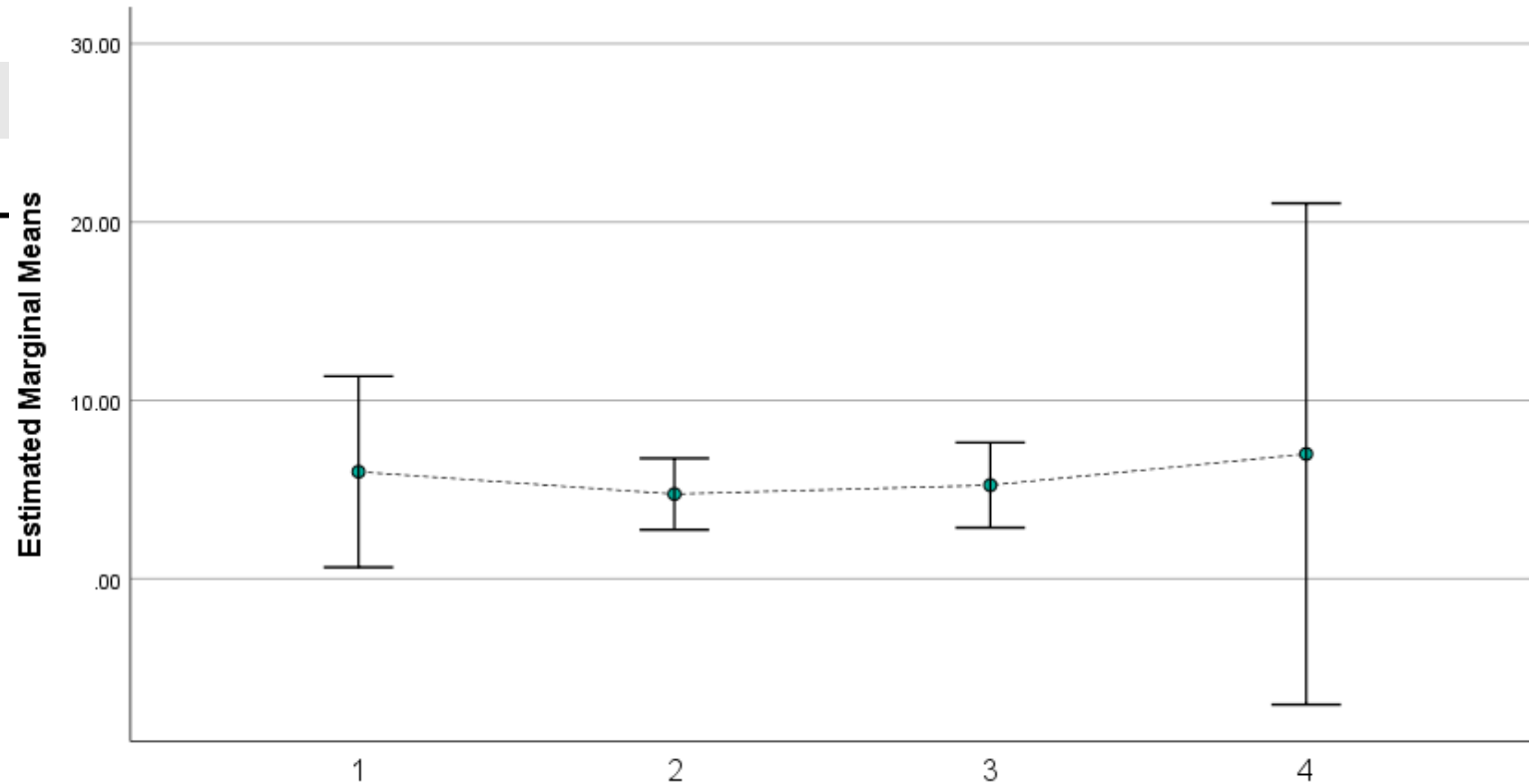
- Em dados longitudinais, o fato de ser a mesma pessoa variáveis vezes aparece no termo do erro.
- Na regressão, os erros não devem ser correlacionados
- Em dados longitudinais, esse pressuposto é quebrado



$$X = \mu + \lambda^* \eta + \varepsilon$$

ANOVA de medidas repetidas

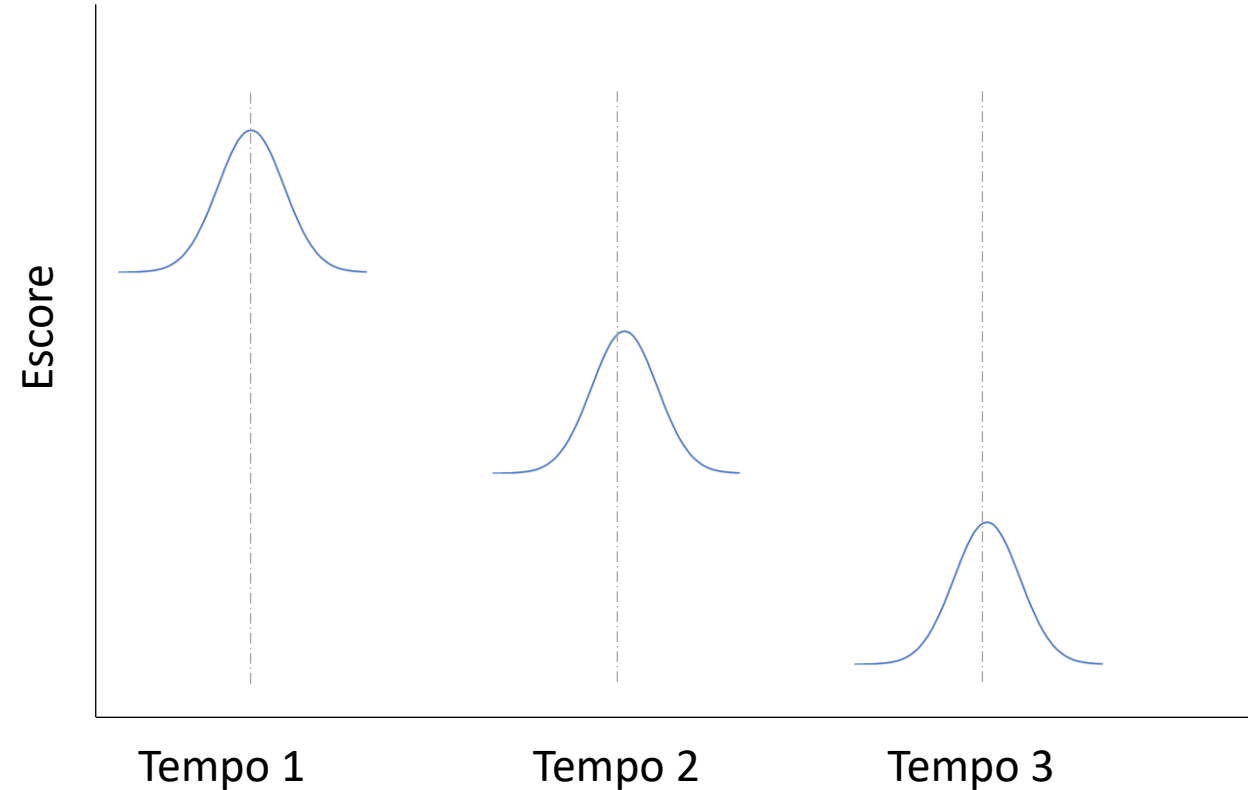
| Id | Medida_av1 | Medida_av2 | Medida_av3 | Medida_avn |
|----|------------|------------|------------|------------|
| 1 | 2 | 3 | 6 | 20 |
| 2 | 5 | 6 | | |
| 3 | 7 | 5 | | |
| 4 | 10 | 5 | | |



ANOVA de medidas repetidas

- Pressupostos

- *Quantitativa e intervalar*
- *Distribuição normal multivariada*
- *Igualdade das variâncias entre os momentos de avaliação*
- *Correlação constante entre os tempos de avaliação (esfericidade)*



Generalized estimating equations

Biometrika (1986), 73, 1, pp. 13–22
Printed in Great Britain

13

Longitudinal data analysis using generalized linear models

BY KUNG-YEE LIANG AND SCOTT L. ZEGER

Department of Biostatistics, Johns Hopkins University, Baltimore, Maryland 21205, U.S.A.

SUMMARY

This paper proposes an extension of generalized linear models to the analysis of longitudinal data. We introduce a class of estimating equations that give consistent estimates of the regression parameters and of their variance under mild assumptions about the time dependence. The estimating equations are derived without specifying the joint distribution of a subject's observations yet they reduce to the score equations for multivariate Gaussian outcomes. Asymptotic theory is presented for the general class of estimators. Specific cases in which we assume independence, m -dependence and exchangeable correlation structures from each subject are discussed. Efficiency of the proposed estimators in two simple situations is considered. The approach is closely related to quasi-likelihood.

Some key words: Estimating equation; Generalized linear model; Longitudinal data; Quasi-likelihood; Repeated measures.

1. INTRODUCTION

Longitudinal data sets, comprised of an outcome variable, y_{it} , and a $p \times 1$ vector of covariates, x_{it} , observed at times $t = 1, \dots, n_i$ for subjects $i = 1, \dots, K$ arise often in applied sciences. Typically, the scientific interest is either in the pattern of change over time, e.g. growth, of the outcome measures or more simply in the dependence of the outcome on the covariates. In the latter case, the time dependence among repeated measurements for a subject is a nuisance. For example, the severity of respiratory disease along with the nutritional status, age, sex and family income of children might be observed once every three months for an 18 month period. The dependence of the outcome variable, severity of disease, on the covariates is of interest.

With a single observation for each subject ($n_i = 1$), a generalized linear model (McCullagh & Nelder, 1983) can be applied to obtain such a description for a variety of continuous or discrete outcome variables. With repeated observations, however, the correlation among values for a given subject must be taken into account. This paper presents an extension of generalized linear models to the analysis of longitudinal data when regression is the primary focus.



Kung-Yee Liang



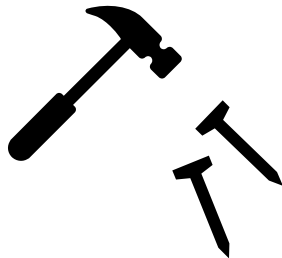
Scott L. Zeger

KUNG-YEE LIANG, SCOTT L. ZEGER, Longitudinal data analysis using generalized linear models, *Biometrika*, Volume 73, Issue 1, April 1986, Pages 13–22, <https://doi.org/10.1093/biomet/73.1.13>

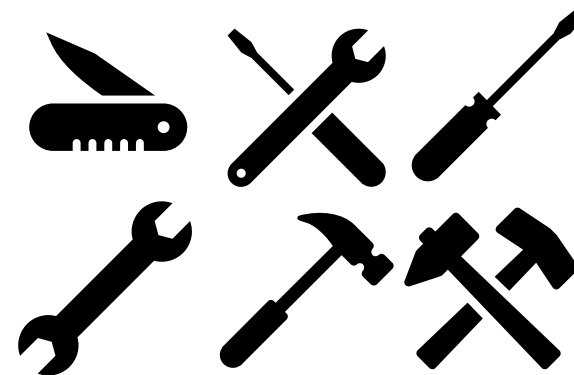
Generalized estimating equations

- **Definição:** GEE é uma extensão de modelos lineares generalizados para estimar estimativas médias para a população controlando para a dependência entre medidas repetidas
- GEE assume um formato de equação em que a variável dependente pode ser de qualquer distribuição, e a interpretação varia de acordo com a função de ligação que escolhemos

ANOVA



GEE



Usos GEE

- Podemos pensar na GEE como uma regressão longitudinal
 - Ou seja, erros correlacionados
- Mas, para isso, é necessário **conhecer a variável dependente**

VD {

- Variável dicotômica: regressão logística
- Variável contínua: (1) distribuição normal (2) distribuição gamma
- Variável de contagem: regressão Poisson
- ...

Pressupostos GEE

- As respostas na variável dependente são correlacionadas (ou seja, um indivíduo responde mais de uma vez)
- Covariáveis existem como um função linear ou não linear, podendo interagir entre si
- A homogeneidade da variância não é satisfeita
- O usuário especifica a matriz de correlação correta dos erros (working correlation matrix)
 - (a) especificar a função de ligação
 - (b) conhecer a distribuição da variável dependente
 - (c) conhecer a distribuição das correlações

Matriz de correlação dos erros

1. Independente (sem correlação entre tempos)
2. Intercambiável (exchangeable, a.k.a simétrica): as correlações são as mesmas entre os tempos
3. Autorregressiva de ordem 1 (AR1): assume que as correlações das observações são dependentes dos valores anteriores apenas, por meio de um processo autorregressivo
4. Matriz não estruturada: não assume *constraints* para as correlações par a par entre os tempos

Independente

$$\begin{array}{c} \text{T1} \\ \text{T2} \\ \text{T3} \end{array} \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

Dados transversais normalmente

Intercambiável (simétrica)

$$\begin{array}{c} \text{T1} \\ \text{T2} \\ \text{T3} \end{array} \begin{array}{ccc} \text{T1} & \text{T2} & \text{T3} \\ \left[\begin{array}{ccc} 1 & \rho & \rho \\ \rho & 1 & \rho \\ \rho & \rho & 1 \end{array} \right] \end{array}$$

Efeitos lineares

Correlação é a mesma entre todos os tempos

Efeito similar à ANOVA

Autorregressiva de ordem 1 (AR1)

$$\begin{matrix} & \begin{matrix} T1 & T2 & T3 \end{matrix} \\ \begin{matrix} T1 \\ T2 \\ T3 \end{matrix} & \begin{bmatrix} 1 & \rho & \rho^2 \\ \rho & 1 & \rho \\ \rho^2 & \rho & 1 \end{bmatrix} \end{matrix}$$

Assume que as correlações das observações são dependentes dos valores anteriores apenas, por meio de um processo autorregressivo

A correlação entre T1 e T3 é menor que entre T1 e T2

Efeitos logaritmos, exponenciais e não monotônicos

Não estruturada

$$\begin{array}{c} \text{T1} \\ \text{T2} \\ \text{T3} \end{array} \begin{array}{ccc} \text{T1} & \text{T2} & \text{T3} \\ \left[\begin{array}{ccc} 1 & \rho_{1-2} & \rho_{1-3} \\ \rho_{1-2} & 1 & \rho_{2-3} \\ \rho_{1-3} & \rho_{2-3} & 1 \end{array} \right] \end{array}$$

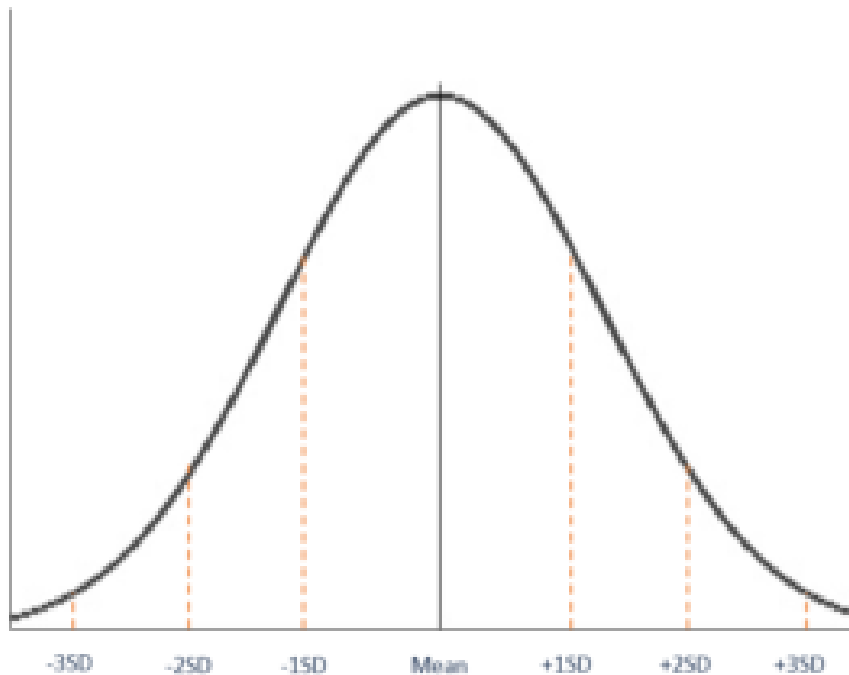
Não sei qual o padrão de mudança e não quero chutar

Ajuste do modelo

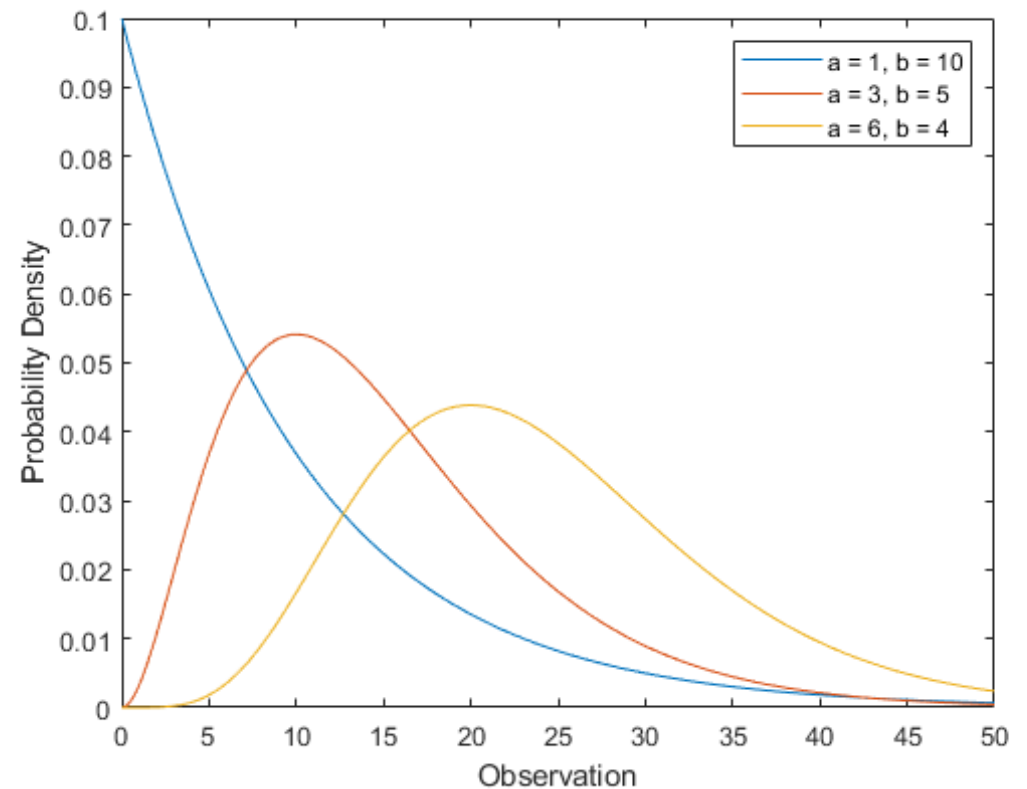
- não tem um índice de ajuste aproximado ou absoluto, porque os modelos são feitos em procedimentos de estimação
- Entretanto, é possível ver testes de qui quadrado para as VI, análise do resíduo (**o resíduo deve aproximar uma normal!**)
- Comparação de modelos usando **QIC** (mesma lógica do AIC, quanto menor melhor)

Diferentes distribuições

Distribuição normal

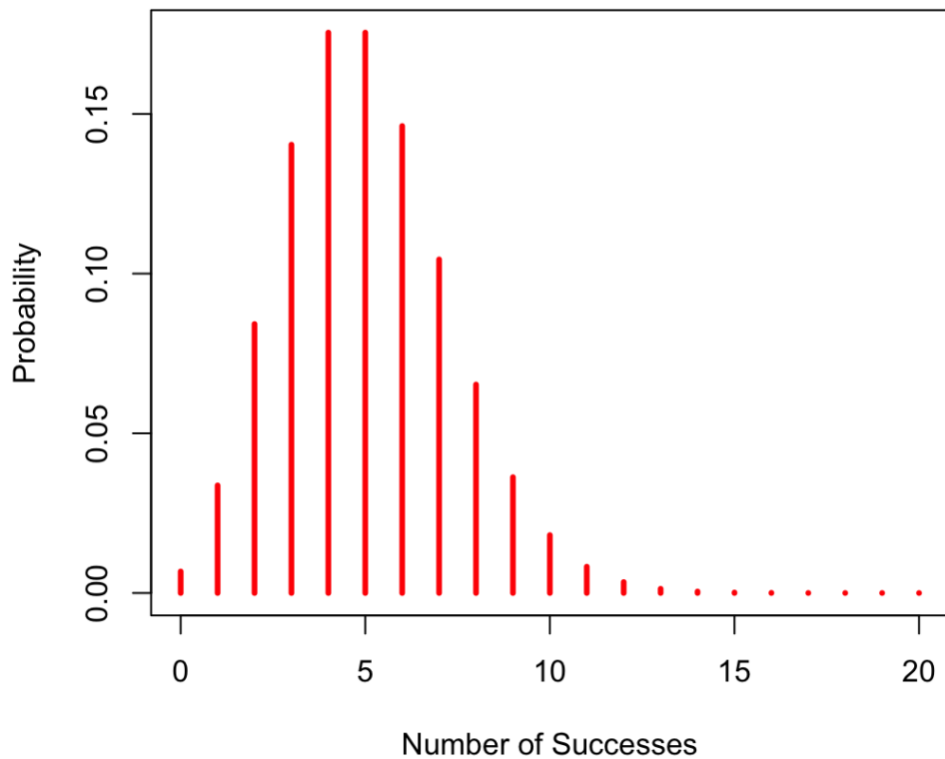


Distribuição Gamma

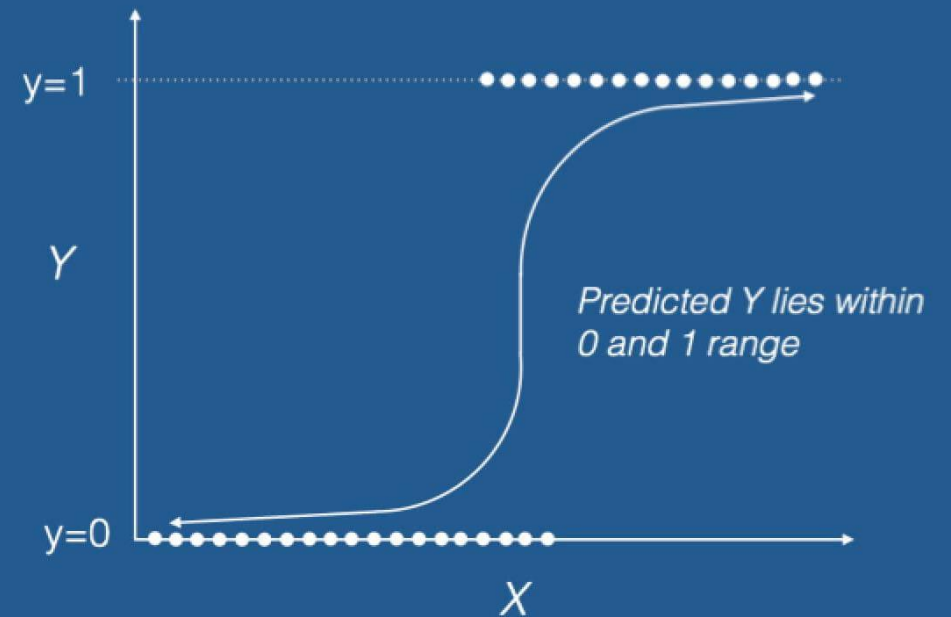


Diferentes distribuições

Poisson Distribution



Logistic Regression



Outras distribuições

- Binomial
- Gamma
- Gaussiana invertida
- Binomial negativa
- Normal
- Poisson
- Tweedie
- Multinomial

Link de ligação

- Como interpretar os resultados?
- Basicamente, como regressão logística ou linear?

Tutorial SPSS

- <https://www.ibm.com/docs/en/spss-statistics/23.0.0?topic=equations-generalized-estimating-type-model>

Aqui seleciona-se a análise

PSS Statistics Data Editor

rm Analyze Graphs Utilities Extensions Window Help

Reports
Descriptive Statistics
Bayesian Statistics
Tables
Compare Means
General Linear Model
Generalized Linear Models
Mixed Models
Correlate
Regression
Loglinear
Neural Networks
Classify
Dimension Reduction
Scale
Nonparametric Tests
Forecasting
Survival
Multiple Response
Missing Value Analysis...
Multiple Imputation
Complex Samples
Simulation...
Quality Control
Spatial and Temporal Modeling...
Direct Marketing

zucubos_t1 zvocabulario_t1 zestresse_t2

-2.85 -3.51 -4.00
-1.37 1.78 -1.00
-0.36 -1.11 -0.50
1.21 -0.39 -1.00
-0.29 -2.09 -0.50
-0.99 0.03 -1.00
2.05 0.60 0.50
-0.29 1.94 0.00
-0.96 -1.03 -1.00
-0.16 -0.76 -0.50
0.61 2.39 -0.50
-0.46 -2.42 -2.00
-0.86 0.10 0.00
-3.04 -2.27 -0.50
0.66 0.51 -1.00
-1.06 -0.01 1.00
-1.08 -1.94 -0.50
-1.11 0.71 0.00
0.37 -0.34 -0.50
-0.97 0.17 0.00
-0.85 1.22 0.00
-1.54 -2.63 -2.00
-0.32 -1.92 -1.00
-1.87 -1.41 -1.00
0.95 1.35 1.00

Generalized Linear Models...
Generalized Estimating Equations...

Generalized Estimating Equations

Repeated Type of Model Response Predictors Model Estimation Statistics EM Means Save Export

Variables:

- sexo
- perfeccionismo
- zcubos_t1
- zvocabulario_t1
- zestresse_t2
- zneuroticismo_t3

Subject variables:

id

Variável dos sujeitos

Within-subject variables:

tempo

Variável de medidas repetidas

☒ Sort cases by subject and within-subject variables

Covariance Matrix

☒ Robust estimator ☐ Model-based estimator

Working Correlation Matrix

Structure: Independent

☒ Adjust estimator by number of non-redundant parameters

Maximum iterations: 100

☒ Update matrix Iterations by

Convergence Criteria

At least one convergence criterion must be specified with a minimum greater than zero.

☒ Change in parameter estimates Minimum: 1E-006 Type: Absolute

☐ Hessian convergence Minimum: Absolute

OK Paste Reset Cancel Help

Working Correlation Matrix

Structure: Independent

☒ Adjust estimator by number of non-redundant parameters

Maximum iterations: 100

☒ Update matrix Iterations by

Convergence Criteria

At least one convergence criterion must be specified with a minimum greater than zero.

☒ Change in parameter estimates Minimum: 1E-006 Type: Absolute


☐ Hessian convergence Minimum: Absolute

Escolha seu veneno: qual a matriz de correlação dos efeitos

Generalized Estimating Equations


Repeated Type of Model Response Predictors Model Estimation Statistics EM Means Save Export

Choose one of the model types listed below or specify a custom combination of distribution and link function.

 Scale Response


☒ Linear

☐ Gamma with log link

 Ordinal Response


☐ Ordinal logistic

☐ Ordinal probit

 Counts

☐ Poisson loglinear


☐ Negative binomial with log link

 Binary Response or Events/Trials Data

☒ Binary logistic


☐ Binary probit

☐ Interval censored survival

 Mixture

☐ Tweedie with log link

☐ Tweedie with identity link

 Custom

☐ Custom

Distribution: Normal Link function: Identity

Power:

Parameter

☒ Specify value

Value:

☐ Estimate value

OK Paste Reset Cancel Help

Escolha seu veneno: qual a distribuição da sua variável dependente

Repeated

Type of Model

Response

Predictors

Model

Estimation

Statistics

EM Means

Save

Export

Variables:

id
tempo
sexo
zcubos_t1
zvocabulario_t1
zestresse_t2
zneuroticismo_t3

Dependent Variable



Dependent Variable:

perfeccionismo

Category order

Ascending

Type of Depend

on Only)

☒ Binary

Reference Category...

☒ Number of events occurring in a set of trials

Trials

☒ Variable

Trials Variable:

☒ Fixed value

Number of Trials:

Scale Weight



Scale Weight Variable:

OK

Paste

Reset

Cancel

Help

Insira sua VD

Generalized Estimating Equations

Repeated Type of Model Response Predictors Model Estimation Statistics EM Means Save Export

Variables:

id

Factors:

tempo
sexo

Options...

Covariates:

zcubos_t1
zvocabulario_t1
zestresse_t2
zneuroticismo_t3

Offset

☒ Variable

Offset Variable:

☐ Fixed value

Value:

OK Paste Reset Cancel Help

Monte seu modelo:

- *Factors*: variáveis independentes categóricas, ordinais ou nominais
- *Covariates*: variáveis independentes intervalares/contínuas

Generalized Estimating Equations

Repeated Type of Model Response Predictors **Model** Estimation Statistics EM Means Save Export

Specify Model Effects

Factors and Covariates:

- tempo
- sexo
- zcubos_t1
- zvocabulario_t1
- zestresse_t2
- zneuroticismo_t3

Build Term(s)

Type:

Interaction

Model:

- tempo
- sexo
- zcubos_t1
- zvocabulario_t1
- zestresse_t2
- zneuroticismo_t3
- tempo*sexo

Number of Effects in Model: 7

Build Nested Term

Term:

By * (Within) Add to Model Clear

☒ Include intercept in model

OK Paste Reset Cancel Help

Monte seu modelo:
Escolha efeitos principais
com interações ou não

Generalized Estimating Equations

Repeated Type of Model Response Predictors Model Estimation **Statistics** EM Means Save Export

Model Effects

Analysis Type: Type III Confidence Interval Level (%): 95

Chi-square Statistics

☒ Wald
☐ Generalized score

Log quasi-likelihood function: Full

Print

☒ Case processing summary
☒ Descriptive statistics
☒ Model information
☒ Goodness of fit statistics
☒ Model summary statistics
☒ Parameter estimates
☐ Include exponential parameter estimates
☐ Covariance matrix for parameter estimates
☐ Correlation matrix for parameter estimates
☐ Working correlation matrix

☐ Contrast coefficient (L) matrices
☐ General estimable functions
☐ Iteration history

Print interval: 1

OK Paste Reset Cancel Help

Aqui é necessário pedir os *exponential parameter estimates* em caso de **regressão logística** para se obter os odds. Caso contrário, pode permanecer tudo no padrão.

Repeated

Type of Model

Response

Predictors

Model

Estimation

Statistics

EM Means

Save

Export

Factors and Interactions:

| M | | Term |
|-------------------------------------|--|------------|
| <input checked="" type="checkbox"/> | | tempo |
| <input checked="" type="checkbox"/> | | sexo |
| <input checked="" type="checkbox"/> | | tempo*sexo |



By *

Scale

- ☒ Compute means for response
☐ Compute means for linear predictor

Adjustment for Multiple Comparisons:

Bonferroni

Display Means for:

| Term | Contrast | Reference Category |
|------------|----------|--------------------|
| tempo | Pairwise | |
| sexo | Pairwise | |
| tempo*sexo | Pairwise | |

Post-hoc: aqui podemos pedir um posthoc igual a uma anova, com correções para comparação múltipla

☐ Display overall estimated mean

OK

Paste

Reset

Cancel

Help

Generalized Estimating Equations

Repeated

Type of Model

Response

Predictors

Model

Estimation

Statistics

EM Means

Save

Export

| Save | Item to Save | Variable Name or Root N... | Categories to Save |
|-------------------------------------|-----------------------------------------------------------------|----------------------------|--------------------|
| <input type="checkbox"/> | Predicted value of mean of response | MeanPredicted | |
| <input type="checkbox"/> | Lower bound of confidence interval for mean of response | CIMeanPredictedLower | |
| <input type="checkbox"/> | Upper bound of confidence interval for mean of response | CIMeanPredictedUpper | |
| <input type="checkbox"/> | Predicted category | PredictedValue | |
| <input type="checkbox"/> | Predicted value of linear predictor | XBPredicted | |
| <input type="checkbox"/> | Estimated standard error of predicted value of linear predictor | XBStandardError | |
| <input checked="" type="checkbox"/> | Residual | Residual | |
| <input type="checkbox"/> | Pearson residual | PearsonResidual | |

Salvar o resíduo para
avaliar normalidade

Existing Variable with Same Name

☒ Add suffix to name of new variable (applies only to default names)

☐ Replace existing variable (applies to both default and user-provided names)

i

If you provide your own variable names, make sure that they do not conflict with existing variables in the active dataset. Select the Replace Existing Variable option if you want to overwrite existing variables with the same user-provided name.

OK

Paste

Reset

Cancel

Help