

Análise de Dados Longitudinais

Modelos de Regressão - Perspectiva Histórica

Enrico A. Colosimo/UFMG

<http://www.est.ufmg.br/~enricoc/>

Revisão para Dados Transversais

1 Características

- Informações amostrais independentes (amostra aleatória simples);
- Uma única observação por indivíduo.

2 Modelos para Dados Transversais

- Linear-Normal: Método de Mínimos Quadrados;
- Lineares Generalizados: Método de Máxima Verossimilhança.

3 Método Máxima Verossimilhança

- Função de Verossimilhança para os parâmetros do modelo β (média) e σ (componentes de variância);
- Estimador de Máxima Verossimilhança (EMV);
- Inferência: propriedades assintóticas do EMV;
- Estatísticas: Wald, Escore e RV.

Modelos para Dados Transversais

1 Resposta Contínua

- Modelo regressão linear-normal.
- A resposta é assumida com distribuição normal.

2 Resposta Categórica/Contagem

- Resposta binária: Modelo de regressão logística.
- Resposta contagem: Modelo de regressão de Poisson.
- Outros modelos: beta, gama, etc.

Como Analisar Dados Longitudinais?

- 1 Reduzir os valores repetidos em uma medida resumo.
 - Diferença (teste t pareado);
 - Média ou mediana;
 - Área sob a curva ou inclinação de reta;
 - Em seguida, analisar como dados transversais.
- 2 Ignorar a correlação entre as observações do mesmo indivíduo.
 - Usar modelos de regressão para dados transversais;
 - Estimadores dos parâmetros da média são consistentes (mais ineficientes);
 - Estimativa dos Componentes de variância não são consistentes. No entanto, podem ser corrigidos utilizando um estimador robusto (Equações de Estimação Generalizadas - GEE).

Como Analisar Dados Longitudinais?

3 Modelo Marginal

- Modelar separadamente a média e a estrutura de covariância.
- Encontrar EMV ou MQG.
- Dificuldades no tratamento de dados longitudinais desbalanceados.

4 Modelo Condicional ou de Efeitos Aleatórios

- Tratar os coeficientes como sendo aleatório para as covariáveis que mudam no tempo (por exemplo, intercepto e coeficiente do tempo);
- As diferenças entre os perfis surgem porque os coeficientes de regressão variam entre indivíduos;
- A correlação entre as medidas no mesmo indivíduo são induzidas pelos efeitos aleatórios.

5 Modelo de Transição

- Útil para predição pois utiliza as respostas nos tempos anteriores.

Notação para Dados Longitudinais

Notação (Estrutura Balanceada)

$$Y_i = (Y_{i1}, \dots, Y_{in})', \quad i = 1, \dots, N,$$

é o vetor de respostas do i -ésimo indivíduo.

- N : número de indivíduos;
- Número total de observações: Nn ;
- $E(Y_i) = ((E(Y_{i1}), \dots, E(Y_{in})))'$;
- $\mu_{ij} = E(Y_{ij})$;
- σ_j^2 : variância de Y_{ij} ;
- σ_{jk} : covariância entre Y_{ij} e Y_{ik} .

Estudos Transversais vs Longitudinais

Vetor de Observações longitudinais para o i -ésimo indivíduo:

$$Y_i = (Y_{i1}, \dots, Y_{in})'$$

- No tempo inicial (linha de base, $j = 1$) foram selecionados indivíduos com diferentes idades;
- Os indivíduos foram acompanhados longitudinalmente;
- Desta forma temos duas fontes da variação da resposta com a idade (transversal e longitudinal)

Qual é a diferença dos efeitos?

- Efeito transversal: variação entre indivíduos. Variação da resposta média em função das idades dos indivíduos medida no tempo inicial.
- Efeito longitudinal: variação intra-indivíduo. Variação da resposta média em função da idade no mesmo indivíduo.
- O efeito de idade em um estudo transversal pode estar potencialmente confundido com efeito de coorte.
- Exemplo anterior do efeito de coorte.

Estudos Transversais vs Longitudinais

Estudo Transversal (sem intercepto): $j = 1$

$$Y_{i1} = \beta_T x_{i1} + \epsilon_{i1} \quad i = 1, \dots, N$$

ou

$$E(Y_{i1}) = \beta_T x_{i1} \quad i = 1, \dots, N$$

- β_T representa a diferença da resposta média entre duas sub-populações que diferem por uma unidade em x .
- Se x é a idade, representa o aumento (diminuição) na média de Y para cada incremento de um ano na idade.

Estudos Transversais vs Longitudinais

Estudo Longitudinal

A resposta média aumenta linearmente com mudanças na idade no mesmo indivíduo:

$$E(Y_{ij} - Y_{i1}) = \beta_L(x_{ij} - x_{i1}),$$

β_L representa a mudança esperada em Y para a mudança em uma unidade em x .

Modelo Linear com componentes transversais e longitudinais

$$E(Y_{ij}) = \beta_T x_{i1} + \beta_L(x_{ij} - x_{i1}).$$

Obs.: É necessário assumir $\beta_L = \beta_T$ para estimar mudança da resposta no tempo em estudos transversais (não existe efeito coorte nem de período).

Efeito de Coorte: Transversal e Longitudinal

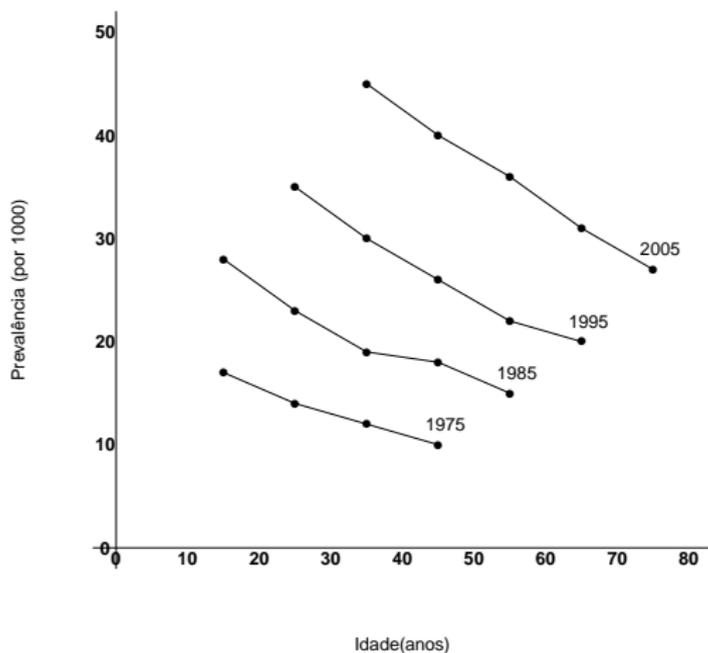


Figura: Estudo Transversal da Prevalência da Doença Y por idade.

Exemplo: Transversais vs Longitudinais// Fitzmaurice e outros (2011, pag. 253)

- Três coortes de crianças com idades iniciais: 5, 6 e 7 anos.
- A resposta foi medida na linha de base e seguida por três anos.
- Suponha que o efeito transversal é linear:

$$E(Y_{i1}) = 0,75 \times \text{idade}_{i1}$$

e que esta relação também vale para $j = 2, 3, 4$.

- Suponha que a resposta média também cresce linearmente com as mudanças na idade em cada coorte. Ou seja

$$E(Y_{ij} - Y_{i1}) = 0,25 \times (\text{idade}_{ij} - \text{idade}_{i1})$$

Exemplo: Estudos Transversais vs Longitudinais

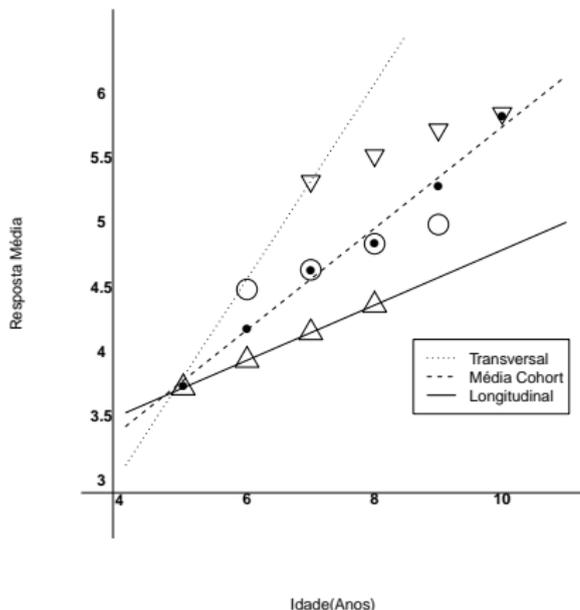


Figura: Resposta Média: transversal vc longitudinal. Transversal: 5,6 e 7 anos. Longitudinal: seguimento por 3 anos. $\beta_T = 0,75$ e $\beta_L = 0,25$.

Erro na figura: os pontos pretos deveriam estar em cima da reta Média Cohort.

Exemplo: Estudos Transversais vs Longitudinais

- Diferença grande entre os efeitos transversal (linha pontilhada) e longitudinal (linha sólida).
- Efeito de corte introduz vício na estimativa transversal quando o efeito longitudinal é ignorado.
- Neste caso o efeito medido é uma combinação ponderada entre β_L e β_T . Ou seja,

$$\hat{\beta} = (1 - w)\hat{\beta}_L + w\hat{\beta}_T$$

em que w depende da proporção de variabilidade (intra e entre indivíduos) e correlação entre as observações intra indivíduo.

Exemplo: Estudos Transversais vs Longitudinais

$$w = \frac{(1 - \rho_y)\rho_x}{(1 - \rho_y) + n\rho_y(1 - \rho_x)}$$

em que:

- ρ_x : proporção de variabilidade em idade devido a variação entre indivíduos;
- ρ_y : proporção de variabilidade na resposta devido a variação entre indivíduos;
- no exemplo $n = 4$ e N indivíduos.
- Para a variação na idade (desenho fixo)

$$\hat{\sigma}_I^2 = \frac{\sum_i \sum_j (x_{ij} - \bar{x}_i)^2}{N(n - 1)} = 5/3$$

e

$$\hat{\sigma}_E^2 = \frac{\sum_i (\bar{x}_i - \bar{x})^2}{N - 1} = 1,0$$

$$\rho_x \approx 0,38$$

Exemplo: Estudos Transversais vs Longitudinais

w varia entre 0, para $\rho_y = 1$, e 0,38, para $\rho_y = 0$.

Ou seja,

$$w = \frac{(1 - \rho_y)0,38}{(1 - \rho_y) + 4\rho_y0,62}$$

- se $\rho_y = 0$ (toda variação intra-indivíduo): $w = 0,38$. Portanto,

$$\hat{\beta} = (1 - 0,38) \times 0,25 + 0,38 \times 0,75 = 0,44$$

- se $\rho_y = 1$ (toda variação entre-indivíduos): $w = 0$. Portanto,

$$\hat{\beta} = (1 - 0) \times 0,25 + 0 \times 0,75 = 0,25$$

Consequências de Ignorar a Correlação em Dados Longitudinais

- Caso mais simples (já avaliado) em que existem somente duas medidas repetidas, digamos nos tempos 1 e 2.
- O objetivo principal do estudo é determinar se existe mudança da média ao longo do tempo. Ou seja

$$\delta = \mu_1 - \mu_2.$$

- Uma estimativa natural para δ é a diferença das médias. Ou seja

$$\hat{\delta} = \hat{\mu}_1 - \hat{\mu}_2.$$

- A variância de $\hat{\delta}$ é

$$\text{Var}(\hat{\delta}) = \frac{1}{N}(\sigma_1^2 + \sigma_2^2 - 2\sigma_{12})$$

Usualmente, $\sigma_{12} > 0$.

Extensão para $n(> 2)$ grupos

Como fazer a comparação para mais de dois grupos?

Exemplos:

- (Dados Longitudinais) A pressão sistólica foi medida, para cada paciente, no tempo inicial (0), após 30 e 60 dias da aplicação da droga.
- (Medidas Repetidas) Três tratamentos foram aplicados de forma aleatória na mesma unidade amostral.

Extensão para $n(> 2)$ grupos

- Interesse é testar a seguinte hipótese:

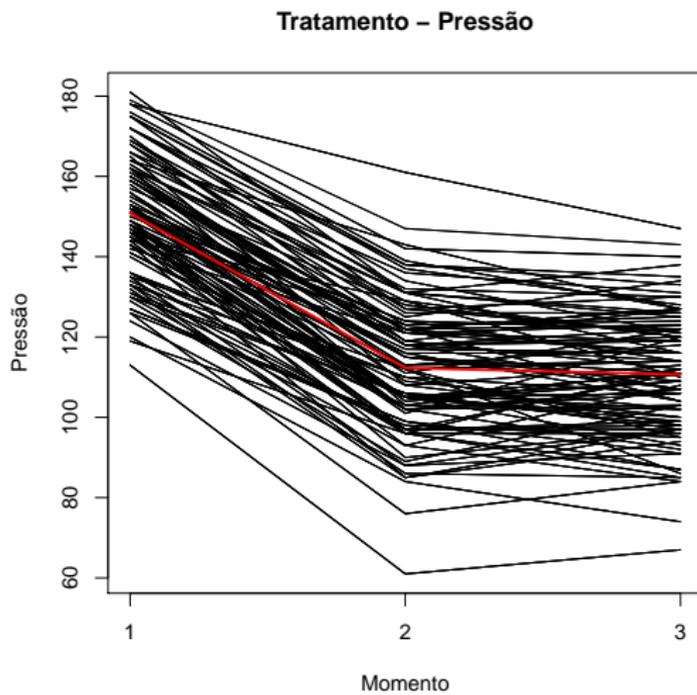
$$\mu_1 = \mu_2 = \dots = \mu_n.$$

- Identificar os grupos diferentes se H_0 for rejeitada.
- Típica situação de planejamento e experimentos. Podemos considerar que cada indivíduo é um bloco e realizar a análise usual de um fator em blocos?

Extensão: Três Medidas por Paciente

- Deseja-se verificar a eficácia de uma certa droga para reduzir a pressão arterial. 100 pacientes hipertensos participaram do estudo.
- A pressão sistólica foi medida no início (tempo 1) do estudo, 30 (tempo 2) e **60 (tempo 3)** dias após os pacientes terem sido submetidos a droga de interesse ($n = 3$).
- O objetivo é avaliar a evolução da pressão ao longo de 60 dias. Ou seja, o interesse é testar a hipótese: $H_0 : \mu_1 = \mu_2 = \mu_3$
- Dados simulados normais:
 - $E(Y_1) = 150, E(Y_2) = E(Y_3) = 110$;
 - $Var(Y_j) = 15, j = 1, 2, 3$;
 - $Cor(Y_j, Y_{j'}) = 0,8, j \neq j'$.

Perfis dos Pacientes



- 1 ANOVA para medidas repetidas;
- 2 MANOVA: análise de variância multivariada.

Análise de Variância

- É uma técnica pela qual a variabilidade total de um conjunto de dados é separada em vários componentes.
- Usualmente, cada um desses componentes de variação está associada a uma fonte específica de variação.
- Em qualquer tipo de experimento é de interesse conhecer a magnitude das contribuições de cada uma dessas fontes para a variação total.

Planejamento de Experimentos - Um fator em bloco

Objetivo: Comparar a resposta média em cada tempo.

$$Y_{ij} = \mu + \alpha_i + \tau_j + \varepsilon_{ij},$$

em que, $\varepsilon_{ij} \sim N(0, \sigma^2)$.

No nosso caso:

- Os blocos são os indivíduos.
- α_i : o efeito do bloco (indivíduo), $i = 1, \dots, N$
- α_i : pode ser tratado como efeito fixo ou aleatório. Neste último caso,

$$\alpha_i \sim N(0, \sigma_\alpha^2)$$

- Os tratamentos são os próprios tempos.
- τ_j : O efeito do tratamento (tempo), $j = 1, \dots, n$

Obs.: Não é possível aleatorizar tratamento dentro do bloco.

Tabela de Análise de Variância - ANOVA

Fonte	SQ	GL	QM	F
Trt. (Tempo)	SQ_{Trat}	$n - 1$	$SQ_{Trat}/(n - 1)$	QM_{Trat}/QM_{Res}
Bloco (Ind.)	SQ_{Bloc}	$N - 1$	$SQ_{Bloc}/(N - 1)$	QM_{Bloc}/QM_{Res}
Erro	SQ_{Res}	$(n - 1)(N - 1)$	$SQ_{Res}/(n - 1)(N - 1)$	
Total	SQ_{Total}	$Nn - 1$	$SQ_{Total}/(Nn - 1)$	

Obs.: Esta tabela ANOVA vale para os dois casos (α fixo e aleatório).

$$SQ_{Total} = \sum_{i=1}^N \sum_{j=1}^n (y_{ij} - \bar{y})^2 \quad \bar{y} = \sum_{i=1}^N \sum_{j=1}^n \frac{y_{ij}}{Nn}$$

$$SQ_{Tratamento} = N \sum_{j=1}^n (\bar{y}_j - \bar{y})^2 \quad \bar{y}_j = \sum_{i=1}^N \frac{y_{ij}}{N}$$

$$SQ_{Bloco} = n \sum_{i=1}^N (\bar{y}_i - \bar{y})^2 \quad \bar{y}_i = \sum_{j=1}^n \frac{y_{ij}}{n}$$

Sob $H_0 : \tau_1 = \dots = \tau_n$,

$$F = \frac{QM_{Trat}}{QM_{Res}} \sim F_{(n-1), (n-1)(N-1)}$$

Ajuste do Modelo - Exemplo Pressão Sistólica

```
> anova<-aov(values ~ factor(grupo)+factor(ident), data=dados4)
> summary(anova)
```

	Df	Sum Sq	Mean Sq	F	value	Pr(>F)
factor(grupo)	2	108150	54075	1358.35	<2e-16	***
factor(ident)	99	52539	531	13.33	<2e-16	***
Residuals	198	7882	40			

Obs.:

- Necessário fazer comparações múltiplas.
- $Cov(y_{ij}, y_{ij'}) = \sigma_{\alpha}^2$ e $Var(Y_{ij}) = \sigma_{\alpha}^2 + \sigma^2$ - Simetria composta.
- Simetria composta pode não ser adequada para dados longitudinais.

Resumo

- Podemos utilizar esta análise para testar a igualdade de mais de duas médias.
- O teste F vale se $Cov(Y_i) = Var((Y_{i1}, \dots, Y_{in})') = \Sigma$ em que:

$$\Sigma = \sigma^2 \begin{bmatrix} 1 & \rho & \dots & \rho \\ \rho & 1 & \dots & \rho \\ \vdots & \vdots & \ddots & \vdots \\ \rho & \rho & \dots & 1 \end{bmatrix}$$

em que, $\rho = \frac{Cov(y_{ij}, Y_{ij'})}{\sigma^2}$,

- Σ é chamada de **simetria composta ou esférica**

Teste: Simetria Composta (esférica)

Teste de Esfericidade (Teste de Mauchly)

$H_0 : \Sigma$ é esférica vs $H_1 : \Sigma$ não é esférica;

Teste da Razão de Verossimilhança

Estatística Teste:

$$W = \det(S) \left(\frac{n+1}{\text{traço}(S)} \right)^{n+1},$$

em que, (1) S : matriz de covariância amostral e (2) sob H_0 , W tem assintoticamente uma distribuição qui-quadrado com $\frac{n(n-1)}{2} - 1$ graus de liberdade.

Obs.: H_0 significa: mesma variância para todos os tempos e mesma correlação entre os diferentes tempos.

Proposta de Solução

- Se não rejeito H_0 , use o teste F e as comparações múltiplas usuais;
- Se rejeito H_0 : corrigir os g.l. e usar a Estatística F. Ou seja, utilize a mesma estatística teste F e sob H_0 , comparar com uma distribuição F com os seguintes graus de liberdade:
 - numerador : $\varepsilon(n - 1)$
 - denominador : $\varepsilon[(n - 1)(N - 1)]$

Exitem duas propostas de correção (estimar ε):

- 1 Greenhouse-Geisser (**GG**)
- 2 Huynh-Feld (**HF**)

Exemplo

```
>Mauchly Tests for Sphericity
```

```
Test statistic p-value  
rfactor 0.98065 0.38378
```

```
>Greenhouse-Geisser and Huynh-Feldt Corrections  
for Departure from Sphericity
```

```
GG eps Pr(>F[GG])  
rfactor 0.98101 < 2.2e-16 ***
```

```
> HF eps Pr(>F[HF])  
rfactor 1.000652 2.372047e-116
```

Ajuste Alternativo - Exemplo Pressão Sistólica

Modelo de Efeito Aleatório (bloco aleatório)

Linear mixed-effects model fit by REML

Random effects: Formula: 1 | ident
(Intercept) Residual
StdDev: 14.2 6.5

Term	Value	Std.Error	DF	t-value	p-value
(Intercept)	151.59	1.5626121	198	97.01064	0
factor(grupo)2	-40.30	0.9236511	198	-43.63119	0
factor(grupo)3	-41.40	0.9236511	198	-44.82212	0

Teste de Friedman (Não Paramétrico)

- É uma alternativa para a ANOVA, quando a suposição de normalidade, igualdade de variâncias ou esfericidade, não for válida.
- Use os postos dos dados ao invés de seus valores observados para obter a estatística de teste.
- Hipóteses:

$$H_0 : med_1 = med_2 = \dots = med_n$$

H_1 : existe pelo menos duas medianas diferentes

Situação: Comparar as medianas em n tempos (tratamentos) do mesmo indivíduo

Teste de Friedman(Não Paramétrico)

- Encontrar os postos para cada bloco (indivíduo) R_{ij} ;
- sob a hipótese de não haver diferença entre os tratamentos (tempos), todas as possíveis ordens ($n!$) devem ser igualmente prováveis.
- Estatística Teste

$$Q = \frac{12N}{n(n+1)} \sum_{j=1}^n (R_j - 0,5(n+1))^2$$

em que $R_j = \sum_{i=1}^N R_{ij} / N$.

Sob H_0 , tem a dist. tabelada de Friedman.

Exemplo

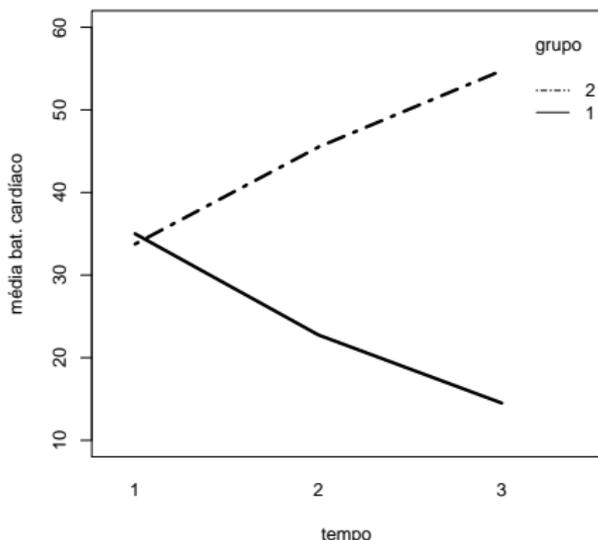
Resultados:

Teste Não-Paramétrico de Friedman

```
> friedman.test(values, grupo, ident)
Friedman rank sum test
Friedman chi-squared = 152.2424, df = 2,
p-value < 2.2e-16
```

Extensão: Comparar grupos ao longo do tempo

Exemplo: Dois grupos ao longo de Três tempos.



Variação: (1) entre grupos; (2) entre indivíduos; (3) entre tempos (intra-indivíduo; e (4) interação grupo*tempo.

Extensão: Comparar grupos ao longo do tempo

- Desenho similar ao split-plot.

```
● > demo4.aov <- aov(pulse ~ group * time + Error(id), data=demo4)
> summary(demo4.aov)
Error: id
Df Sum Sq Mean Sq F value Pr(>F)
group 1 2542.0 2542 629 2.65e-07 ***
Residuals 6 24.3 4
```

```
Error: Within
Df Sum Sq Mean Sq F value Pr(>F)
time 2 0.5 0.079 0.925
group:time 2 1736 868.2 137.079 5.44e-09 ***
Residuals 12 76 6.3
```

- Tutorial: <https://stats.idre.ucla.edu/r/seminars/repeated-measures-analysis-with-r/>

Limitações - ANOVA

- 1 Não se aplica em situações desbalanceadas;
- 2 Trata o tempo como categórico;
- 3 Usualmente a correlação tende a diminuir a medida que aumentamos a distância temporal. Ou seja, a estrutura esférica não é adequada;
- 4 Difícil (impossível?) ser utilizado na presença de covariáveis contínuas.
- 5 Resposta com distribuição Normal.

Razões Históricas - Planejamento de Experimentos

- 1 A matriz de simetria composta tem uma justificativa em termos da aleatorização em Planejamento de Experimentos.
- 2 Usualmente, não tem a dimensão temporal e, simplesmente, medidas repetidas.
- 3 Facilidade computacional em termos históricos. Basta uma calculadora para construir a ANOVA.

MANOVA - Análise Multivariada

- 1 O foco é a resposta multivariada.
- 2 Usualmente para respostas de diferente natureza.

MANOVA: é uma ANOVA multivariada para $n - 1$ diferenças entre os tempos subsequentes. A ideia básica é obter um novo conjunto de variáveis baseado em combinação linear das originais.

T^2 de Hotelling é o teste multivariado mais conhecido baseado na normal multivariada. Pode-se dizer que é o teste-t multivariado.

MANOVA tem, essencialmente, as mesmas limitações da ANOVA em relação à dados longitudinais e medidas repetidas.

MANOVA - Análise Multivariada

- 1 Usualmente, na forma proposta na literatura, as respostas multivariada são de diferentes naturezas;
- 2 Exemplo: pressão, frequência cardíaca, colesterol, etc medidas no mesmo indivíduo;
- 3 MANOVA pode ser adaptada para dados longitudinais, definindo as respostas de acordo com o objetivo do estudo;
- 4 Ver um exemplo em FLW (2011), pgs 79 a 83.

Modelagem para Dados Longitudinais - Resposta Multivariada ($n = 3$).

$$y_i \sim N_3(X_i\beta, \Omega) \quad i = 1 \dots N.$$

Modelando as Médias

$$E(Y_{i1}) = \beta_0$$

$$E(Y_{i2}) = \beta_0 + \beta_1$$

$$E(Y_{i3}) = \beta_0 + \beta_2$$

ou em termos do modelo

$$Y_{ij} = \beta_0 + \beta_1 \text{lg}1_j + \beta_2 \text{lg}2_j + \epsilon_{ij} \quad i = 1, \dots, N; j = 1, 2, 3.$$

em que $\text{lg}1_j = 1$, se $j = 2$, $\text{lg}2_j = 1$, se $j = 3$ e $= 0$, caso contrário.

Modelagem via Dados Longitudinais

E podemos tomar uma forma geral para a matriz de covariância Σ . Ou seja,

$$\begin{aligned}\epsilon_{ij} &\sim N(0, \sigma_j^2), j = 1, 2, 3; \\ \text{Cov}(\epsilon_{ij}, \epsilon_{ij'}) &= \sigma_{jj'}/2, j \neq j', j = 1, 2, 3.\end{aligned}$$

Interesse em testar $\beta_1 = \beta_2 = 0$.

Este é o **modelo marginal**, bastante utilizado em Dados Longitudinais.