Planejamento e Análise Estatística de Experimentos: um único fator

Prof. Caio Azevedo

- Em relação às questões levantadas anteriormente, uma solução é a utilização de modelos estatísticos apropriados.
- Modelos estatísticos: representações do experimento, levando em consideração características de interesse, e o comportamento estocástico dos dados.
- Estrutura geral: parte sistemática e parte aleatória.
- Modelos de regressão normais lineares (MRNL ou MNL): parte sistemática (modela a média de forma linear em função dos fatores) e parte aleatória (distribuição normal).

Suposíções

- Suponha que queremos comparar as médias de vários grupos.
- Dispomos de uma amostra de tamanho n_i para cada grupo.
- Distribuição normal com mesma variância para todas os grupos.

Modelo para um único fator com vários níveis (vários grupos): modelo de médias

$$Y_{ij} = \mu_i + \xi_{ij}, i = 1, 2, .., k$$

(grupos); $j = 1, ..., n_i$(unidades experimentais)

- Erros $\xi_{ij} \overset{i.i.d}{\sim} N(0, \sigma^2)$, μ_i não aleatório.
- $\mathcal{E}_{\xi_{ij}}(Y_{ij}) = \mu_i, \mathcal{V}_{\xi_{ij}}(Y_{ij}) = \sigma^2$.
- μ_i : média populacional relacionada ao i-ésimo fator.
- $Y_{ij} \overset{ind.}{\sim} N(\mu_i, \sigma^2)$.

Estimação (método dos mínimos quadrados)

Defina $Q(\boldsymbol{\mu}) = \sum_{i=1}^{k} \sum_{j=1}^{n_i} \xi_{ij}^2 = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij} - \mu_i)^2$ (soma dos quadrados dos erros) $\boldsymbol{\mu} = (\mu_1, ..., \mu_k)$.

Objetivo: minimizar $Q(\boldsymbol{\mu})$ em $\boldsymbol{\mu}$. Assim, derivando-se com relação à cada μ_i, tem-se as equações normais:

$$\frac{\partial}{\partial \mu_i} Q(\boldsymbol{\mu}) = -2 \sum_{j=1}^{n_i} (y_{ij} - \mu_i) = 0$$

Precisamos resolver as k equações em função de $\mu_i, i = 1, 2, ..., k$, ou seja

$$-2 \sum_{j=1}^{n_i} (y_{ij} - \widehat{\mu}_i) = 0$$

Estimadores

- $\widehat{\mu}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} Y_{ij} = \overline{Y}_{i.}$. Neste caso $\widehat{\mu}_i \sim N(\mu_i, \sigma^2/n_i)$.
- Estimador não viciado para σ^2,

$$\widehat{\sigma}^2 = \frac{1}{n-k}\sum_{i=1}^{k}\sum_{j=1}^{n_i}(Y_{ij} - \widehat{\mu}_i)^2 = \frac{1}{n-k}\sum_{i=1}^{k}\sum_{j=1}^{n_i}(Y_{ij} - \overline{Y}_i)^2,$$

  $n = \sum_{i=1}^{k} n_i$. Além disso, $\frac{(n-k)\widehat{\sigma}^2}{\sigma^2} \sim \chi^2_{(n-k)}$.
- Note que $\widehat{\sigma}^2 = QMR = SQR/(n-k)$.

Tabela de análise de variância

- Para testar a igualdade simultânea das médias

FV	SQ	GL	QM	Estatística F	pvalor
Fatores	SQF	k-1	QMF=SQF/(k-1)	F=QMF/QMR	$\min(P(F > f\|H_0), P(F < f\|H_0)$
Resíduo	SQR	n-k	QMR=SQR/(n-k)		
Total	SQT	n-1			

FV: fonte de variação, SQ: soma de quadrados, Gl: graus de liberdade,

QM: quadrado médio.

Indicações para análise de um planejamento com um único fator

1. Análise descritiva.
2. Propor e ajustar o modelo (estimar os parâmetros).
3. Verificar as suposições do modelo.
4. Análise de variância.
5. Testar sub-hipóteses de interesse.
6. Ajustar modelo reduzido e apresentar os resultados.

OBS: Não se pode utilizar um modelo para analisar um determinado conjunto de dados para o qual as suposições não se verifiquem e/ou não seja robusto à ausência das suposições.

Robustez às suposições do MNL

1. Relativamente robusto à ausência de normalidade (teorema central do limite).
2. Pouco robusto à ausência de homocedasticidade.
3. Pouco robusto à ausência de independência.
4. Todos os aspectos de interesse, em princípio, devem ser considerados no modelo (modelagem da média).

Voltando ao exemplo das árvores

- Análise descritiva já foi realizada.
- Aparentemente, as suposições de normalidade e homocedasticidade não se verificam.
- Em princípio, outro modelo deveria ser utilizado.
- Transformações de variáveis não são aconselháveis: perda da interpretabilidade, interferência na natureza dos dados, estimadores com propriedades indesejáveis.

Tabela ANOVA

FV	SQ	GL	QM	Estatística F	pvalor
Tipo de árvore	1363,27	1	1363,27	15,19	0,0002
Resíduo	5203,90	58	89,72		
Total	6657,17	59			

Rejeita-se H_0.

Parametrizações do modelo

$$Y_{ij} = \mu + \alpha_i + \xi_{ij}, i = 1, 2, .., k$$

(grupos); $j = 1, ..., n_i$(unidades experimentais)

- Erros $\xi_{ij} \overset{i.i.d}{\sim} N(0, \sigma^2)$, μ, α_i não aleatório.
- $\mathcal{E}_{\xi_{ij}}(Y_{ij}) = \mu_i, \mathcal{V}_{\xi_{ij}}(Y_{ij}) = \sigma^2$.
- $\mu + \alpha_i$: média populacional relacionada ao i-ésimo fator.
- $Y_{ij} \overset{ind.}{\sim} N(\mu + \alpha_i, \sigma^2)$.

Problema: tem-se k médias e $k + 1$ parmâmetros (falta de identificabilidade)

Parametrizações dentro da estrutura admitida

- Desvios sem restrição: não se coloca nenhuma restrição. Neste caso, somente as funções estimáveis podem ser estimadas.
- Desvios com restrição: $\sum_{i=1}^{k} \alpha_i = 0$. Modelo identificado.
- Casela (cela de referência): igualar um único α_i à 0, por exemplo α_1. Modelo identificado.
- Para interpretação dos parâmetros basta lembrar que $\mu_i = \mu + \alpha_i$

Interpretações dos modelos (parametrização)

- Desvios com restrição: $\sum_{i=1}^{k} \alpha_i = 0$.
  - μ : média das médias de cada os grupo $\mu = \overline{\mu} = \frac{1}{k} \sum_{i=1}^{k} \mu_i$.
  - α_i : incremento (positivo ou negativo) da média do grupo i com relação à média das médias, $\alpha_i = \mu_i - \overline{\mu}$.

Interpretações dos modelos (parametrização)

- Casela de referência: $\alpha_1 = 0$.
  - μ : média do grupo 1 (grupo de referência), $\mu = \mu_1$.
  - α_i : incremento (positivo ou negativo) da média do grupo i com relação à média do grupo 1 (grupo de referência), $\alpha_i = \mu_i - \mu_1$.
- Em geral, esta será a parametrização utilizada no curso.

Utilidades das parametrizações

- Facilidade de interpretação dos parâmetros de interesse (principalmente na presença de mais de um fator e planejamentos mais complexos).
- Testes de hipótese mais simples. Casela de referência/desvios com restrição (hipótese de igualdade de médias)

$$H_0 : \alpha_2 = \alpha_3 = ... = \alpha_k = 0$$

- Mais facilidade na identificação dos efeitos significativos (interação).

Estimadores de mínimos quadrados (casela de referência)

- Pode-se provar que $\widehat{\mu} = \overline{Y}_{1.}$, $\widehat{\alpha}_i = \overline{Y}_{i.} - \overline{Y}_{1.}$.
- O estimador para σ^2 permanece o mesmo.
- Tem-se ainda que $\mathcal{V}(\widehat{\mu}) = \frac{\sigma^2}{n_1}$ e $\mathcal{V}(\widehat{\alpha}_i) = \sigma^2 \left(\frac{1}{n_i} + \frac{1}{n_1} \right)$.
- Além disso $\widehat{\alpha}_i \sim N\left(\alpha_i, \sigma^2 \left(\frac{1}{n_i} + \frac{1}{n_1} \right) \right)$

Voltando ao exemplo das árvores

- Estimativas (erros-padrão) : $\mu = 12,77(1,77)$, $\alpha_2 = 9,53(2,44)$.
- Teste t usual para nulidade de cada parâmetros (pvalor)

  $\mu(\leq 0,0001)$, $\alpha_2(0,0002)$.
- Hipótese de igualdade de médias $H_0 : \alpha_2 = 0$, rejeitada.

Revisão normal multivariada

- Dizemos que $Y = (Y_1, ..., Y_p) \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ se sua fdp é dada por

$$f_Y(y) = |\boldsymbol{\Sigma}|^{-1/2}(2\pi)^{-p/2} \exp\left\{-\frac{1}{2}(y-\boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1}(y-\boldsymbol{\mu})\right\} \mathbb{1}_{\mathcal{R}^p}(y)$$

- $\boldsymbol{\mu}$ é o vetor de médias e $\boldsymbol{\Sigma}$ é a matriz de covariâncias.

Parâmetros

- $\boldsymbol{\mu} = \mathcal{E}(Y) = \begin{bmatrix} \mathcal{E}(Y_1) \\ \mathcal{E}(Y_2) \\ \vdots \\ \mathcal{E}(Y_p) \end{bmatrix} = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_p \end{bmatrix}$
- $\boldsymbol{\Sigma} = Cov(Y) = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1p} \\ \sigma_{12} & \sigma_2^2 & \cdots & \sigma_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{1p} & \sigma_{2p} & \cdots & \sigma_p^2 \end{bmatrix}$

Propiedades

- Fechada sob marginalização: $Y_i \sim N(\mu_i, \sigma_i^2)$.
- $Y_i \perp Y_j, \forall i \neq j \Leftrightarrow \sigma_{ij} = 0$.
- Se $A_{(q\times p)}$ for uma matriz não aleatória, então
  $V = AY \sim N_q(A\boldsymbol{\mu}, A\boldsymbol{\Sigma}A')$.
- Se $A_{(p\times p)}$ for uma matriz não aleatória, simétrica e idempotente de rank = p, $\boldsymbol{\mu} = \mathbf{0}$ e $\boldsymbol{\Sigma} = \sigma^2 I_{(p\times p)}$, então
  $V = \frac{1}{\sigma^2} Y'AY \sim \chi^2_r, r = tr(A)$.
- Se $A_{(p\times p)}$ for uma matriz não aleatória, então
  $\mathcal{E}(Y'AY) = tr(A\boldsymbol{\Sigma}) + \boldsymbol{\mu}'A\boldsymbol{\mu}$.

Propiedades (cont.)

- Se $A_{(p\times p)}$ for uma matriz não aleatória, simétrica e idempotente de rank = p, $\boldsymbol{\mu} = \mathbf{0}$, então $V = Y'AY \sim \chi^2_r, r = tr(A) \Leftrightarrow A\boldsymbol{\Sigma}A = A$.
- Sejam $A_{(q\times p)}$ e $B_{(s\times p)}$ matrizes não aleatórias e $V = AY$ e $W = BY$. Então $V \perp W \Leftrightarrow A\boldsymbol{\Sigma}B' = \mathbf{0}$.
- Sejam $A_{(p\times p)}$ e $B_{(p\times p)}$ matrizes não aleatórias e simétricas e $V = Y'AY$ e $W = Y'BY$. Então $V \perp W \Leftrightarrow A\boldsymbol{\Sigma}B = \mathbf{0}$.

Notação matricial para o MRNL

- Considere o modelo de médias $Y_{ij} = \mu_i + \xi_{ij}$. Note que

$$
\begin{aligned}
Y_{11} &= \mu_1 + \xi_{11} \\
Y_{12} &= \mu_1 + \xi_{12} \\
&\vdots \\
Y_{1n_1} &= \mu_1 + \xi_{1n_1} \\
Y_{21} &= \mu_2 + \xi_{21} \\
Y_{22} &= \mu_2 + \xi_{22} \\
&\vdots \\
Y_{2n_2} &= \mu_2 + \xi_{2n_2} \\
&\vdots \\
Y_{k1} &= \mu_k + \xi_{k1} \\
&\vdots \\
Y_{kn_k} &= \mu_k + \xi_{kn_k}
\end{aligned}
$$

Notação matricial para o MRNL

- Assim, o modelo anterior pode ser escrito na seguinte estrutura matricial:

$$Y = X\beta + \xi$$

$$Y = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}, X = \begin{bmatrix} X_{11} & \cdots & X_{1p} \\ X_{21} & \cdots & X_{2p} \\ \vdots & \ddots & \vdots \\ X_{n1} & \cdots & X_{np} \end{bmatrix}, \beta = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{bmatrix}, \xi = \begin{bmatrix} \xi_1 \\ \xi_2 \\ \vdots \\ \xi_n \end{bmatrix}$$

- Suposição $\xi \sim N_n(\mathbf{0}, \sigma^2 I_n)$.
- Estimador de mínimos quadrados de β: minimizar $(Y - X\beta)'(Y - X\beta)$, em β.

Cont.

- Solução $\widehat{\boldsymbol{\beta}} = (X'X)^{-1}X'Y$. Em geral $X'X$ é positiva definida.
- Das suposições do modelo, temos que $\widehat{\boldsymbol{\beta}} \sim N_p(\boldsymbol{\beta}, \sigma^2 \left(X'X\right)^{-1})$.
- Assim $\widehat{\boldsymbol{\beta}}_i \sim N(\boldsymbol{\beta}_i, \sigma^2\sigma_{\boldsymbol{\beta}_i}), i = 1, ..., p$, em que $\sigma_{\boldsymbol{\beta}_i}$ é o i-ésimo elemento da diagonal principal de $\left(X'X\right)^{-1}$.

Modelo de médias : matriz de planejamento

$$X = \begin{bmatrix} 1 & 0 & \dots & 0 \\ 1 & 0 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 0 & \dots & 0 \\ 0 & 1 & 0\dots & 0 \\ 0 & 1 & 0\dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 1 & 0\dots & 0 \\ 0 & 0 & 0\dots & 1 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0\dots & 1 \end{bmatrix}; \boldsymbol{\beta} = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_k \end{bmatrix}$$

Casela de referência : matriz de planejamento

$$X = \begin{bmatrix} 1 & 0 & \dots & 0 \\ 1 & 0 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 0 & \dots & 0 \\ 1 & 1 & 0\dots & 0 \\ 1 & 1 & 0\dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & 0\dots & 0 \\ 1 & 0 & 0\dots & 1 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 0 & 0\dots & 1 \end{bmatrix}; \boldsymbol{\beta} = \begin{bmatrix} \mu \\ \alpha_2 \\ \vdots \\ \alpha_k \end{bmatrix}$$

EMQ : modelo de médias

- Modelo de médias

$$\widehat{\boldsymbol{\beta}} = \begin{bmatrix} \overline{Y}_{1.} \\ \overline{Y}_{2.} \\ \vdots \\ \overline{Y}_{k.} \end{bmatrix}$$

- Casela de referência

$$\widehat{\boldsymbol{\beta}} = \begin{bmatrix} \overline{Y}_{1.} \\ \overline{Y}_{2.} - \overline{Y}_{1.} \\ \vdots \\ \overline{Y}_{k.} - \overline{Y}_{1.} \end{bmatrix}$$

Somas de quadrados do resíduo

- Erro $\boldsymbol{\xi} = Y - X\boldsymbol{\beta}$. Valor predito $Y = X\widehat{\boldsymbol{\beta}}$.
- Resíduo (valor predito para o erro)
  $\widehat{\boldsymbol{\xi}} = Y - X\widehat{\beta} = Y - X\left(X'X\right)^{-1}XY = \left[I - H\right]Y$.
  $H = X\left(X'X\right)^{-1}X$ (matriz "hat" ou matriz de projeção).
- SQR $= \widehat{\boldsymbol{\xi}}'\widehat{\boldsymbol{\xi}} = Y'\left[I - H\right]\left[I - H\right]Y = Y'AY$.
- A matriz $A = I - H$ é simétrica, idempotente de posto=traço = n-k (p=k).
- SQR: é a parte da variabilidade não explicada pelo modelo (devida à outros fatores).

Somas de quadrados do modelo

- Considere o modelo $Y_{ij} = \mu + \xi_{ij}$, ou seja $Y = X^*\beta^* + \boldsymbol{\xi}$ (modelo a ser testado).
- Note que $\beta^* = \mu$ e $X^* = \mathbf{1}_n = \mathbf{1}, n = \sum_{i=1}^{k} n_i$.
- $SQR^* = Y' \left[I - \mathbf{1} \left(\mathbf{1}'\mathbf{1} \right)^{-1} \mathbf{1}' \right] Y = Y' \left[I - n^{-1} J \right] Y$, $J = \mathbf{1}\mathbf{1}'$.
- Objetivo na comparação de médias: medir $SQF = SQR^* - SQR$ (soma de quadrados dos fatores ou do modelo) (exercício).

Cont.

- Portanto

$$SQF = Y' \left[I - n^{-1} J \right] Y - Y'[I - H]Y = Y' \left[H - n^{-1} J \right] Y.$$

- Fato: independentemente da parametrização (modelo de médias, casela de referência e desvios com restrição), o valor predito é dado por

$$\widehat{Y} = X\widehat{\beta} = \begin{bmatrix} \overline{Y}_{1.} \\ \vdots \\ \overline{Y}_{1.} \\ \overline{Y}_{2.} \\ \vdots \\ \overline{Y}_{2.} \\ \overline{Y}_{k.} \\ \vdots \\ \overline{Y}_{k.} \end{bmatrix}$$

Cont.

- Ou seja:

$$H = X\left(X'X\right)^{-1}X' = \begin{bmatrix} n_1^{-1}\mathbf{1}'_{n_1} & \mathbf{0}'_{n_2} & \dots & \mathbf{0}'_{n_k} \\ \vdots & \vdots & \ddots & \vdots \\ n_1^{-1}\mathbf{1}'_{n_1} & \mathbf{0}'_{n_2} & \dots & \mathbf{0}'_{n_k} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0}'_{n_1} & n_2^{-1}\mathbf{1}'_{n_2} & \dots & \mathbf{0}'_{n_k} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0}'_{n_1} & n_2^{-1}\mathbf{1}'_{n_2} & \dots & \mathbf{0}'_{n_k} \\ \vdots & \vdots & \vdots & \vdots \\ \mathbf{0}'_{n_1} & \mathbf{0}'_{n_2} & \dots & n_k^{-1}\mathbf{1}'_{n_k} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0}'_{n_1} & \mathbf{0}'_{n_2} & \dots & n_k^{-1}\mathbf{1}'_{n_k} \end{bmatrix}$$

Cont.

- Portanto: $n^{-1}HJ = n^{-1}J$ (exercício). Seja $B = \left[H - n^{-1}J\right]$
- Além disso, se $P = n^{-1}J$, temos que $PP = P$.
- Logo $BB = [H - HP - HP + PP] = [H - P] = B$.
- Assim, B é uma matriz simétrica, idempotente de posto $k - 1$.
- Além disso, $AB = [H - H - P + PH] = \mathbf{0}_{n \times n}$.

Anova matricial: resumo

- Tem-se que $SQF = Y'BY$ e $SQR = Y'AY$, em que $B = H - n^{-1}J$ e $A = I - H$.
- Sob H_0, $Y_{ij} \sim N(\mu, \sigma^2)$ (modelo reduzido).
- Assim, sob H_0, $SQF/\sigma^2 \sim \chi^2_{(k-1)}$, $SQR/\sigma^2 \sim \chi^2_{(n-k)}$ e $SQF \perp SQR$.
- Logo, sob H_0, $F = \frac{SQF/(k-1)}{SQR/(n-k)} \sim F_{(k-1,n-k)}$.

Tabela de ANOVA (matricial)

- Para testar a igualdade simultânea das médias

FV	SQ	GL	QM	Estatística F	pvalor
Fatores	$SQF = Y'BY$	k-1	$QMF = \frac{SQF}{k-1}$	$F = \frac{QMF}{QMR}$	$\min(P(F > f\|H_0), P(F < f\|H_0)$
Resíduo	$SQR = Y'AY$	n-k	$QMR = \frac{SQR}{n-k}$		
Total	SQT	n-1			

FV: fonte de variação, SQ: soma de quadrados, Gl: graus de liberdade,

QM: quadrado médio.

Em termos dos parâmetros

- Suponha que queremos testar $H_0 : \beta_i = \beta_{0i}$, vs $H_1 : \beta_i \neq \beta_{0i}$, β_{0i} conhecido.
- Lembrando que $\widehat{\boldsymbol{\beta}} \sim N_p(\boldsymbol{\beta}, \sigma^2 \left(X'X\right)^{-1})$ e $\widehat{\boldsymbol{\beta}}_i \sim N(\boldsymbol{\beta}_i, \sigma^2 \sigma_{\boldsymbol{\beta}_i}), i = 1, ..., p$, em que $\sigma_{\boldsymbol{\beta}_i}$ é o i-ésimo elemento da diagonal principal de $\left(X'X\right)^{-1}$.
- Assim, tem-se que, sob H_0

$$T = \frac{\widehat{\beta}_i - \beta_{0i}}{\sqrt{\widehat{\sigma}^2 \left(\sigma_{\boldsymbol{\beta}_i}\right)}} \sim t_{(n-k)}$$

em que $\widehat{\sigma}^2 = QMR = \frac{SQR}{n-k}$.

- Para $n - k$ suficientemente grande, temos que $T \approx N(0, 1)$.