Revista da Estatística UFOP, Vol III(3), 2014, ISSN 2237-8111
Edição especial: 59ª Reunião Anual da Regional Brasileira da Sociedade Internacional de Biometria - RBRAS
Trabalhos apresentados na versão poster.

Abordagem geométrica para imputação de dados em modelos lineares

Leandro da Silva Pereira [1]
Lucas Monteiro Chaves [2]
Devanil Jaques Souza [3]
[4]

1 Introdução

Experimentos em que se tem perda de parcelas ou dados faltantes são bastante comuns. Nestas situações, é usual se imputar o valor destes. Este trabalho apresenta uma interpretação geométrica, em termos de subespaços vetoriais e projeções ortogonais, para um método de imputação de dados, proposto por Kruskall(1961).

2 Material e métodos

Um conjunto de dados com n observações pode ser entendido como a realização de um vetor aleatório em $\mathbb{R}^n$, denotado por $\mathbf{Y} = (\mathrm{Y}_1, ..., \mathrm{Y}_n)'$. Suponha, sem perda de generalidade, que as primeiras m coordenadas de $\mathbf{Y}$ sejam observáveis e as restantes $n-m$ sejam não observáveis ou faltantes. Desta forma se tem $\mathbf{Y} = \mathbf{Y}^1 + \mathbf{Y}^2$ em que $\mathbf{Y}^1$ é a parte observável e $\mathbf{Y}^2$ a não observável. Para o modelo $\mathbf{Y}_{n\times 1} = \mathbf{X}_{n\times p}\beta_{p\times 1} + \varepsilon_{n\times 1}$ que leva em conta todos os dados, a matriz $\mathbf{X}$ pode ser vista como a transformação linear $\mathbf{X} : \mathbb{R}^p \mapsto \mathbb{R}^n = \mathbb{R}^m \oplus \mathbb{R}^{n-m}$. Se $Im(\mathbf{X})$ é o subespaço definido pela imagem da transformação e $\mathrm{P_w}$ é a projeção ortogonal em um subespaço W, ficam definido os subespaços $\Omega_1 = \mathrm{P}_{\mathbb{R}^n}(\mathrm{Im(X)})$ e $\Omega_2 = \mathrm{P}_{\mathbb{R}^{n-m}}(\mathrm{Im(X)})$. Observe que, em geral, $\mathrm{Im(X)} \neq \Omega_1 + \Omega_2$. O que se tem é que $\mathrm{Im(X)} \subset \Omega_1 + \Omega_2$. Vamos considerar a seguinte hipótese: $dim(\mathrm{Im(X)}) = dim(\Omega_1)$. Estatisticamente, tal fato significa que o número de parâmetros do modelo, mesmo com a ocorrência de dados faltantes, não se altera. $\mathrm{P}_{\mathbb{R}^m} : \mathrm{Im(X)} \mapsto \Omega_1$ é linear e sobrejetiva. $\mathrm{P}_{\mathbb{R}^m} : \mathrm{Im(X)} \mapsto \Omega_1$ Como $dim(\mathrm{Im(X)}) = dim(\Omega_1)$ então a projeção também é injetiva, isto é, $ker(\mathrm{P}_{\mathbb{R}^m})$ restrito à $\mathrm{Im(X)}$ é o subespaço nulo. Logo, se

[1]DEX-UFLA. e-mail: *lspleandro_2@hotmail.com.br*
[2]DEX - UFLA. e-mail: *lucas@dex.ufla.br*
[3]DEX - UFLA. e-mail: *devaniljaques@dex.ufla.br*
[4]Agradecimento à FAPEMIG pelo apoio financeiro.

$\mathbf{v}_1 + \mathbf{v}_2 \in \mathrm{Im(X)}$ e $\mathbf{v}_1 + \mathbf{v'}_2 \in \mathrm{Im(X)}$ então

$$\begin{aligned} (\mathbf{v}_1 + \mathbf{v}_2) - (\mathbf{v}_1 + \mathbf{v'}_2) &\in ker(\mathrm{P}_{\mathbb{R}^m}) \\ \mathbf{v}_2 - \mathbf{v'}_2 &= 0 \\ \mathbf{v}_2 &= \mathbf{v'}_2, \end{aligned}$$

isto é, para cada vetor $\mathbf{v}_1$ em Ω_1 existe um único vetor $\mathbf{v}_2$ em Ω_2 tal que $\mathbf{v}_1 + \mathbf{v}_2 \in \mathrm{Im(X)}$, e portanto o *kernel* da projeção pertence à $\mathrm{Im(X)}$. Pode-se então definir uma transformação linear $\mathbf{A} : \mathbb{R}^{\mathbf{n}} \mapsto \mathbb{R}^n$ tal que $\mathbf{A}(\mathbf{v}_1) = \mathbf{v}_2$ e $\mathbf{A}(\mathbf{x}) = \mathbf{0}$ se $\mathbf{x} \in \Omega_1^{\perp}$.

Seja a transformação linear $\mathbf{I} : \Omega_1 \mapsto \mathbb{R}^n$, inclusão, isto é, $\mathbf{I}(\mathbf{v}_1) - \mathbf{v}_1$, $\mathbf{v}_1 \in \Omega_1$. Desta forma, pode-se escrever, utilizando um certo abuso de notação, que $\mathrm{Im(X)} = (\mathbf{I} + \mathbf{A})\Omega_1$.

Se μ é o vetor de médias geral, seu estimador de Gauss-Markov é $\hat{\mu} = \mathrm{P}_{\mathrm{Im(X)}}(\mathbf{Y})$. Como as m últimas componentes de $\mathbf{Y}$ são dados faltantes, não é possível obter $\hat{\mu}$. A estratégia então é obter o estimador de Gauss-Markov para os dados observados utilizando o sub-modelo dado por Ω_1, isto é, $\hat{\mu}^1 = \mathrm{P}_{\mathrm{Im(X)}}(\mathbf{Y}^1)$, sendo $\mu^1 = \mathrm{E}[\mathbf{Y}^1]$. Para $\hat{\mu}^1$ existe um único $\hat{\mu}^2$ dado por $\hat{\mu}^2 = \mathbf{A}(\hat{\mu}^1)$. Imputando-se o valor $\hat{\mu}^2$ obtém-se os dados com imputação $\mathbf{Y}^1 + \hat{\mu}^2$. Toma-se agora como estimador de μ o vetor obtido por

$$\hat{\mu} = \mathrm{P}_{\mathrm{Im(X)}}(\mathbf{Y}^1 + \hat{\mu}^2) \tag{1}$$

A representação destes vetores e suas projeções nos subespaços discutidos podem ser observados na Figura 1.

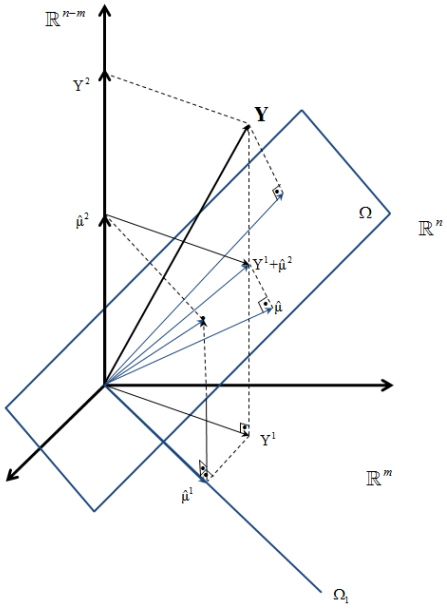

Figura 1: Projeções para dados faltantes

3 Resultados e discussão

A abordagem geométrica permite demonstrar de maneira trivial o resultado.

Proposição 1. *O processo de imputação (1) diminui a soma de quadrado dos erros.*

Demonstração. No subespaço Im(X), o ponto mais próximo de $\left(\mathbf{Y}^1+\hat{\mu}^2\right)$ é dado pela projeção ortogonal $\hat{\mu}=\mathrm{P}_{\mathrm{Im(X)}}\left(\mathbf{Y}^1+\hat{\mu}^2\right)$. Note que $\hat{\mu}^1+\hat{\mu}^2\in\mathrm{Im(X)}$.

Logo,
$\left\| \underbrace{\left(\mathbf{Y}^1+\hat{\mu}^2\right)}_{\textit{dados com imputacão}} - \underbrace{\left(\hat{\mu}^2+\hat{\mu}^1\right)}_{\in \mathrm{Im(X)}} \right\| = \left\|\mathbf{Y}^1-\hat{\mu}^1\right\|$. Assim, qualquer ponto que pertença à Im(X) e que seja diferente de $\hat{\mu}$ terá uma distância maior até o vetor $\left(\mathbf{Y}^1+\hat{\mu}^2\right)$, o que implica em

$$\left\|\left(\mathbf{Y}^1+\hat{\mu}^2\right)-\hat{\mu}=\mathrm{P}_{\mathrm{Im(X)}}\left(\mathbf{Y}^1+\hat{\mu}^1\right)\right\| \leq \left\|\mathbf{Y}^1-\hat{\mu}^1\right\|.$$

□

Exemplo de aplicação: delineamento inteiramente casualizado (DIC) com 3 tratamentos e 4 repetições $y_{ij}=\mu+\tau_i+\varepsilon_{ij}$, com $i=\{1,2,3\}$ e $j=\{1,2,3,4\}$. O vetor de dados

$$\mathbf{y}=(\mathrm{y}_{11},\mathrm{y}_{12},...,\mathrm{y}_{14},\mathrm{y}_{21},...,\mathrm{y}_{24},\mathrm{y}_{31},...,\mathrm{y}_{34})'.$$

Suponha a última observação perdida. O vetor de dados observado

$$\mathbf{y}^1=(\mathrm{y}_{11},\mathrm{y}_{12},...,\mathrm{y}_{14},\mathrm{y}_{21},...,\mathrm{y}_{24},\mathrm{y}_{31},...,\mathrm{y}_{33},0)'.$$

Tem-se

$$\mathrm{Im(X)}=\left\{(a,a,a,a,b,b,b,b,c,c,c,c)',a,b,c\in\mathbb{R}\right\},$$
$$\Omega_1=\left\{(a,a,a,a,b,b,b,b,c,c,c,0)',a,b,c\in\mathbb{R}\right\},$$
$$\Omega_2=\left\{(0,0,0,0,0,0,0,0,0,0,0,d)',d\in\mathbb{R}\right\}.$$

A transformação linear natural para este problema é $\mathbf{A}:\Omega_1\mapsto\Omega_2$.
$\mathbf{A}\left((a,a,a,a,b,b,b,b,c,c,c,0)'\right)=(0,0,0,0,0,0,0,0,0,0,0,c)'$.
$\bar{y}_{1\cdot}=\frac{\mathrm{y}_{11}+\cdots+\mathrm{y}_{14}}{4}$, $\bar{y}_{2\cdot}=\frac{\mathrm{y}_{21}+\cdots+\mathrm{y}_{24}}{4}$ e $\bar{y}'_3=\frac{\mathrm{y}_{31}+\cdots+\mathrm{y}_{33}}{3}$,

$$\mathrm{P}_{\Omega_1}\left(\mathbf{y}^1\right)=\left(\overbrace{\bar{y}_{1\cdot},...,\bar{y}_{1\cdot}}^{4},\overbrace{\bar{y}_{2\cdot},...,\bar{y}_{2\cdot}}^{4},\overbrace{\bar{y}'_3,...,\bar{y}'_3}^{3},0\right)'=\hat{\mu}^1.$$

$\mathrm{A}\left(\hat{\mu}^1\right)=\left(0,0,0,0,0,0,0,0,0,0,0,\bar{y}'_3\right)'=\hat{\mu}^2$.

O vetor de dados com imputação

$$\begin{aligned}\mathbf{y}^* &= \hat{\mu}^2 + \mathbf{y}^1 \\ &= (0,0,...,0,\bar{y}_{3\cdot})' + (y_{11},...,y_{14},y_{21},...,y_{24},y_{31},...,y_{33},0)' \\ &= \left(y_{11},...,y_{14},y_{21},...,y_{24},y_{31},...,y_{33},\bar{y}_3'\right)'.\end{aligned}$$

Portanto, $P_{Im(X)}\mathbf{y}^* = \left(\overbrace{\bar{y}_{1\cdot},...,\bar{y}_{1\cdot}}^{4},\overbrace{\bar{y}_{2\cdot},...,\bar{y}_{2\cdot}}^{4},\overbrace{\alpha,...,\alpha}^{4}\right)' = \hat{\mu}$, em que

$$\begin{aligned}\alpha &= \frac{y_{31}+y_{32}+y_{33}+\frac{y_{31}+y_{32}+y_{33}}{3}}{4} \\ &= \frac{\frac{3(y_{31}+y_{32}+y_{33})+(y_{31}+y_{32}+y_{33})}{3}}{4} \\ &= \frac{y_{31}+y_{32}+y_{33}}{3}.\end{aligned}$$

4 Conclusões

O método geométrico para abordagem de dados faltantes em modelos lineares é intuitivo e natural, permitindo uma manipulação eficiente de soma de quadrados.

Referências

[1] SAVILLE, D. J; WOOD, G. L. **Statistical Methods: The geometric aproach**. New York: Springer-Verlag,1991. 560 p.

[2] KRUSKAL, W. **The coordinate-free approach to Gauss-Markov estimation and its aplication to missing and extra observations.**. Proc. Fourth Berkeley Symp. Math. Statist. Probab., v.1, p.435-451, 1961.

[3] SILVEIRA, F. G; COSTA, L. A; PEREIRA, L. S.; et. al.; **Uma demonstração geométrica para uma identidade de Fisher para o modelo de dois fatores**. Revista Brasileira de Biometria. v. 30, n. 2, p. 199-222, 2012.