应用概率统计　第 34 卷
第 1 期　2018 年 2 月

Chinese Journal of Applied Probability and Statistics
Feb. 2018, Vol. 34, No. 1, pp. 95-100
doi: 10.3969/j.issn.1001-4268.2018.01.008

Panel Data 模型基于 Within 估计的最优设计 *

程　靖*
(安徽农业大学理学院统计系, 合肥, 230036)

岳荣先
(上海师范大学数理学院, 上海, 200234)

摘　要: 本文研究二元 Panel Data 模型基于 Within 估计在恒等设计类中的 D-, A-, I- 和 E- 最优设计. 文章首先证明了二元 Panel Data 模型基于 Within 估计在恒等设计类中的几类最优设计可在设计域的顶点处获得; 研究发现模型的 D-, A- 和 E- 最优设计均为设计域顶点上的等权重设计, I- 最优设计为对称设计; 结果还表明等权重设计具有较高的 I- 效.

关键词: Panel Data 模型; Within 估计, 最优设计

中图分类号: O212.6

英文引用格式: Cheng J, Yue R X. Optimal designs of panel data model based on within estimator [J]. Chinese J. Appl. Probab. Statist., 2018, 34(1): 95–100. (in Chinese)

§1. 引　言

在经济学、管理学、社会学、心理学等科学领域的科学研究中, Panel Data 模型 (面板数据模型) 常被用于对一个总体中给定样本在一段时间进行多重观测所得的数据进行统计分析. 由于Panel Data 模型中随机效应项的存在, 很多情况下对模型未知参数的估计比较复杂, 在实际应用中最佳线性无偏估计很难实现. 通常的处理方法有两种: 一种是构造两步估计 [1], 另一种就是设法获得 Panel Data 模型不含未知参数的估计量 [2], Within 估计就是其中影响比较大、使用比较广的估计量. 一方面 Within 估计保证了估计量中不含有未知参数; 一方面 Within 估计消除了不可观测的个体效应, 是无偏和一致的; 同时它是 Panel Data 模型一个子模型的最佳线性无偏估计 (BLUE). 因此 Within 估计在计量经济学中应用广泛, 本文也将基于 Within 估计来研究二元 Panel Data 模型的最优设计问题.

我们知道未知参数估计的精确度极大依赖于获取数据的试验方案的设计, 目前专门研究 Panel Data 模型的最优设计的成果不多, 但对一般随机系数回归模型的最优设计研究成果已有不少, 如文献 [3–7]. 这些现有研究成果中大多数情况下都是假定随机效应项的方差是已知的, 所获得的最优设计很多情况下也依赖于随机效应项的方差值. 事实上, 随机效应项的方差通常是无法获知的, 因此相应结论还有一定的局限性.

本文考虑二元 Panel Data 模型基于 Within 估计在恒等设计类中的最优设计, 由于 Within 估计消除了未知的个体效应, 不依赖于随机效应项的方差, 由此避免了最优设计对

*国家自然科学基金项目 (批准号: 11401056、11471216)、安徽省环保厅项目 (批准号: 2015-5) 和安徽农业大学稳定与引进人才项目 (批准号: yj2015-27) 资助.

*通讯作者, E-mail: chengjing@ahau.edu.cn.

本文 2016 年 9 月 23 日收到, 2017 年 5 月 5 日收到修改稿.

随机效应项的依赖. 文章首先证明了恒等设计类中二元 Panel Data 模型基于 Within 估计的几类经典最优设计均可在设计域的顶点上获得, 进而文章分别研究了二元 Panel Data 模型基于 Within 估计在恒等设计类中的 D-, A-, I- 和 E- 最优设计. 结果表明 D-, A- 和 E- 最优设计均为设计域顶点上的等权重设计, 而 I- 最优设计为对称设计, 并且等权重设计具有较高的 I- 效.

§2. Panel Data 模型的 Within 估计

Panel Data 模型的一般形式为

$$y_{it}=\beta_0+x_{it1}\beta_1+x_{it2}\beta_2+\cdots+x_{itk}\beta_k+\mu_i+e_{it},\qquad i=1,2,\ldots,N;\ t=1,2,\ldots,T. \tag{1}$$

这里 y_{it} 表示第 i 个个体在 t 时刻的观测值, x_{itj} 表示第 i 个个体上第 j 个自变量在 t 时刻的取值, $\beta_1,\beta_2,\ldots,\beta_k$ 为待参数, μ_i 为第 i 个个体效应. 通常这 N 个个体是从一个大的总体中随机抽取的, 个体效应是随机的. e_{it} 为随机误差, 假定所有的 μ_i 和 e_{it} 都互不相关, 且 $\mu_i\sim(\mu,\sigma_\mu^2)$, $e_{it}\sim(0,\sigma_e^2)$, 其中 σ_μ^2 和 σ_e^2 未知.

记

$$Y=\begin{pmatrix} y_{11}\\ \vdots\\ y_{1T}\\ \vdots\\ y_{N1}\\ \vdots\\ y_{NT}\end{pmatrix},\qquad X=\begin{pmatrix} x_{11}\\ \vdots\\ x_{1T}\\ \vdots\\ x_{N1}\\ \vdots\\ x_{NT}\end{pmatrix},\qquad e=\begin{pmatrix} e_{11}\\ \vdots\\ e_{1T}\\ \vdots\\ e_{N1}\\ \vdots\\ e_{NT}\end{pmatrix},$$

$$\beta=\begin{pmatrix}\beta_1\\ \vdots\\ \beta_k\end{pmatrix},\qquad \mu=\begin{pmatrix}\mu_1\\ \vdots\\ \mu_N\end{pmatrix},\qquad x_{it}=(x_{it1},x_{it2},\ldots,x_{itk}).$$

则模型模型 (1) 可以表示为

$$Y=\mathbf{1}_{NT}\beta_0+X\beta+(\boldsymbol{I}_N\otimes\mathbf{1}_T)\mu+e. \tag{2}$$

这里 $\mathbf{1}_{NT}$ 表示元素全为 1 的 NT 维列向量, $\boldsymbol{I}_N$ 为 N 阶单位阵, 则有

$$\mathsf{Cov}\,(Y)=\sigma_\mu^2(\boldsymbol{I}_N\otimes\boldsymbol{J}_T)+\sigma_e^2\boldsymbol{I}_{NT}=\sigma_1^2P_1+\sigma_e^2Q+\sigma_1^2\overline{\boldsymbol{J}}_{NT}.$$

上式中 $\sigma_1^2=T\sigma_\mu^2+\sigma_e^2$, $P_1=P-\overline{\boldsymbol{J}}_{NT}$, $P=\boldsymbol{I}_N\otimes\overline{\boldsymbol{J}}_T$, $Q=\boldsymbol{I}_{NT}-P$, $\overline{\boldsymbol{J}}_T=\mathbf{1}_T\mathbf{1}_T'/T$. 矩阵 P, Q, P_1 和 $\overline{\boldsymbol{J}}_{NT}$ 都是对称幂等阵, 且两两正交. 将模型 (2) 两侧同时左乘矩阵 Q

$$QY=QX\beta+Q[(\boldsymbol{I}_N\otimes\mathbf{1}_T)\mu+e], \tag{3}$$

由此得到模型 (3) 中参数向量 β 的最佳线性无偏估计为

$$\widehat{\beta}_W=(X'QX)^{-1}X'QY, \tag{4}$$

估计量 (4) 就称为 Panel Data 模型 (2) 的 Within 估计, 且有

$$\mathsf{Cov}\,(\widehat{\beta}_W)=\sigma_e^2(X'QX)^{-1}.$$

§3. 最优设计

目前随机系数回归模型的最优设计研究多集中在一些具有应用背景的模型讨论上. 本部分仅就 $k=2$ 的情形讨论 Panel Data 模型 (2) 基于 Within 估计 (4) 的 D-, I-, A- 和 E-最优设计. 该模型具有较为广泛的应用背景, 早在 1958 年 Grunfeld 就应用此种模型建立了美国 10 家大型制造业公司 (年度) 实际总投资与公司实际总价值和资本存量价值的投资方程 ([2; p.20]). 下面将讨论该类模型的最优设计, 为了方便改用记号

$$y_{it}=\beta_0+x_{it}\beta_1+z_{it}\beta_2+\mu_i+e_{it},\qquad i=1,2,\ldots,N;\ t=1,2,\ldots,T. \tag{5}$$

不失一般性, 假设 $x_{it}\in[0,1]$, $z_{it}\in[0,1]$. 在很多实际情况中, 不同个体的观测点往往选择相同, 本文中也仅考虑恒等设计, 即所有个体选择相同的观测点, 此时有 $x_{it}=x_{jt}$, $z_{it}=z_{jt}$, $i,j=1,2,\ldots,N$; $t=1,2,\ldots,T$. 考虑近似恒等设计

$$\xi=\begin{pmatrix}(x_1,z_1) & (x_2,z_2) & \cdots & (x_m,z_m)\\ \delta_1 & \delta_2 & \cdots & \delta_m\end{pmatrix}, \tag{6}$$

其中, $(x_i,z_i)\in[0,1]^2$, $\delta_i>0$, $\sum\limits_{i=1}^m\delta_i=1$, $2\leqslant m\leqslant T$. 从 Panel Data 模型的 Within 估计的协方差阵可以发现, σ_e^2 虽为未知参数, 但不影响设计方案的选择, 不失一般性可假定 $\sigma_e^2=1$. 那么在设计 (6) 下基于 Within 估计的信息阵为

$$M(\xi)=\begin{pmatrix}\sum\limits_{i=1}^m x_i^2\delta_i-\Big(\sum\limits_{i=1}^m x_i\delta_i\Big)^2 & \sum\limits_{i=1}^m x_iz_i\delta_i-\sum\limits_{i=1}^m x_i\delta_i\sum\limits_{i=1}^m z_i\delta_i\\ \sum\limits_{i=1}^m x_iz_i\delta_i-\sum\limits_{i=1}^m x_i\delta_i\sum\limits_{i=1}^m z_i\delta_i & \sum\limits_{i=1}^m z_i^2\delta_i-\Big(\sum\limits_{i=1}^m z_i\delta_i\Big)^2\end{pmatrix}.$$

由此可得如下结论

定理 1　对 Panel Data 模型 (5) 基于 Within 估计考虑恒等设计, 则对任一形如 (6) 的近似设计, 存在一个定义在设计域顶点上的近似设计

$$\xi^*=\begin{pmatrix}(0,0) & (1,0) & (0,1) & (1,1)\\ \omega_1 & \omega_2 & \omega_3 & \omega_4\end{pmatrix},\qquad 0\leqslant\omega_j\leqslant1,\ \sum_{j=1}^4\omega_j=1, \tag{7}$$

使得 $M(\xi^*)\geqslant M(\xi)$.

证明: 易见

$$M(\xi^*)=\begin{pmatrix} (\omega_2+\omega_4)-(\omega_2+\omega_4)^2 & \omega_4-(\omega_2+\omega_4)(\omega_3+\omega_4) \\ \omega_4-(\omega_2+\omega_4)(\omega_3+\omega_4) & (\omega_3+\omega_4)-(\omega_3+\omega_4)^2 \end{pmatrix}.$$

令 $\omega_1=\sum\limits_{i=1}^{m}(1-x_i)(1-z_i)\delta_i$; $\omega_2=\sum\limits_{i=1}^{m}x_i(1-z_i)\delta_i$; $\omega_3=\sum\limits_{i=1}^{m}(1-x_i)z_i\delta_i$; $\omega_4=\sum\limits_{i=1}^{m}x_iz_i\delta_i$. 显然满足 $0\leqslant\omega_j\leqslant 1$, $j=1,2,3,4$, 且有 $\sum\limits_{j=1}^{4}\omega_j=1$. 这时

$$M(\xi^*)-M(\xi)=\begin{pmatrix} \sum\limits_{i=1}^{m}x_i\delta_i-\sum\limits_{i=1}^{m}x_i^2\delta_i & 0 \\ 0 & \sum\limits_{i=1}^{m}z_i\delta_i-\sum\limits_{i=1}^{m}z_i^2\delta_i \end{pmatrix}\geqslant 0.$$

即 $M(\xi^*)\geqslant M(\xi)$. □

对于经典的 D-, A-, I- 和 E- 最优设计准则均满足 Loewner 偏序性质 ([8; p.101]), 即对 $M_1\geqslant M_2$, 则准则函数 Φ 满足 $\Phi(M_1)\leqslant\Phi(M_2)$. 最优设计就是对给定的准则函数 Φ, 寻找使得 $\Phi(M(\xi))$ 达到最小的设计 ξ. 结合定理 1 的结论和 D-, A-, I- 和 E- 最优设计准则的 Loewner 偏序性质, 我们可以立刻得出结论: 对 Panel Data 模型 (5) 基于 Within 估计考虑恒等设计, 则 D-, A-, I- 和 E- 最优设计可以在设计类 (7) 中获得. 此时, 信息阵为

$$M(\xi^*)\triangleq\begin{pmatrix} a-a^2 & c-ab \\ c-ab & b-b^2 \end{pmatrix},\qquad a=\omega_2+\omega_4,\ b=\omega_3+\omega_4,\ c=\omega_4.$$

则有

$$M^{-1}(\xi^*)=\frac{1}{(a-a^2)(b-b^2)-(c-ab)^2}\begin{pmatrix} b-b^2 & ab-c \\ ab-c & a-a^2 \end{pmatrix}.$$

在 D- 最优设计准则下

$$\Phi_D(M(\xi^*))=\det(M^{-1}(\xi^*))=\frac{1}{(a-a^2)(b-b^2)-(c-ab)^2}.$$

当 $a=b=1/2$, $c=1/4$ 时上式取得最小值, 因此有

定理 2 对 Panel Data 模型 (5) 基于 Within 估计考虑恒等设计, 则 D- 最优设计为等权重设计

$$\xi_D^*=\begin{pmatrix} (0,0) & (1,0) & (0,1) & (1,1) \\ 1/4 & 1/4 & 1/4 & 1/4 \end{pmatrix}.$$

在 A- 最优准则下, 由于

$$\Phi_A(M(\xi^*))=\mathrm{tr}(M^{-1}(\xi^*))=\frac{(a-a^2)+(b-b^2)}{(a-a^2)(b-b^2)-(c-ab)^2}.$$

分别关于 a, b, c 求偏导, 得到上式极小值点满足

$$(1-2a)(b-b^2)+2b(c-ab)=0;\quad (1-2b)(a-a^2)+2a(c-ab)=0;\quad 2(c-ab)=0.$$

即当 $a=b=1/2$, $c=1/4$ 时上式取得最小值. 故有

定理 3　对 Panel Data 模型 (5) 基于 Within 估计考虑恒等设计, 则 A- 最优设计为等权重设计

$$\xi_A^*=\begin{pmatrix}(0,0) & (1,0) & (0,1) & (1,1)\\ 1/4 & 1/4 & 1/4 & 1/4\end{pmatrix}.$$

对于考虑预测方差优良性的 I- 最优设计,

$$\Phi_I(M(\xi^*))=\int_0^1\int_0^1(x,z)M^{-1}(\xi^*)(x,z)^{\mathrm{T}}\mathrm{d}x\mathrm{d}z=\frac{2(a-a^2)+2(b-b^2)-3(c-ab)}{6[(a-a^2)(b-b^2)-(c-ab)^2]}.$$

上式中 $\Phi_I(M(\xi^*))$ 的最小值在 $\omega_2=\omega_3=0.1371$, $\omega_1=\omega_4=0.3629$ 处获得, 故有

定理 4　对 Panel Data 模型 (5) 基于 Within 估计考虑恒等设计, 则 I- 最优设计为对称设计

$$\xi_I^*=\begin{pmatrix}(0,0) & (1,0) & (0,1) & (1,1)\\ 0.3629 & 0.1371 & 0.1371 & 0.3629\end{pmatrix}.$$

E- 最优设计是使得信息阵逆矩阵最大特征值达到最小的设计, 即

$$\Phi_E(M(\xi^*))=\lambda_1(M^{-1}(\xi^*)).$$

可以等价地考虑使得信息阵最小特征值达到最大的设计. 这里 $M(\xi^*)$ 的特征值为

$$\lambda_{1,2}(M(\xi^*))=\frac{a-a^2+b-b^2}{2}\pm\sqrt{(c-ab)^2-(a-a^2)(b-b^2)+\frac{(a-a^2+b-b^2)^2}{4}}.$$

其中信息阵 $M(\xi^*)$ 的最小特征值 $\lambda_2(M(\xi^*))$ 在 $a=b=1/2$, $c=1/4$ 时取得最大值, 故有

定理 5　对 Panel Data 模型 (5) 基于 Within 估计考虑恒等设计, 则 E- 最优设计为等权重设计

$$\xi_E^*=\begin{pmatrix}(0,0) & (1,0) & (0,1) & (1,1)\\ 1/4 & 1/4 & 1/4 & 1/4\end{pmatrix}.$$

§4.　讨　　论

对于 $k=2$ 的这类较为常见的 Panel Data 模型 (5), 上面的分析发现, 其基于 Within 估计的几类最优恒等设计均可在设计域的四个顶点上获得, 且具有良好的性质. 其中 D-, A-, E- 最优设计为四个设计谱点上的等权重设计, 而 I- 最优设计具有对称性, 且在 $(0,0)$, $(1,1)$ 两个对称点上的权重为 0.3629 明显大于 0.25, 在 $(1,0)$, $(0,1)$ 两个对称点上的权重为 0.1371 明显小于 0.25. 等权重设计的 I- 效为

$$\mathrm{Eff}_I(\xi_{1/4}^*)=\frac{\Phi_I(M(\xi_I^*))}{\Phi_I(M(\xi_{1/4}^*))}=0.83.$$

等权重设计的 I- 效还是比较理想的.

综上所述, 对于 Panel Data 模型 (5), 在采用 Within 估计作为参数估计方法时, 可采用设计域顶点上的等权重设计方案 (特别是观测次数较少时), 此时等权重设计 $\xi^*_{1/4}$ 不仅具有 D-, A- 和 E- 最优性质, 还具有比较理想的 I- 效.

参 考 文 献

[1] 王松桂, 范永辉. Panel 模型中两步估计的优良性 [J]. 应用概率统计, 1998, **14(2)**: 177–184.

[2] Hsiao C. *Analysis of Panel Data* [M]. 2nd ed. Cambridge: Cambridge University Press, 2003.

[3] Prus M, Schwabe R. Optimal designs for the prediction of individual effects in random coefficient regression [M] // Ucinski D, Atkinson A C, Patan M. *mODa 10 – Advances in Model-Oriented Design and Analysis*, Switzerland: Springer, 2013: 211–218.

[4] Cheng J, Yue R X, Liu X. Optimal designs for random coefficient regression models with heteroscedastic errors [J]. *Comm. Statist. Theory Methods*, 2013, **42(15)**: 2798–2809.

[5] Graßhoff U, Doebler A, Holling H, et al. Optimal design for linear regression models in the presence of heteroscedasticity caused by random coefficients [J]. *J. Statist. Plann. Inference*, 2012, **142(5)**: 1108–1113.

[6] Liu X, Yue R X, Chatterjee K. *R*-optimal designs in random coefficient regression models [J]. *Statist. Probab. Lett.*, 2014, **88**: 127–132.

[7] 程靖, 岳荣先. 两变量随机系数回归模型的最优设计 [J]. 应用概率统计, 2012, **28(3)**: 225–234.

[8] Pukelsheim F. *Optimal Design of Experiments* [M]. New York: Wiley, 1993.

Optimal Designs of Panel Data Model Based on Within Estimator

CHENG Jing

(*College of Science, Anhui Agricultural University, Hefei, 230036, China*)

YUE RongXian

(*College of Mathematics and Sciences, Shanghai Normal University, Shanghai, 200234, China*)

Abstract: In this paper, optimal identical designs of bivariate panel data model based on within estimator are discussed. It is proved that D-, A-, E- and I-optimal design of bivariate panel data model can be obtained on the vertexes of the design region. Equal-weight design is proved to be D-, A-, E-optimal of bivariate panel data model in the paper. I-optimal design is also obtained in the paper and it shows that equal-weight design has high I-efficiency.

Keywords: panel data model; within estimator; optimal design

2010 Mathematics Subject Classification: 62K05