

异方差的估计与实证^{*}

孙 凤

(清华大学社会学系, 北京 100084)

摘 要:“异方差”是微观计量建模中不可回避的问题, 尽管计量经济学理论从残差项出发给出大量的检验方法, 但对如何估计异方差则没有给出明确答案。从数据分布出发, 认为极大似然法在估计异方差方面具有独特的优势。首先建立异方差的极大似然估计模型, 然后对一实际调查数据进行异方差估计。

关键词: 异方差; 极大似然估计; 微观计量分析

中图分类号: O212 1 **文献标识码:** A **文章编号:** 0529-6579 (2005) S2-0103-04

1 问题的提出

微观经济计量建模研究中一个不可回避的问题就是“异方差性”(Heteroskedasticity)问题。计量经济理论认为如果存在异方差还用最小二乘法去估计参数, 会产生以下严重后果: ①参数估计值不再具有最小方差的特性; ②解释变量的显著性检验失效; ③模型预测精度降低。

针对异方差问题, 涌现出大量的检验方法, 常见的有: 图示检验法、等级相关检验法、Glejser 检验、Battlett 检验、Breusch-Pagan 检验、Goldfeld-Quandt 检验、Wald 检验、拉格朗日乘数检验、似然比检验和 White 大样本检验。这些检验的共同思想是设法通过误差的估计量来检验误差方差与解释变量间是否存在相关性。若存在明显的相关, 则原模型存在异方差性; 否则, 认为原模型满足方差齐性条件。

由上可见, 传统计量经济学理论对异方差的认知集中在对残差项的认识上, 透过残差项的异方差性来反映变量的方差特征。这带给我们的问题是, 既然因变量 y_i 或自变量 x_i 是已知数据, 由参数 β 和 σ_i^2 共同决定着 y_i 或 x_i 的实际分布状况, 为什么不直接从因变量 y_i 或自变量 x_i 的分布了解齐方差或异方差性呢?

事实上, 如果尝试从变量的角度去思考异方差问题, 发现异方差的存在其实是客观事实。在微观计量分析中, 数据多为个体数据, 由于个体的差异性, 异方差的存在应该是常态, 而齐方差性则是特例。如果我们仅仅为了研究问题的便利, 而要求所

有变量的方差都是相同的, 是违背客观事实的。

因此, 本文认为异方差性的存在是客观事实, 同时认为判别异方差性固然重要, 但如何估计异方差, 并揭示出异方差性的存在对客观事物的影响程度有多大则更为必要。鉴于此, 本文尝试以极大似然法研究该问题。

2 异方差的极大似然估计

传统计量经济模型中, 对于存在异方差的模型, 是通过一些方法剔除掉异方差性, 如加权最小二乘法, 广义最小二乘法等, 目前使用最为广泛的是广义最小二乘法 (GLS):

$$\hat{\beta} = (X' \Omega^{-1} X)^{-1} X' \Omega^{-1} Y \quad (1)$$

(1) 式中, Ω 是残差项的方差-协方差矩阵。

尽管 GLS 法能够纠正由于异方差的存在而使最小二乘法存在偏差的估计量, 但该方法并不能测算出有关异方差性影响程度大小的结果, 这限制了我们对于个体差异性影响程度的研究。本文以因变量 y 服从正态分布展开对问题的讨论。

当方差为齐性时:

$$y_i := \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2} \left[\frac{(y_i - x_i \beta)^2}{\sigma^2} \right]} \quad (2)$$

将方差加入下标, 可以得到异方差的表达式:

$$y_i := \frac{1}{\sqrt{2\pi\sigma_i^2}} e^{-\frac{1}{2} \left[\frac{(y_i - x_i \beta)^2}{\sigma_i^2} \right]} \quad (3)$$

(3) 式与 (2) 式相比, 每一个观察值都有一个方差, 由 (3) 式产生两个问题:

第一个问题是模型的识别问题。因为在该模型

* 收稿日期: 2005-09-09

基金项目: 国家自然科学基金资助项目 (70173040)

作者简介: 孙 凤 (1964 年生), 女, 博士, 副教授, E-mail: sunfeng@tsinghua.edu.cn

中，待估参数为 $k+N$ 个，其中 k 为 β 的个数，而 N 为 σ_i^2 的个数。当待估参数值多于观察值时，模型是不可识别的，参数亦估计不出来。为了克服该问题，我们建立 σ_i^2 与一个可观察的外生变量和相关参数的关系模型，此外生变量的维度从 N 降到 s ，当 $k+s$ 小于或等于 N ，则该模型即可识别，参数亦可识别。

第二个问题是 σ_i^2 的约束条件应为非负。尽管许多函数形式可以满足该约束条件，但简单与可操作性则是我们首先需要的。为了达到该目标，假定 z_i 是一个 $1 \times s$ 的外生变量，影响回归模型的方差，假定 γ （希腊字母）是 $s \times 1$ 的系数向量，则

$$\sigma_i^2 = e^{z_i \gamma} \tag{4}$$

(4) 式的技术处理可以保证 σ_i^2 的非负性。

我们通过用参数替代 σ_i^2 ，建立极大似然函数模型：

$$y_i := \frac{1}{\sqrt{2\pi e^{z_i \gamma}}} e^{\frac{1}{2} \left[\frac{(y_i - x_i \beta)^2}{e^{z_i \gamma}} \right]}$$

$$\begin{aligned} \ln L(y/x, z, \beta, \gamma) &= \sum_{i=1}^N \ln \left(\frac{1}{\sqrt{2\pi e^{z_i \gamma}}} e^{\frac{1}{2} \left[\frac{(y_i - x_i \beta)^2}{e^{z_i \gamma}} \right]} \right) = \\ &= -\frac{N}{2} \ln(2\pi) - \sum_{i=1}^N \frac{1}{2} \ln e^{z_i \gamma} - \frac{1}{2} \sum_{i=1}^N \left[\frac{(y_i - x_i \beta)^2}{e^{z_i \gamma}} \right] \\ &= -\frac{N}{2} \ln(2\pi) - \sum_{i=1}^N \frac{1}{2} z_i \gamma - \frac{1}{2} \sum_{i=1}^N \left[\frac{(y_i - x_i \beta)^2}{e^{z_i \gamma}} \right] \end{aligned}$$

通过求偏导、求极大值，我们就可以得到参数的估计值。本例由于存在非线性关系 $z_i \gamma$ ，需要借助计算机迭代完成。一旦得到参数的估计值，方差就可以通过公式（4）简单计算求得。

3 一个实例

本文运用实际调查数据对异方差的极大似然估计方法做一考察，数据来源于中国人民大学 2002 年北京市居民医疗消费行为意愿调查^①。本次调查的有效样本是 620 个，调查的时段为自此调查前 10 个月内是否到过医院。在此调查时段内曾经到过医院的样本有 402 个，本文仅对到过医院的样本数据做一分析。本次调查反映居民就诊满意度的指标有 9 个，分别是：TT4 总体满意度；TT5 总体收费合理性；TT6 药品费用合理性；TT7 治疗费用合理性；TT8 就诊手续方便度；TT9 环境卫生满意度；TT10 硬件设施满意度；TT11 医务人员专业水平；TT12 医务人员服务态度。该问卷通过设计 10 级量表，判断满意程度，其中“1”为特别不满意，“10”为特别满意。在本文的变量分析中，TT4 为因变量，其他变量为自变量。

如果我们假定个体的方差都相同，这意味着被调查者对 9 个医疗消费满意度指标的内在评判标准相同。本文认为由于被调查者所处的社会经济环境不同，其内在评判标准是不同的，个体之间内在标准的差异性会对问题的回答产生影响，这种内在标准的差异即为异方差。因此，就本文来说，估计出异方差，意味着估计出人们的内在评判标准，这为深入研究居民的医疗消费行为提供了重要信息。

由于本次调查存在缺失数据，本文运用预期最大似然法（EM）对缺失数据进行了填充^⑧。我们首先通过建立异方差的极大似然估计模型估计异方差，然后将包含异方差的模型与普通最小二乘法回归模型的进行比较分析。

本文的分析软件是 R1.9.1。R 软件是当前国际上社会科学研究中运用非常广泛的软件，该软件在进行极大似然估计和贝叶斯估计上具有独特的功能。

R 软件需要写程序，在下列程序中，med 为数据库名，absTT5—TT12 是为了估计出各个变量的异方差，而对变量重命名，MLhetreg 是建立异方差的极大似然估计模型，lm 是建立普通最小二乘法回归模型。

```
# Heteroskedastic Linear Regression
> med<-load("E:/我的文档/med.RData")
> attach(med)
> names(med)
[1] "TT4" "TT5" "TT6" "TT7" "TT8" "TT9" "TT10"
   "TT11" "TT12"
> absTT5<-abs(TT5)
> absTT6<-abs(TT6)
> absTT7<-abs(TT7)
> absTT8<-abs(TT8)
> absTT9<-abs(TT9)
> absTT10<-abs(TT10)
> absTT11<-abs(TT11)
> absTT12<-abs(TT12)
HR1<-MLhetreg(TT4, cbind(TT5, TT6, TT7, TT8, TT9,
TT10, TT11, TT12), cbind(absTT5, absTT6, absTT7,
absTT8, absTT9, absTT10, absTT11, absTT12))
> summary(lm(TT4 ~ TT5+TT6+TT7+TT8+TT9+
TT10+TT11+TT12))
> summary(HR1)
```

根据以上程序，我们可以得到二个计算结果，

① 调查问卷见易丹辉：《北京市居民医疗消费行为及意愿研究》，中国人民大学出版社 2004 年版。

表 1 为最小二乘法回归系数, 表 2 为自变量异方差的最大似然估计值。表的下方为模型的拟和优度和检验结果。表 2 的 Z constant 代表 (4) 式中异方差模型的常数值, absTT5—absTT12 代表反映异方差的各个外生变量 γ 的参数估计值。

表 1 OLS 回归结果的系数

Tab 1 OLS Regression Coefficients

Estimate	Std. Error	t value	Pr(> t)
(Intercept)	5. 00973	0. 12931	38. 74 < 2e- 16 ***
TT5	− 0. 47244	0. 01832	− 25. 79 < 2e- 16 ***
TT6	0. 30682	0. 01642	18. 68 < 2e- 16 ***
TT7	0. 50168	0. 01448	34. 64 < 2e- 16 ***
TT8	− 0. 29290	0. 01429	− 20. 49 < 2e- 16 ***
TT9	− 0. 93437	0. 03376	− 27. 68 < 2e- 16 ***
TT10	0. 62444	0. 02997	20. 83 < 2e- 16 ***
TT11	0. 77531	0. 01677	46. 23 < 2e- 16 ***
TT12	− 0. 27601	0. 01662	− 16. 61 < 2e- 16 ***

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1; Residual standard error: 0.3267 on 393 degrees of freedom; Multiple R-Squared: 0.9079, Adjusted R-squared: 0.906; F-statistic: 484 on 8 and 393 DF, p-value:< 2.2e-16

表 1 回归模型的计算结果显示:TT5(总体收费合理性),TT8(就诊手续方便度),TT9(环境卫生满意度),TT12(医务人员服务态度)均对人们的医院满意度判断形成负影响。而 TT6(药品费用合理性),TT7

(治疗费用合理性),TT10(硬件设施满意度),TT11(医务人员专业水平)则对医院满意度产生正影响。

该结论中对药品费用合理性的认识与当前人们普遍对医院药品费用过高的看法有悖,因此需要结合异方差的极大似然估计纠正由于被调查者评判标准的差异性而产生的模型偏差。

表 2 通过引入异方差分析来矫正普通回归模型的偏差。表 2 的极大似然回归模型计算结果显示:尽管各个自变量对医院满意度的影响方向与回归模型相同,但引入异方差后,TT6(药品费用合理性)、TT11(医务人员专业水平)的异方差对回归方程均有显著负影响,这意味着被调查者对药品费用合理性、医务人员专业水平的评判标准存在较大的差异性,特别是如果将药品费用和药品费用异方差的影响合计起来,有关药品费用对医院总体看法的影响系数应为负。普通回归模型由于没有考虑该变量的异方差性而得出对医院的总体满意度评判偏高的结论。另外,TT9(环境卫生满意度)的异方差进一步降低了普通回归模型中该变量的影响程度,而 TT10(硬件设施满意度)的异方差则进一步提高了普通回归模型中该变量的影响程度。以上计算结果和分析结果显示:考虑了变量异方差极大似然估计法的回归模型较普通最小二乘法的回归模型更有现实解释力。

表 2 极大似然回归结果的系数

Tab 2 MLhetreg Regression Coefficients

	Estimate	Std. Error	z-value	Pr(> z)
Constant	4. 64786515	0. 15394854	30. 1910309	3. 106250e− 200
TT5	− 0. 43121689	0. 01914767	− 22. 5205914	2. 608638e− 112
TT6	0. 30263126	0. 01341344	22. 5618012	1. 028487e− 112
TT7	0. 48172392	0. 01240307	38. 8390745	0. 000000e+ 00
TT8	− 0. 32204098	0. 01498148	− 21. 4959370	1. 699302e− 102
TT9	− 0. 80702950	0. 03982539	− 20. 2641952	2. 662783e− 91
TT10	0. 55071865	0. 02900868	18. 9846172	2. 285961e− 80
TT11	0. 73723119	0. 01699107	43. 3893335	0. 000000e+ 00
TT12	− 0. 21997782	0. 01644568	− 13. 3760283	8. 349102e− 41
Z Constant	3. 03835592	0. 89743209	3. 3856110	7. 101999e− 04
absTT5	0. 11548189	0. 11222467	1. 0290241	3. 034683e− 01
absTT6	− 0. 66241545	0. 11252590	− 5. 8867819	3. 937879e− 09
absTT7	0. 04407974	0. 08280143	0. 5323548	5. 944803e− 01
absTT8	0. 02372137	0. 07824447	0. 3031699	7. 617604e− 01
absTT9	− 0. 95201520	0. 12945686	− 7. 3539187	1. 924788e− 13
absTT10	0. 54078971	0. 10361818	5. 2190618	1. 798317e− 07
absTT11	− 0. 19390596	0. 08148769	− 2. 3795738	1. 733267e− 02
absTT12	− 0. 03617569	0. 07646876	− 0. 4730780	6. 361575e− 01

Log− Likelihood: 331. 0112; Wald statistic: 121. 6169 with 8 degrees of freedom; p= 0

传统计量经济学理论仅仅关注总体异方差存在与否的检验,而对是由于哪些变量存在异方差以及异方差程度有多大没有给予特别的关注. 本文通过对北京市居民医疗消费调查数据的分析,提出该问题的研究思路与研究方法. 异方差极大似然估计方法的目的不仅仅如此,实际上,在行为经济学研究中,该方法对于研究微观个体行为差异性具有独特的功能.

参考文献:

[1] MULIER H G, STADTMULLER U. Estimation of Heteroscedasticity in Regression Analysis[J]. The Annals of Statistics, 1987, 15(2): 610— 625.
[2] LONG J S, ERVIN L H. Using Heteroscedasity Consistent

Stand Errors in the Linear Regression Model[J]. American Statistician, 2000, 54: 217— 224.
[3] GRAGG J G. More Efficient Estimation in the Presence of Heteroscedasticity of Unknown fom[J]. Econometrica, 1983, 51(3): 751— 764.
[4] 兰嘉庆, 余宛泠. 异方差的游程检验[J]. 中山大学学报: 自然科学版, 2004, 43(增刊): 9— 11.
[5] 高辉. 异方差的诊断方法及数据属性影响的实证分析[J]. 山西师范大学学报. 自然科学版, 2003, 4: 1— 6.
[6] 李子奈. 计量经济学[M]. 北京: 高等教育出版社, 2000.
[7] 易丹辉. 北京市居民医疗消费行为及意愿研究[M]. 北京: 中国人民大学出版社, 2004.
[8] 孙凤. 微观计量分析中缺失数据的极大似然估计. 数量经济技术经济研究, 2005, 22(11): 51— 58.

Maximum Likelihood Estimates for Heteroskedastiy

SUN Feng

(Department of Sociology, Tsinghua University, Beijing 100084, China)

Abstract: Studies in the micro-econometric frequently suffer from heteroskedastiy. Traditional methods focus on how to test heteroskedastiy, and hardly discuss how to estimate it. Maximum likelihood can be an effective and practical method for handling heteroskedastiy. This paper illustrates the basic principles of ML under heteroskedastiy; then uses the ML to estimate heteroskedastic coefficients based on a concrete survey data.

Key words: heteroskedastiy; maximum likelihood function; micro-econometric analysis