DOI:10.13546/j.cnki.tjyjc.2013.05.012

多重共线性的解决:剔除变量的新标准

刘　明[a,b]

（兰州商学院 a.甘肃经济发展数量分析研究中心;b.统计学院,甘肃 兰州 730020）

摘　要:当使用剔除变量法解决线性回归模型的多重共线性问题时,根据方差膨胀因子的大小来选择被剔除变量是存在缺陷的。解释变量显著性检验的t统计量的绝对值大小反映了该解释变量对被解释变量的贡献程度的大小,因此可以将t统计量绝对值作为剔除解释变量的依据,从而得到一类多重共线性的解决办法。

关键词:多重共线性;t统计量

中图分类号:O212　　**文献标识码:**A　　**文章编号:**1002-6487(2013)05-0082-02

0　引言

线性回归模型的多重共线性的本质是解释变量之间存在线性相关。多重共线性的解决有多种经验性方法,这些方法因模型和样本数据的不同而各异,其中一类比较常用而且简单的办法是“剔除变量法”,即剔除引起多重共线性的解释变量,以达到解决多重共线性问题的目的。实施剔除变量法的关键是确定哪一个或哪些变量应该被剔除,因此需要确立剔除依据。文献[1,2]认为可以根据方差膨胀因子(VIF)的大小来选择被剔除变量,VIF最大的变量应首先剔除。该依据的理由是,VIF最大的变量与其余变量的相关性最强,因而是多重共线性的罪魁祸首,因此应首先剔除。为考察这种方法的效果,首先看一个实例,这也是研究的出发点。

1　剔除变量法的一个实例:以方差膨胀因子为准则

为展示以方差膨胀因子为准则的剔除变量的方法,这里利用朗利数据构造一个例子。数据如下表,其中Y=被雇佣人数(千人),X_1=GNP价格缩减指数,X_2=GNP(百万美元),X_3=失业人数(千人),X_4=服役人数(千人),X_5=14岁以上非编制人口,X_6=时间。原数据参见文献[3]。

利用上述数据,以Y为被解释变量,其余变量为解释变量构建线性回归模型如下:

$$\hat{Y}=\underset{(3.4332)}{77270.12}+\underset{(0.1774)}{1.5062}X_1-\underset{(-1.0695)}{0.0358}X_2-\underset{(-4.1364)}{2.0202}X_3-\underset{(-4.8220)}{1.0332}X_4-\underset{(-0.2261)}{0.0511}X_5+\underset{(4.0159)}{1829.15}X_6$$

其中括号内为t检验统计值,为节约篇幅,其余统计量均未给出。此模型整体拟合效果较好,可决系数$R^2=0.9955$接近于1,但部分解释变量不显著,因而可能存在多重共线性问题,经过进一步诊断,模型确实受到共线性问题干扰。考虑使用剔除变量法解决多重共线性问题,依据该方法的思想,选择方差膨胀因子最大的解释变量予以首先剔除。解释变量的方差膨胀因子计算结果依次为:

表1　朗利数据

Y	X_1	X_2	X_3	X_4	X_5	X_6
60323	830	234289	2356	1590	107608	1
61122	885	259426	2325	1456	108632	2
60171	882	258054	3682	1616	109773	3
61187	895	284599	3351	1650	110929	4
63221	962	328975	2099	3099	112075	5
63639	981	346999	1932	3594	113270	6
64989	990	365385	1870	3547	115094	7
63761	1000	363112	3578	3350	116219	8
66019	1012	397469	2904	3048	117388	9
67857	1046	419180	2822	2857	118734	10
68169	1084	442769	2936	2798	120445	11
66513	1108	444546	4681	2637	121950	12
68655	1126	482704	3813	2552	123366	13
69564	1142	502601	3931	2514	125368	14
69331	1157	518173	4806	2572	127852	15
70551	1169	554894	4007	2827	130081	16

$$VIF_{X_1}=135.14\ VIF_{X_2}=1666.67\ VIF_{X_3}=33.67\ VIF_{X_4}=3.59\ VIF_{X_5}=400\ VIF_{X_6}=769.23$$

显然,X_2的方差膨胀因子最大,先将其剔除。剔除后重新构建的回归模型为:

$$\hat{Y}=\underset{(5.9693)}{94498.64}-\underset{(-0.6753)}{4.3917}X_1-\underset{(-9.5381)}{1.5263}X_3-\underset{(-4.8563)}{0.9258}X_4-\underset{(-2.0060)}{0.2526}X_5+\underset{(5.2463)}{1438.62}X_6$$

经检验,该模型仍存在多重共线性问题,继续实施剔除变量法,选择该模型中方差膨胀因子最大的解释变量予以剔除,剔除后继续构建回归模型并检验是否存在多重共线性问题,若存在,继续按上述过程剔除变量,直到无多重共线性问题存在为此。最终得到的模型是:

$$\hat{Y}=\underset{(13.8378)}{50662.81}+\underset{(3.2266)}{2.2647}X_3+\underset{(3.0211)}{2.8474}X_4$$

该模型的可决系数$R^2=0.5608$,相对偏小,而且模型中仅剩余两个解释变量X_3、X_4,因此该模型没有达到对原问题的正确表述。

作者简介:刘　明(1981-),男,安徽霍邱县人,硕士,讲师,研究方向:统计理论与方法、经济计量分析。

DOI:10.13546/j.cnki.tjyjc.2012.19.015

基于一元线性回归模型异方差对加权最小二乘法的考察

刘　明[a,b]

（兰州商学院 a.统计学院 b.甘肃经济发展数量分析研究中心，兰州 730020）

摘　要：作为普通最小二乘法的改进，加权最小二乘法用于存在异方差问题的线性回归模型的参数估计。文章通过对加权最小二乘估计量、加权最小二乘变换的分析，并结合实际例证研究发现，加权最小二乘法在应用中存在一些不足之处，因而当发现模型存在异方差时使用加权最小二乘法是存在风险的。

关键词：异方差；加权最小二乘法；模型参数估计与检验

中图分类号：O212　　**文献标识码**：A　　**文章编号**：1002-6487(2012)19-0011-04

1　问题的提出

在经典回归分析中，为保证线性回归模型的普通最小二乘估计量是最佳线性无偏估计量，通常要求模型满足所谓的“高斯假定”，同方差即是高斯假定条件之一。在高斯假定条件下，线性回归模型的普通最小二乘估计量表现出了良好的统计性质：线性、无偏性和最小方差性，即它是最佳线性无偏估计量，这一结论即为“高斯——马尔科夫定理”所表述。然而对于现实应用中的线性回归模型，并非总能满足高斯假定，这时模型的普通最小二乘估计量很可能不再是最佳线性无偏估计量，异方差即是违背高斯假定的情形之一。异方差现象在线性回归模型的实际应用中较为普通，尤其对于截面数据模型。当模型存在异方差时，参数估计和假设检验一般都会出现不良后果，导致模型不能正常应用于经济问题的分析和研究。为避免异方差这种不良影响，人们在对存在异方差问题的模型进行参数估计时引入了加权最小二乘法，这种参数估计方法是解决线性回归模型异方差问题的有效途径。但现实问题是，由加权最小二乘法所得到的估计量即加权最小二乘估计量一定优于普通最小二乘估计量吗？当模型存在异方差问题时都必须加以解决吗？使用加权最小二乘法估计线性回归模型参数时没有任何成本吗？诸如此类的一些问题摆在了理论研究者和应用研究者的面前。本文将通过对异方差问题阐述及对加权最小二乘原理的分析，从总体及样本两个角度对加权最小二乘法及其在现实中的应用问题进行分析论证，以对上述问题或类似上述问题做出回答。为简化问题的分析过程，本文将以一元线性回归模型为例进行研究，从而可以避免复杂的矩阵运算，所得结论亦不失一般性。

2　异方差及加权最小二乘原理

异方差是与同方差相对而言的，同方差即是指线性回归模型的随机干扰项的方差全部等于一个有限的常数，该常数通常被视为参数，而异方差现象可表述为线性回归模型中随机干扰项的方差不再是某一相等的常数，而是随着观察点的变化而变化。

对于如下的一元线性回归模型：

$$y_i=\beta_0+\beta_1 x_i+\mu_i \tag{1}$$

若随机干扰项μ_i的方差$\mathrm{Var}(\mu_i)=\sigma^2$（$\sigma^2$为常数参数）对于任意观测点$i$均成立，则认为该一元线性回归模型是同方差的；若$\mu_i$的方差$\mathrm{Var}(\mu_i)=\sigma_i^2$，$\sigma_i^2$表明其取值随着观察点$i$的变化而变化，则认为该一元线性回归模型存在异方差问题。

在一般情形下，存在异方差的线性回归模型的普通最小二乘估计量不再是一个有效估计量。为了能得到一个有效估计量，通常对存在异方差问题的线性回归模型实施加权最小二乘法。加权最小二乘法是解决线性回归模型异方差问题的较为凑效的一类参数估计方法，它是普通最小二乘法的改进。加权最小二乘法在估计过程中也是通过对线性回归模型残差平方求和再求最小，但在残差平方求和过程中，加权最小二乘法不是像普通最小二乘法那样对残差平方进行简单求和，而是考虑到了异方差的影响，对残差平方进行加权求和。具体说，存在异方差情形时，不同的样本点对于样本回归直线的影响作用是不同的，为体现这种不同，在残差平方求和过程中就应考虑对不同样本点所形成的残差的平方赋予不同的权重，权重通常选择对应样本点所形成的随机干扰项方差的倒数。

考虑上述一元线性回归模型存在异方差，即$\mathrm{Var}(\mu_i)=\sigma_i^2$，设该模型由样本所形成的残差为$e_i$，权重为$\omega_i$，它为随机干扰项方差的倒数，即$\omega_i=1/\sigma_i^2$。可见，随机干扰项方差越大，对应的权重就越小，残差平方求和过程中该样本点残差平方的作用就越小，正体现了加权最小二

作者简介：刘　明（1981-），男，安徽霍邱人，硕士，讲师，研究方向：统计理论与方法、经济计量分析。

DOI:10.13546/j.cnki.tjyjc.2012.04.003

普通最小二乘法的几何分析

刘　明[a,b]

（兰州商学院 a.甘肃经济发展数量分析研究中心；b.统计学院，兰州 730020）

摘　要：普通最小二乘估计法是在目标函数残差平方和的达到最小的条件下求得参数估计量，从向量的角度来说，普通最小二乘法将被解释变量分解成了相互正交的两部分，通过空间向量理论和几何分析方法，可以在欧氏空间内对普通最小二乘估计量进行求解，这种分析过程使普通最小二乘法变得更直观。

关键词：普通最小二乘法；投影；向量分解

中图分类号：F222.1　　**文献标识码**：A　　**文章编号**：1002-6487(2012)0090-02

普通最小二乘法是线性回归模型最基本最重要的参数估计方法之一。最小，即残差平方和达到最小；二乘，即残差的二次方。它通过构造目标函数——残差平方和，该函数以模型参数估计量为变量，以函数值达到最小来确定参数估计量的取值。普通最小二乘法在数学上构造计技巧性强，且便于理解和操作，在满足高斯假定情况下普通最小二乘估计量又具有良好的性质，因此为人们所推崇。普通最小二乘法简单易行，但很多人对其可以灵活掌握却不能达到充分理解，例如，为什么“残差向量与解释变量正交”和“残差平方和最小”两种情形是等价的？究其原因，是这种方法在本质上是一种数学方法，需要从数学的角度进行分析和进一步的认识。本文从普通最小二乘估计的方法出发，在矩阵分析的基础上，提出了一种几何方法求解普通最小二乘估计量，从几何的角度考察分析普通最小二乘法。

1　普通最小二乘法的矩阵分析

普通最小二乘法的实现过程较为简单，首先是构造残差平方和函数，再利用微分学中求极值的办法构造残差平

作者简介：刘　明(1981-)，男，安徽霍邱人，硕士，讲师，研究方向：统计计量分析。

件发生，属于股票市场正常调整。

通过以上实验可以看到，大部分异常点都对应着重大事件和消息，其余则对应着相对较大的波动，证明了本文采用的方法能够准确有效地检测异常点，并避免了“遮蔽效应”对异常点检测的影响，取得了良好的效果。

4　结论

本文首先使用GARCH(1,1)模型对股票数据收益率进行残差估计。残差数据反映了股票市场走势对均值的偏离，但直接对其进行异常点检测，则无法避免“遮蔽效应”。本文通过对残差数据进行haar小波变换得到高频系数进行异常点检测，能够准确地检测异常点，且很好地克服了“遮蔽效应”。最后分析证明了我们的方法效果良好，具有很好的理论和应用价值。

参考文献：

[1]王宏鼎，童云海，谭少华，唐世渭，杨冬青.异常点挖掘研究进展[J].智能系统学报，2006.

[2]陶运信，皮德常.屏蔽输入参数敏感的异常点检测新方法[J].计算机科学，2008.

[3]刘晓艳，王丽珍，杨志强，陈红梅.基于数学形态学的模糊异常点检测[J].计算机研究与发展，2009，46.

[4]陶运信，皮德常.基于邻域和密度的异常点检测算法[J].吉林大学学报，2008.

[5]R.Engle. Autoregressive Conditional Heteroskedasticity with Estimates of the Variance of U. K. Inflation[J]. Econometrica,1982,50(4).

[6]T.Bollerslev. Generalized Autoregressive Conditional Heteroskedasticity[J]. Journal of Economics,1986,31(3).

[7]R. F. Engle, D. Lilien, R. P. Robins. Estimating Time Varying Risk Premia in the Term Structure: The ARCH-M Model [J]. Econometrica, 1987,55(2).

[8]张世英，柯柯.ARCH模型体系，系统工程学报，2002，(3).

[9]Aurea Granéa, Helena Veiga. Wavelet Based Detection of Outliers in Financial time Series[J]. Computational Statistics and Data Analysis, 2010,54(11).

[10]D.Pena,F.Prieto.Multivariate Outlierdetection and Robust Covariance Matrix Estimation[J].Technometrics, 2001, 43(3).

[11]傅强，彭选华，毛一波.金融时间序列变点探测的小波模极大值线方法[J].重庆大学学报(自然科学版)2007.

[12]周大镯，刘月芬，马文秀.时间序列异常检测[J].计算机工程与应用，2008.

[13]X. Zhang,M.King.Influence in Generalized Autoregressive Conditional Heteroscedasticity Processes[J].Journal of Business & Economic Statistics,2005,118~129.

[14]高铁梅.计量经济学建模与教程第二版[M].北京：清华大学出版社，2009.

文章编号: 1674-1730(2012) 03-0129-03

统计学专业计量经济学教学中的问题探讨

——以兰州商学院统计学专业计量经济学课程为例

刘 明[1,2]

(1. 兰州商学院 甘肃经济发展数量分析研究中心,甘肃 兰州 730020; 2. 兰州商学院 统计学院,甘肃 兰州 730020)

摘 要: 结合兰州商学院统计学专业的计量经济学课程的教学状况讨论了统计学专业的计量经济学教学的主要问题,他们表现在课时安排、课程教学内容、教学方法和软件学习等方面; 针对这些问题,在统计学专业背景下提出相应的解决办法。

关键词: 统计学专业; 计量经济学教学; 问题解决办法
中图分类号: G642. 42 文献标识码: A

计量经济学是一门经济学科,研究的是现实经济问题,以对经济现象的透彻认识为目的; 由于其在经济领域的广泛应用,因而在经济学科中占据了极为重要的地位,计量经济学已成为经济学专业教育中的重点学科。1998 年,我国教育部经济学学科教学指导委员会将计量经济学定为高等学校经济学门类各专业的核心课程之一。目前,国内大部分学校已将计量经济学作为经济管理类专业的重要基础课程。兰州商学院计量经济学课程于 1998 年开设,经过十多年的建设和发展,计量经济学在兰州商学院诸多经济类专业的核心课程中已有着重要影响力,已经形成一支结构合理的教学队伍和科研队伍。统计学作为一个应用经济学专业①,计量经济学在专业课程设置中居于重要地位,统计学专业的培养模式强调数理分析能力,因而在课程教学中对统计学专业的计量经济学教学有更高要求。当前,兰州商学院统计学专业的计量经济学教学已开始逐步采用双语化教学,选用国外经典的计量经济学原版教材,力争该课程的教学水平上升一个新台阶。笔者是兰州商学院计量经济学教学团队中的一员,多年来承担着统计学专业的计量经济学的教学任务,在教学过程中发现统计学专业的计量经济学教学并非尽如人意,也存在一些问题及需要改进之处。本文从统计学专业计量经济学教学中存在的问题出发,进一步提出相应的解决方法和改进方向,为广大计量经济学教育工作者提供借鉴。

1 统计学专业计量经济学教学中存在的主要问题

对于统计学专业而言,计量经济学不仅是一门经济类核心课程,更是一门极具统计特色的专业主干课程,计量经济学在统计学专业各课程的学习中几乎居于中心地位。因此,计量经济学的教学水平的高低和教学效果的优劣直接关系到统计学专业人才的培养。然而,统计学专业计量经济学在实际教学中存在着诸多方面的不足。

首先是课时数安排不能满足教学的实际需要。现在一般的高等院校为本科生开设了许多课程以增加学生的知识宽度,必然导致一些重要课程学时数的压缩。兰州商学院也是如此。2007 年前,兰州商学院统计学专业的计量经济学课时总数为 85 学时,每周四个小时的课堂理论讲授外加 1 小时的上机实验,而现在的学时总数仅为 51,且没有上机实验,这使得授课教师难以安排教学内容。从事计量经济学教学的人都

① 当前,统计学已划归到理学之下,但就财经类院校而言,统计学专业培养模式和课程设置仍偏重于应用经济学,本文所述的也正是应用经济学下的统计学专业。

收稿日期: 2011 - 05 - 22
基金项目: 兰州商学院教改项目和兰州商学院双语教学示范课程建设项目的研究成果(20100225)
作者简介: 刘明(1981—) ,男,安徽霍邱人,讲师,硕士,主要从事统计理论与方法、经济计量分析的教学与研究。

统计学专业计量经济学课程教学的思考

刘 明[1,2]

(1.兰州商学院 统计学院 2.甘肃经济发展数量分析研究中心,甘肃 兰州 730020)

[摘要] 对于统计学专业而言,计量经济学课程既是专业核心课,又是专业主干课,在统计学专业教学中居于承上启下的中心地位。通过分析统计学专业的课程设置状况,指出了计量经济学课程与相关统计学方法类课程的联系,由此进一步提出了计量经济学在统计学专业教学中应实现四大功能,以此为计量经济学教学内容的设计提供思路。

[关键词] 统计学专业;计量经济学;教学功能;教学内容设计

[中图分类号] G642.3 [文献标识码] A [文章编号] 1674-3288(2012)02-0119-03

[收稿日期] 2012-01-15

[基金项目] 兰州商学院教改项目(编号:20100225)和兰州商学院双语教学示范课程建设项目的研究成果

[作者简介] 刘明(1981-),安徽霍邱人,兰州商学院统计学院讲师,甘肃经济发展数量分析研究中心特聘研究员,研究方向:统计理论与方法、经济计量分析。

一、引言

计量经济学是教育部规定核心课程,是经济类专业 8 门核心课程之一,是经济类专业的专业必修课。计量经济学作为一门经济学学科,在经济类的各个专业的教学中占有非常重要的地位。正如诺贝尔经济学奖获得者克莱因(R. Klein)所评价的,"在大多数大学和学院中,计量经济学的讲授已经成为经济学课程表中最有权威的一部分"。另一位诺贝尔经济学奖获得者萨缪尔森(P. Samuelson)甚至说,"二战后的经济学是计量经济学的时代"。[1]计量经济学是一门方法论学科,它的主要特点是理论与实际应用并重,既要详细阐述基本理论知识,又要注重经济计量方法在实践中的应用。在教学过程中,通过学习、掌握计量经济学的基本原理和常用方法,训练学生的创造性思维,并在此基础上培养学生运用定量方法分析和计量经济学模型解决现实经济问题的能力。课程在内容与应用上与微积分、线性代数、概率论与数理统计、统计学、经济学等课程有紧密的联系。[2]从学科性质来说,计量经济学是一门经济学科,或者说是经济学的一个分支学科,属于应用经济学范畴的一门文理渗透的交叉学科,而不是单纯的应用数学或统计学。

统计学作为一个应用经济学专业①,计量经济学在专业课程设置中居于重要地位,对于统计学专业而言,计量经济学不仅是一门经济类核心课程,更是一门极具统计特色的专业主干课程。因此,计量经济学是统计学专业课程的重中之重。显然,就教学来说,统计学专业的计量经济学课程建设勿容忽视。课程建设中最为重要的,是研究和设计合理科学的课程内容体系。和其他经济学科专业相比,统计学专业更注重培养学生的数理分析能力,统计学专业学生具有良好的数学基础,因此,强调数理分析方法的计量经济学课程在教学内容设置上应体现出统计学的专业特点,而不应与其他经济类专业的计量经济学教学内容体系相一致。本文通过对统计学专业所设置的主要课程进行分析,结合统计学专业的培养模式,分析统计学专业计量经济学课程教学的作用,为教学内容体系设计提供思路,供广大统计学教育工作者参考。

二、统计学专业的方法类课程及其与计量经济学的关系

经济学的研究离不开数理分析方法,而计量经济学是数理分析方法中最重要的部分,因为它实现了经济

① 当前,统计学已划归到理学之下,但就财经类院校而言,统计学专业培养模式和课程设置仍偏重于应用经济学,本文所述的也正是应用经济学下的统计学专业。

【统计理论与方法】

线性回归模型的统计检验关系辨析

刘 明
(兰州商学院 统计学院，甘肃 兰州 730020)

摘要：拟合优度检验(调整的可决系数)、模型整体显著性检验(F 检验)和单变量显著性检验(t 检验)是建立线性回归模型所需的三类基本统计检验方法，这三类检验既相互区别又紧密联系。通过推理论证发现：调整的可决系数和 F 统计量存在严格的对应关系，F 统计量也可用于反映线性回归模型的拟合效果；F 检验和 t 检验的检验对象和检验功能虽不同，但可看作是线性回归模型参数为零的约束检验的两种极端情况。

关键词：统计检验；调整的可决系数；F 统计量；t 统计量

中图分类号：O212 **文献标志码**：A **文章编号**：1007－3116(2011)04－0021－04

一、问题的提出

回归分析中对于线性回归模型存在三类基本的统计检验：模型的拟合优度检验、模型整体显著性检验和单变量(或称单参数)显著性检验。拟合优度检验是完全依赖样本的检验，它是由样本出发，检验样本回归直线对样本点的拟合效果的优劣，通俗地说，就是用于检验样本回归直线对所有样本点的综合代表性的好坏程度；模型整体显著性检验是一种联合检验，它检验所有的解释变量整体上对被解释变量的影响是否显著；单变量的显著性检验则是检验某单个解释变量对被解释变量有无显著的影响作用。模型的拟合优度检验通常是计算调整的可决系数 $R^{2①}$，据其大小对拟合效果的优劣性进行判定，而整体显著性检验和单变量显著性检验则分别使用 F 检验和 t 检验方法。R^2 的拟合优度检验、F 检验和 t 检验有着不同的检验对象、适用条件和范围，存在着明显的区别，这是人们所熟知的，但它们之间也存在着紧密的数理联系，在一定条件下可以通用。本文旨在通过对此三类检验方法之间的区别和适用范围做出系统阐述，重点以严格的推导证明过程来论证三类检验所依赖的统计量 R^2、F 统计量和 t 统计量之间的数理联系。

二、拟合优度检验、F 检验和 t 检验的构造及区别

考虑含有 k 个解释变量的线性回归总体模型和样本模型：

$$y_i = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_j x_j + \cdots + \beta_k x_{ki} + \mu_i$$

$$y_i = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_j x_j + \cdots + \beta_k x_{ki} + \mu_i$$

定义总离差平方和 TSS(Total Sum of Squares)、可解释的平方和 ESS(Explained Sum of Squares)和剩余平方和 RSS(Residual Sum of Squares)为：

$$\mathrm{TSS} = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

$$\mathrm{ESS} = \sum_{i=1}^{n} (\hat{y_i} - \bar{y})^2$$

$$\mathrm{RSS} = \sum_{i=1}^{n} (y_i - \hat{y})^2$$

其中 $\hat{y_i} = y_i - \mu_i$ 为样本拟合值，$\bar{y}$ 为样本均值，n 为样本容量。容易证明：TSS、ESS、RSS 的自由度分别为 $n-1$，k，$n-k-1$；TSS = ESS + RSS。

模型的拟合优度检验使用调整的可决系数 R^2，它被定义为：

收稿日期：2010－12－26

作者简介：刘 明，男，安徽霍邱人，讲师，经济学硕士，研究方向：经济计量分析。

①对于一元线性回归模型的拟合优度检验可用可决系数 R^2。拟合优度检验常用的指标还包括赤池信息准则(AIC)和施瓦茨准则(SC)，本文只讨论调整的可决系数 R^2。

DOI:10.13546/j.cnki.tjyjc.2013.09.003

线性回归模型几何分析：基于估计结果及检验统计量的考察

黄恒君[ab]，刘　明[ab]

（兰州商学院 a.统计学院，b.甘肃省经济发展数量分析研究中心，兰州 730020）

摘　要：文章对普通最小二乘估计中形成的相关系数、回归系数、可决系数R^2、F统计量，以及多重共线性问题作出几何分析，指出该分析方法表现为向量的长度和角度关系。这种分析过程使普通最小二乘法及估计结果变得更直观。

关键词：普通最小二乘估计；检验统计量；欧氏空间

中图分类号：F222.1　　**文献标识码**：A　　**文章编号**：1002-6487(2013)09-0017-03

0　引言

线性回归模型最基本的参数估计方法是普通最小二乘法，在最小二乘法下线性回归模型具有良好的统计性质，从数理的角度来分析，最小二乘估计方法及估计结果的数学特征也非常显著，用数学思想对最小二乘估计方法及估计结果进行分析和研究，能够将其直观化、系统化。我们曾利用正交和投影对普通最小二乘法原理作出了几何解释[1]，在此基础上，本文利用“长度”和“角度”等几何概念，分别对普通最小二乘估计形成的回归系数、相关系数、可决系数R^2、F统计量作出几何分析，以期用数学语言表

作者简介：黄恒君(1981-)，男，浙江温州人，博士研究生，讲师，研究方向：调查技术与统计分析。
刘　明(1981-)，男，安徽霍邱人，硕士，讲师，研究方向：经济计量分析。

3.2　校正的PC方法和校正的BCa方法比较

对表2～表4校正的boot-p、boot-pi和boot-i策略的PC方法和BCa方法的包含率进行整理，结果如表5所示。

表5　校正的boot-p、boot-pi和boot-i策略的PC和BCa方法估计的置信区间包含率

	PC	BCa	Bootstrap策略	方差分量置信区间	分布
1	0.769	0.774	boot-p	CI(vc.p)	Normal
2	0.768	0.773	boot-p	CI(vc.p)	Dichotomous
3	0.783	0.794	boot-p	CI(vc.p)	Polytomous
4	0.737	0.740	boot-p	CI(vc.p)	$\beta=-2.0$
5	0.775	0.778	boot-p	CI(vc.p)	$\beta=-1.0$
6	0.767	0.766	boot-p	CI(vc.p)	$\beta=0$
7	0.799	0.802	boot-pi	CI(vc.i)	Normal
8	0.796	0.847	boot-pi	CI(vc.i)	Dichotomous
9	0.671	0.682	boot-pi	CI(vc.i)	Polytomous
10	0.672	0.672	boot-pi	CI(vc.i)	$\beta=-2.0$
11	0.716	0.726	boot-pi	CI(vc.i)	$\beta=-1.0$
12	0.710	0.706	boot-pi	CI(vc.pi)	$\beta=0$
13	0.847	0.846	boot-i	CI(vc.pi)	Normal
14	0.799	0.815	boot-i	CI(vc.pi)	Dichotomous
15	0.838	0.843	boot-i	CI(vc.pi)	Polytomous
16	0.812	0.812	boot-i	CI(vc.pi)	$\beta=-2.0$
17	0.784	0.782	boot-i	CI(vc.pi)	$\beta=-1.0$
18	0.786	0.786	boot-i	CI(vc.pi)	$\beta=0$

使用配对样本T检验比较表5中PC和BCa方法估计的置信区间包含率，结果表明两种方法差异显著且效果量较大(t=2.197，p=0.420<0.05，d=0.842)。因此，可以认为，对于概化理论方差分量置信区间，校正的Bootstrap的BCa方法要优于校正的Bootstrap的PC方法。

4　结论

(1)与未校正的方法相比，校正的Bootstrap的PC和BCa方法估计概化理论的方差分量置信区间更为可靠。

(2)校正的Bootstrap的BCa方法估计概化理论的方差分量置信区间，要优于校正的Bootstrap的PC方法。

参考文献：

[1]Gao, X. H.,Brennan, R. L.Variability of Estimated Variance Components and Related Statistics in a Performance Assessment[J].Applied Measurement in Education, 2001,14(2).

[2]Efron, B. Bootstrap Method: another Look at the Jackknife[J]. Annals of Statistics,1979,7(1).

[3]Fan, X. T. Using Commonly Available Software for Bootstrapping in Both Substantive and Measurement Analyses[J]. Educational and Psychological Measurement , 2003,63 (1).

[4]Efron, B. The Jackknife, the Bootstrap and other Resampling Plans [C]. SIAM CBMS-NSF Monograph 38,1982.

[5]张敏强. 教育与心理统计学(第三版)[M].北京：人民教育出版社，2010.

[6]Efron, B.,Tibshrani, R. An Introduction to the Bootstrap[M].New York: Chapman and Hall,1993.

[7]黎光明，张敏强. 基于概化理论的方差分量变异量估计[J].心理学报，2009,41(9).

(责任编辑/亦　民)

线性回归模型统计检验方法体系构建

刘　明[1,2]，李明莉[3]

（1.兰州商学院 甘肃经济发展数量分析研究中心，兰州 730020；2.中国人民大学 统计学院，北京 100872；
3.中石油第二建设公司，甘肃 兰州 730000）

摘　要：文章一方面依据统计检验过程中检验统计量的分布形态、检验指标的特征，另一方面依据模型的预设条件、参数及模型的显著性、拟合效果等检验对象，构建了以检验统计量为主导和以模型为主导的两组线性回归模型的检验方法体系。

关键词：线性回归模型；统计检验

中图分类号：O212　　**文献标识码**：A　　**文章编号**：1002-6487(2014)02-0008-004

0　引言

线性回归模型是最基础、最常用的计量经济学模型，统计检验是计量经济学建模过程中的重点环节，一个好的计量经济学模型必须首先经得起各类统计检验，线性回归模型的构建亦须如此。线性回归模型的统计检验方法很多，在这些检验方法中，有很多是相互关联的，有着相似甚至相同的特征，本文通过探寻这些特征并进一步研究检验方法间共同的规律，依据这些相似或共同的特征及规律，构造线性回归模型的检验方法体系，以期解决线性回归模型统计检验方法冗杂的问题。

1　经典参数检验方法

总的来看，线性回归模型经典参数检验方法主要是指最常见的t分布、F分布、χ^2分布等三大分布检验，即通常所说的t检验、F检验和χ^2检验。这三类检验在统计学中居于基础性的地位，在线性回归模型的各类检验中，此三类检验是最为常见的，且称之为经典参数检验方法。标准正态分布的Z检验也是基础检验，但在线性回归模型的检验中很少用到，因此本文不对Z检验进行单列详述。

1.1　t检验

t检验是通过构造服从t分布的检验统计量来完成的假设检验。在线性回归模型的检验中，t统计量的形式一般为：

$$t=\frac{\hat{\beta}-\beta^*}{se(\hat{\beta})}$$

其中$\hat{\beta}$为参数β的点估计量，$se(\hat{\beta})$为$\hat{\beta}$的样本标准差，β^*是待检验的参数值，检验目标通常是$\beta=\beta^*$，$\beta>\beta^*$及$\beta<\beta^*$。

在线性回归模型中，t检验最常用的功能是检验单个解释变量对被解释变量是否存在显著的影响作用，即检验(偏)回归系数是否为零，也就是通常所说的单参数显著性检验；t检验还常用于检验(偏)回归系数是否大于、等于或小于某一给定值，以考察参数取值特征；在约束性检验中，t检验可以用以诊断模型中的参数是否相等。用于检验方程联立性问题的豪斯曼设定检验(Hausman Specification Test)[1-2]，就可以通过t检验来完成的。

1.2　F检验

简单的说，F检验就是通过构建服从F分布的统计量来完成的检验。一般的F检验都可以归结为线性约束条件的检验，检验统计量为：

$$F=\frac{(RSS^*-RSS)/p}{RSS/(n-k-1)}$$

式中，RSS表示原模型的剩余平方和，RSS^*为受约束模型剩余平方和，k为模型变量个数，n是样本容量，p是约束条件个数。此F统计量还可以写为：

$$F=\frac{(R^2-R^{*2})/p}{(1-R^2)/(n-k-1)}$$

其中R^2是原回归模型的样本可决系数，R^{*2}是受约束回归模型的样本可决系数。

最为常见的F检验当属模型整体显著性检验，即检验所有解释变量联合起来(在整体上)对被解释变量是否存在影响作用，它是线性约束检验的一种特殊形式，其检验统计量可表示为：

$$F=\frac{ESS/k}{RSS/(n-k-1)}=\frac{R^2/k}{(1-R^2)/(n-k-1)}$$

其中ESS为待检验模型的回归平方和，其余符号和前文相同。在线性回归模型中，有很多检验，诸如异方差检验(格德菲尔德-匡特检验(Goldfeld-Quandt Test，G-Q Test)[3]、帕克检验(Park Test)[4]、怀特检验(White Test))、格兰杰因果关系检验(Granger Causality Test)[5]、邹氏检验

作者简介：刘　明(1981-)，男，安徽霍邱人，博士研究生，讲师，研究方向：统计理论与方法、经济计量分析。

【统计理论与方法】

一类新的多重共线性检验方法

刘 明[a,b]

(兰州商学院 a. 统计学院 b. 甘肃经济发展数量分析研究中心, 甘肃 兰州 730020)

摘要:解释变量间的相关性导致了多元线性回归模型的多重共线性问题,由于考察相关性的角度和方法不同,产生了不同的多重共线性的检验方法。由阿达马不等式可以构建多个变量的综合相关性度量指标,将该指标用于度量多元线性回归模型的解释变量的综合相关程度,用以作为多元线性回归模型多重共线性的一类检验方法。

关键词:多重共线性;阿达马不等式;检验方法

中图分类号:O212 文献标志码:A 文章编号:1007-3116(2012)10-0014-03

线性回归模型的多重共线性的本质是解释变量之间存在线性相关,它是一种样本现象,是由样本数据引起的。由于经济数据之间的相关性,多重共线性普遍存在于多元线性回归模型中,因此人们所关注的是多重共线性是否严重这一问题而非是否存在。多重共线性的严重程度实质上表现为解释变量间线性相关的密切程度,解释变量间的相关性越强,多重共线性问题越严重。对于多重共线性严重程度的检验方法有多种,这些检验方法都存在着一定的缺陷,笔者依据数学中行列式运算的阿达马不等式,提出一类新的多重共线性的检验方法。

一、常用的多重共线性检验方法

进行多重共线性检验最简单的方法就是利用解释变量间的简单线性相关系数,通过其大小就可以判断两个解释变量间是否存在多重共线性[1]117-124。这种检验方法简单直观,直接依据多重共线性的本质,但检验结论并不完全可靠,因为只考察了解释变量两两之间的相关性,没有综合考虑,也就是说,如果相关系数很大,则一定存在多重共线性;如果相关系数很小,不一定没有多重共线性。

解释变量之间存在多重共线性就是至少存在某一个解释变量可以近似地由其他解释变量线性表出。显然,寻找这种线性表达式的方法就是将每个解释变量对其余解释变量进行回归,得到 k 个回归模型(即所谓的辅助回归模型,k 为解释变量的个数),进一步计算出每一个辅助回归模型的可决系数 R^2,如果其中最大的一个 R^2 接近于 1,则模型存在多重共线性,这种方法通常被称为辅助回归模型检验法。辅助回归模型的可决系数 R^2,从本质上来说是某解释变量与其余解释变量间的复相关系数的平方,复相关系数考虑了所有的解释变量,因此比利用简单相关系数进行检验更可靠,但它没有全面考虑到解释变量间线性组合的相关性,因而也不能全面衡量多重共线性问题。在辅助回归模型基础上可以进一步利用方差膨胀因子进行多重共线性检验,方差膨胀因子为:

$$\mathrm{VIF}_i = \frac{1}{1-R_i^2}$$

方差膨胀因子越大,表明解释变量之间的多重共线性越严重。方差膨胀因子检验和辅助回归模型检验本质上是相同的。

利用解释变量矩阵的特征值也可以进行多重共线性检验。解释变量的样本数据矩阵 $|X'X|$ 可以分解为特征值 $\lambda_i(i=1,2,\cdots,k)$ 的连乘积的形式,如果 λ_i 中有一个或几个近似等于 0,则 $|X'X|\approx 0$,即解释变量存在多重共线性。利用特征值构建指标:

$$\mathrm{CN}=\frac{\max(\lambda_i)}{\min(\lambda_i)}$$

收稿日期:2012-02-08

作者简介:刘 明,男,安徽霍邱人,讲师,经济学硕士,研究方向:统计理论与方法,经济计量分析。

【统计理论与方法】

异方差 White 检验应用的几个问题

刘 明[a,b]

(兰州商学院 a. 统计学院; b. 甘肃经济发展数量分析研究中心, 甘肃 兰州 730020)

摘要:White 检验通常由 LM 检验来完成,现实中也可以使用更简单的 F 检验来替代 LM 检验;White 检验过程中需要构筑辅助回归模型,它可以有不同的形式以应对实际问题的需要;讨论了四类不同的辅助回归模型,理论分析表明,它们在应用中各有所长;最后通过一个实际例子,验证了理论分析的结论,展示了 White 检验应用的灵活性。

关键词:异方差;White 检验;辅助回归模型

中图分类号:O212 文献标志码:A 文章编号:1007-3116(2012)06-0045-05

异方差是计量经济学中所讨论的一个重要问题,异方差的检验方法在异方差问题的讨论中居于核心地位。在线性回归模型中,异方差的检验方法有很多,依据不同的前提条件,这些检验又各有所倚。在诸多异方差检验方法中,最有效、最常用的当属 White 检验了,这种检验方法提出以后,以其快捷的检验过程和有效的检验结论而得到广大学者的认可,并迅速成为异方差检验的经典方法。笔者在计量经济学课程教学过程中发现,White 检验是初学者的一个学习难点,很多初学者只懂得 White 检验的思路和方法,而不会对其加以灵活应用。本文将对 White 检验的使用技巧进行总结提炼,为初学者提供学习便利,为研究者提供借鉴。

一、White 检验方法概述

White 检验是 Halbert White 在 1980 年提出的,他在随机项存在异方差的情形下,构造出参数估计量方差协方差矩阵的一致估计,并根据这个估计结果,导出了一个服从 χ^2 分布的统计量,即拉格朗日乘数(LM),用此统计量可完成异方差的检验[1]。在统计学中,通过构造 LM 统计量进行的检验也可简称为 LM 检验。为避免繁杂的数学推导公式,这里借用普通教科书中的关于 White 检验的表述内容,并以此作为讨论的开始。对于更严谨的、一般化的论述,可参见 Halbert White 所述[1]。

White 检验基于这样的一个前提:随机干扰项的异方差和解释变量有关。经验证明这是合乎实际的。以二元线性回归模型为例的 White 检验方法如下[2]413-414:

设二元线性回归模型:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \mu_i \quad (1)$$

该模型存在异方差:$\mathrm{var}(\mu_i) = \sigma_i^2$

第一步,构造检验辅助回归模型:

$$\sigma_i^2 = \alpha_0 + \alpha_1 X_{1i} + \alpha_2 X_{2i} + \alpha_3 X_{1i}^2 + \alpha_4 X_{2i}^2 + \alpha_5 X_{1i} X_{2i} + \upsilon_i \quad (2)$$

式(2)表明随机项的方差受到解释变量的影响。

第二步,提出假设 $H_0: \alpha_i = 0, (i = 1,2,3,4,5)$,即原模型不存在异方差;$H_1: \alpha_i (i = 1,2,3,4,5)$ 不同时等于零,即原模型存在异方差。

第三步,估计辅助回归模型。首先用 OLS 方法估计式(1)得到残差 e_i 并取平方得 e_i^2,再利用 e_i^2 估计出辅助回归模型。进一步计算出辅助回归模型的多重可决系数 R^2。

第四步,构造并计算检验统计量 nR^2,即拉格朗日乘数。该统计量服从自由度为约束条件个数的 χ^2 分布。式中 n 为样本容量,R^2 为辅助回归模型的多重可决系数。此处约束条件个数为 5。

第五步,做出检验决策。如果 nR^2 大于 χ^2 分布

收稿日期:2012-02-18

作者简介:刘 明,男,安徽霍邱人,经济学硕士,讲师,研究方向:统计理论与方法,经济计量分析。

DOI:10.13546/j.cnki.tjyjc.2012.20.012

最小一乘法与最小二乘法:基于例证的比较

刘　明[a,b]

（兰州商学院 a.统计学院;b.甘肃经济发展数量分析研究中心,兰州 730020）

摘　要:最小一乘法和最小二乘法在估计思想上有着相同的渊源,而在实现路径上有所不同:最小一乘法属于中位数回归而最小二乘法属于均值回归。由此,两者在回归系数的计算、回归直线的性质和估计结果等方面均存在较大差异。文章在理论分析的基础上进一步通过例证,将两类估计方法在计算、优劣势和应用范围做出了比较和分析。

关键词:最小一乘法;最小二乘法;比较分析

中图分类号:O212　　**文献标识码**:A　　**文章编号**:1002-6487(2012)20-0012-03

1　文献回顾

最小二乘法是参数估计中最基础、最重要的方法之一,该方法以其估计量的优良的统计性质和简便的数学运算而著称,在数学、统计学、经济学中有着广泛的应用。与最小二乘法具有相似估计思想的最小一乘法,在线性回归模型参数估计中也有着重要应用。事实上,最小一乘法的提出比最小二乘法还要早,1760年波斯科维奇在研究子午线长度问题时提出最小一乘法;45年后,即1805年,法国数学家列让德提出了最小二乘法。对最小二乘法做出最大贡献的当属高斯,他对最小二乘的估计方法、估计量的性质等问题做了详细研究,构筑了近代最小二乘法在统计学中广泛应用的基础。最小二乘法发展至今已经较为成熟,而最小一乘法在其提出以后的近200年中几乎没有发展,最根本的原因是最小一乘法在估计值的计算上较为困难,直到上世纪5、60年代,人们才开始对最小一乘法予以关注,原因是计算数学的兴起和计算机技术的发展,使得最小一乘法的参数估计值的计算得以实现。在我国,对于最小一乘法的研究始于上世纪80年代。陈希孺对最小一乘准则下线性回归模型回归系数的计算、分布形态、性质以及假设检验等问题进行了详细的研究,包括一元线性回归模型和多元线性回归模型,他的研究结论包括使用规划方法计算回归系数、最小一乘估计值的渐进正态分布以及构建F统计量进行回归系数的显著性检验等[1]。在此之后的研究主要集中在最小一乘法参数估算方法方面。董祺提出了"残差绝对值和最小"准则的松弛算法,这是最小一乘准则下的参数估计值计算方法[2]。袁修贵、李显方等分别对最小一乘准则下的回归系数的计算方法目标规划法、搜索算法进行了论述和探讨[3,4]。李德志针对一元线性回归模型的最小一乘估计,提出双样本点算法,即构造两个虚拟的样本点,通过此二样本点确定最小一乘回归直线。徐龙封讨论了 L^P 空间上线性回归方程回归系数的估计问题,他将回归系数的估计转化为数学规划模型,通过计算机程序进行求解,最小一乘法和最小二乘法正是他所讨论的问题的特殊形式[6]。王福昌根据最小一乘的性质,把最小一乘问题变为组合优化问题,将模拟退火算法用在最小一乘模型的求解上,取得了较好的计算结果[7]。在最小一乘估计结果的分析和评价方面,谢开贵等论证了最小一乘回归直线的一些性质,并通过实例说明了最小一乘估计结果的稳健性[8]。朱春浩、冯守平等对最小一乘法和最小二乘法的发展和计算作出了概括性的介绍和比较[9~10]。关于最小一乘估计结果的分析和评价的文献相对较少,这方面的研究相对薄弱。本文将以最小一乘法和最小二乘法的思想脉络为研究路径,在考察二者的思想起源、实现方法及在现实中的应用的基础上,将它们进行比较,得出相关研究结论,以期更清楚地认识它们。

2　最小一乘法与最小二乘法的思想及实现路径

最小一乘法与最小二乘法最初来源于数学中对线性方程组求解问题的研究。考虑一个 k 元非齐次线性方程组(1):

$$\begin{cases} a_{11}x_1+a_{21}x_2+\cdots+a_{k1}x_k=b_1 \\ a_{12}x_1+a_{22}x_2+\cdots+a_{k2}x_k=b_2 \\ \cdots\cdots \\ a_{1n}x_1+a_{2n}x_2+\cdots+a_{kn}x_k=b_n \end{cases} \quad (1)$$

在数学上讨论方程组(1)时,通常是 $n\leqslant k$ 的情形,此时方程组(1)有解的充分必要条件是系数矩阵[a_{ij}]的秩与增广矩阵[$a_{ij}\ b_j$]的秩相等。当 $n>k$ 时,该方程组是无解的(不考虑方程组中存在等价方程,因为此时可视为 $n\leqslant k$ 的情形)。在实际问题的研究中,这种情形是普遍存在的。由于此时无法得到一组确切的解,使得方程组中每一等式同时成立,于是只有退而求其次,寻找方程组中未知数的一组解,使得每一方程左右两边的数值尽可能的接近。为了实现这一目的,有两种方法可供使用,第一种方

作者简介:刘　明(1981-),男,安徽霍邱人,硕士,讲师,研究方向:统计理论与方法、经济计量分析。

【统计理论与方法】

基于 B-样条基底展开的曲线聚类方法

黄恒君[a,b]

(兰州商学院 a. 统计学院;b. 甘肃省经济发展数量分析研究中心,甘肃 兰州 730020)

摘要:随着大数据时代的来临,近年来函数型数据分析方法成为研究的热点问题,针对曲线的聚类分析方法引起了学界的关注。给出一种曲线聚类的方法:以 L^2 距离作为亲疏程度的度量,在 B-样条基底函数展开表述下,将曲线本身信息、曲线变化信息引入聚类算法构建,并实现了曲线聚类与传统多元统计聚类方法的对接。作为应用,以城乡收入函数聚类实例验证了该曲线聚类方法,结果表明,在引入曲线变化信息的情况下,比仅考虑曲线本身信息能够取得更好的聚类效果。

关键词:函数型数据;大数据;曲线聚类;B-样条

中图分类号:O212.4 文献标志码:A 文章编号:1007-3116(2013)09-0003-06

一、问题的提出

在统计研究和统计工作中,人们经常面临截面数据、时序数据,以及相对复杂的面板数据、纵向数据的处理和分析问题。随着大数据时代的来临,在统计数据结构越来越复杂的同时,数据采集的密集化程度也越来越高,随之出现了一种有别于传统形式、具有明显函数特征的数据类型。如心理学研究中的脑电信号数据、生物技术中的基因微序列数据、化学计量中的光谱分析数据、经济研究中的股票分时成交价数据、地区的人均收入水平数据等,都表现出明显的函数特征。当前文献中将这种数据类型称为函数型数据[1]1-18。

对于这种数据类型,人们通常能够采用传统的统计工具(如多元统计方法)进行分析。但由于未能考虑到数据的函数特性,传统统计工具除了计算过程中可能伴随奇异矩阵处理的技术问题外,更加重要的是,函数信息丢失使得使用传统统计工具不利于函数型数据中所包含信息的全面挖掘,因而往往不能取得很好的分析效果。

为此,人们有针对性地开发函数型数据分析方法[2-3]。就目前的研究来看,函数型数据分析的大多数研究内容可以看成是有限维多元分析方法向无限维函数的延伸和拓展:一方面,从延伸性角度讲,对函数型数据进行分析的过程中,往往要将无限维对象投影到有限维空间进行分析,多元统计中的绝大多数方法都能在函数型数据场合找到类似模型。另一方面,从拓展性视角看,正是由于数据本身的函数特征,函数型数据分析中许多内容,诸如数据平滑处理技术(通常表现为非参数回归)、泛函分析工具运用(通常表现为 Hilbert 空间上的线性算子)以及针对函数的一些假定(如平方可积性)等,是多元统计分析无法涉及的。

聚类分析作为一种重要的多元统计方法,在许多领域得到广泛的应用,其基本思想是依据数据之间的亲疏程度,对样品或变量进行归类。经典的聚类算法包括系统聚类、动态聚类等[4]228-252。依据前面的描述,可以将多元统计中聚类分析的思想推广到函数型数据的情形。当前研究中,曲线聚类的方法引起了人们的研究兴趣,包括曲线的分割归类和曲线的整体聚类[5]。本文针对后者进行讨论。

无论数据本身是否具有函数特征,其观测值往往是离散的。在曲线得到拟合的基础上,目前对曲线聚类主要包括两种做法:一是在对所拟合曲线协方差算子谱分解(Karhumen-Loêve)展开)的基础上,进行降维聚类算法构建[6];二是将无限维函数投影到基底函

收稿日期:2013-04-21;修复日期:2013-06-18

基金项目:国家社会科学基金项目《中国现行社会福利保障制度下城镇贫困人口的统计研究》(11BTJ002)

作者简介:黄恒君,男,浙江温州人,经济学博士,副教授,研究方向:调查技术与统计分析。

DOI:10.13546/j.cnki.tjyjc.2012.06.012

基于t检验的逐步回归的改进

刘　明[1],王仁曾[2]

(1.兰州商学院 统计学院,兰州 730020;2.华南理工大学 经济贸易学院,广州 510006)

摘　要:传统的逐步回归是依据偏回归平方和所构造的F统计量进行F检验来完成的,F检验的目的是判断变量是否应该引入或剔除。文章通过论证发现,在普通最小二乘估计下,逐步回归中的F检验和对应解释变量的显著性t检验是等价的,利用t检验同样可以完成逐步回归,t检验下的逐步回归结果和F检验下的逐步回归结果一致。

关键词:逐步回归;F检验;t检验

中图分类号:O212　　**文献标识码**:A　　**文章编号**:1002-6487(2012)06-0016-04

1　问题的提出

逐步回归是线性回归分析中重要的一种分析方法,主要用来解决多元线性回归模型中解释变量个数较多时如何选择解释变量,以使得在回归方程中包含所有对被解释变量影响显著的解释变量而不包含影响不显著的解释变量的问题。逐步回归正是为解决这类问题而设计的一种回归方法。它的主要思路是在所考虑的全部解释变量中按对被解释变量的贡献大小逐个引入回归方程,已被引入回归方程的变量在引入新变量后也可能失去重要性,而需要从回归方程中剔除出去。引入一个变量或者从回归方程中剔除一个变量都要进行F检验,以保证在引入新变量前回归方程中只含有对被解释变量影响显著的变量,而不显著的变量已被剔除[1]。

在逐步回归中每剔除和引入一个变量都需要计算F统计量的值,这需要一定的工作量。同时,逐步回归中所用的F检验对于众多初学者和应用者来说也难以理解和把握,而单个参数显著性t检验是人们所熟知的。笔者通过研究发现,F统计量和t统计量存在紧密的联系,逐步回归中的F检验和参数显著性t检验是等价的,因此可以转而考虑使用t检验。相比较而言,t统计量的计算要比F统计量的计算简便得多,F统计量需要计算复杂的偏回归平方和及剩余平方和,而t统计量只需要计算回归系数的估计值及其估计量的标准差的古计量即可。现代常用的统计软件一般都会计算显示回归模型参数的t检验值,而很少会给出用于逐步回归的F检验值,即便使用计算机,F统计量也不易计算。本文考虑用t检验准则替代F检验准则对多元线性模型进行逐步回归,以简化逐步回归的计算过程。要实现这一目标,需分析逐步回归中的F检验,并完成其与t检验的等价性的证明。

2　逐步回归中的F检验及其与t检验的等价关系

考虑含有k个解释变量的线性总体回归模型式(1)和普通最小二乘法(本文均在普通最小二乘法下讨论样本回归模型)下的样本回归模型式(2):

$$y_i=\beta_0+\beta_1x_{1i}+\beta_2x_{2i}+\ldots+\beta_kx_{ki}+\mu_i \quad (1)$$

$$y_i=\hat{\beta}_0+\hat{\beta}_1x_{1i}+\hat{\beta}_2x_{2i}+\ldots+\hat{\beta}_kx_{ki}+\hat{\mu}_i \quad (2)$$

首先定义总离差平方和TSS(Total Sum of Squares)、可解释的平方和ESS(Explained Sum of Squares)和剩余平方和RSS(Residual Sum of Squares):

$$TSS=\sum_{i=1}^{n}(y_i-\bar{y})^2,\ ESS=\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2,\ RSS=\sum_{i=1}^{n}(y_i-\hat{y})^2$$

其中 $\hat{y}_i=y_i-\hat{\mu}_i$ 为样本拟合值,$\bar{y}$ 为样本均值,n 为样本容量。

再定义偏回归平方和。不含 x_k 的样本回归模型(为方便分析,在每一步对解释变量的考察中,本文均以 x_k 为研究代表)

$$y_i=\hat{\beta}_0+\hat{\beta}_1x_{1i}+\hat{\beta}_2x_{2i}+\ldots+\hat{\beta}_{k-1}x_{k-1i}+\hat{\mu}_{ri} \quad (3)$$

设式(3)的可解释的平方和为 ESS^*,剩余平方和为 RSS^*,则 x_k 的偏回归平方和定义为:

$$ESSP_k=ESS-ESS^*$$

按此法即可定义其他解释变量的偏回归平方和。不难看出,$ESSP_k=RSS^*-RSS$。

逐步回归中引入(剔除)解释变量的标准是偏回归平方和最大(最小)。在某一引入步骤中,设模型中已含有 $k-1$ 个解释变量(全部显著),需引入第 k 个解释变量,原模型和引入变量之后的样本模型即可分别表示为式(3)和式(2)。

这样由第 k 个解释变量 x_k 的偏回归平方和 $ESSP_k$ 构造的F统计量为:

$$F=\frac{ESSP_k}{(TSS-ESS)/(n-k-1)}=\frac{RSS^*-RSS}{RSS/(n-k-1)}\sim F(1,n-k-1) \quad (4)$$

作者简介:刘　明(1981-),男,安徽霍邱人,硕士,讲师,研究方向:统计理论与方法、经济计量分析。

DOI:10.13546/j.cnki.tjyjc.2013.20.001

基于函数型主成分的收入分布变迁特征探索

黄恒君[a,b]

（兰州商学院 a.统计学院；b.甘肃省经济发展数量分析研究中心，兰州 730020）

摘　要：文章提出基于函数的收入分布变迁分析方法：在拟合收入分布函数序列的基础上，进行函数型主成分分析。利用该建模思路，对2000~2010年中国城镇居民收入分布变迁规律进行探索分析。

关键词：函数型主成分；收入分布变迁；收入差距

中图分类号：O212.1　　**文献标识码**：A　　**文章编号**：1002-6487(2013)20-0024-03

0　引言

居民收入问题的研究历来是国内外经济学界关注的焦点。近年来，收入差距在一定程度上迅速扩大，其引发的诸多经济社会问题倍受瞩目，相关研究广泛而深入。拟合函数(如收入分布、洛伦兹曲线)，估算、分解并分析不平等指标，是目前关于收入不平等测度研究的重要思路。

事实上，收入变动的最新研究强调对整个收入函数进行分解，这意味着跳过不平等指标的估算，直接对收入函数进行分析，已有学者重视基于收入函数的研究[1,2]。这些研究直接在收入函数上进行分析，避免了不平等指标估算过程中的信息损失。然而，位置-尺度估计仍是将在汇总方式下建立收入分布之间的联系，并非收入分布函数之间逐点对应的直接联系。因此，诸如收入分布的变迁由哪些收入群体引起，不同收入群体对收入分布变迁的贡献程度，难以作出回答。

为了刻画这些问题，需要引入以函数为元素的空间，在收入函数上建立统计推断模型。从这个角度讲，基于位置-尺度的分析还是初步的，不能完全刻画收入函数的动态特征。鉴于此，本文利用泛函分析的思路，对历年收入分布变迁规律进行探索分析。

1　函数型数据生成及分析模型

本文借鉴函数型数据分析思想，对收入分布变迁的探索分析可分为两个步骤：第一，生成函数型数据(收入分布函数)作为“样本”；第二，利用函数型主成分探索收入分布函数的变迁信息。

然而，由于分布函数取值的特殊性，分布函数的非参数拟合往往是一个带约束条件的优化问题，很难找到关于一组函数的线性组合去表述分布函数，而函数型主成分分析需要在线性函数空间下才能进行。为解决这一矛盾，本文处理思路如下：第一，采用B-样条基底，以类似于Logistic分布的形式拟合收入分布函数；第二，将拟合的分布函数进行变换，转化为取值无约束的函数；第三，对取值无约束的函数进行函数型主成分分析；第四，将分析结果代回收入分布函数。具体模型如下：

1.1　函数型数据生成——收入分布函数序列拟合

记收入变量为y，第i年($i=1,2,\cdots,N$)的收入分布函数为$F_i(y)$，形式为

$$F_i(y)=1-[1+\exp(w_i(y))]^{-1} \tag{1}$$

其中$w_i(y)$为待估函数。

该收入分布可以视为对一组函数基底线性组合的Logit变换：根据式(1)的表述，有$w_i(y)=\ln[F_i(y)/(1+F_i(y))]$，根据逼近理论，可以选取一组B-样条基底$\phi_k(y)(k=1,2,\cdots,K)$使得$w_i(y)$近似表述为

$$w_i(y)\doteq\sum_{k=1}^{K}\beta_k\phi_k(y) \tag{2}$$

其中，β_k为待估参数。采用最小二乘法可得到β_k的估计$\hat{\beta}_k$，并根据式(2)得到$w_i(y)$的估计，将结果代回式(1)，可以计算收入分布函数。该方法在数据拟合上具有“自适应”能力，就本文所使用的数据而言，其拟合优度高于常规的参数方法，且避免了对不同年份数据分布函数形式不统一的问题。

1.2　函数型主成分分析——收入分布函数特征提取方法

函数型主成分分析基本思想与传统主成分分析类似，是有限维度函数线性组合对无限维的近似。结合上述收入分布函数拟合过程，本文采用基底函数算法求解函数型主成分。

前面指出，式(1)关于基底函数$\phi_k(y)$是非线性的，而式(2)关于基底$\phi_k(y)$是线性的，为了对收入分布函数的波动特征进行分析，本文首先对$w_i(y)$进行函数型主成分分析，将分析结果代入式(1)，即可得到收入分布函数的波动特征

基金项目：国家社会科学基金资助项目(11BTJ002)

作者简介：黄恒君(1981-)，男，浙江温州人，博士研究生，讲师，研究方向：调查技术与统计分析。

Durbin-Watson 自相关检验应用问题探讨

刘 明 王永瑜

（1. 兰州商学院统计学院；
2. 兰州商学院甘肃经济发展数量分析研究中心）

【摘要】通过对 Durbin-Watson 自相关检验方法的阐述，进一步讨论这类检验方法在应用时先设条件存在的原因，并由此引出另一些不被注意的先设条件，随后对 Durbin-Watson 检验中的临界值和判别区间问题进行探讨，给出临界上限取值大于 2 情形下的检验方法并结合案例进行分析。

关键词 Durbin-Watson 检验 先设条件 临界上限 判别区间

中图分类号 O212 **文献标识码** A **JEL 分类号** C12

DOI:10.13653/j.cnki.jqte.2014.06.011

Exploring in Durbin-Watson Autocorrelation Test

Abstract: Through the description of Durbin-Watson autocorrelation test method, this paper gives a further discussion about the reasons of using conditions of Durbin-Watson autocorrelation test method in application, and some other using conditions which is not noticed. It explores the critical value and discriminant interval problem of Durbin-Watson test, and gives Durbin-Watson test method with an example when the upper critical value is greater than 2.

Key words: Durbin-Watson Test; Using Condition; Upper Critical Value; Discriminant Interval

引 言

在一般的计量经济学教科书中，Durbin-Watson 自相关检验（D-W 检验）是重点介绍的自相关检验方法。这种检验方法在实际应用中虽然有诸多弊端，但因其检验过程简单、便于操作而受到人们的广泛关注及使用。而现实中，人们在正确使用这种方法进行计量经济学模型自相关检验的时候，却很少有人去探究“为什么需要这样”等问题，例如为什么进行 D-W 检验的模型需要满足一些先设条件？或许这些问题并不会影响建模者对该检验方法的正确使用，但我们认为，弄清楚这些问题不仅仅可以帮助人们对 D-W 检验全面认识与理解，更能帮助使用者在实际中对这一检验方法的灵活运用。本文从教科书中的D-W检验方法入手，详细探究这类检验方法在实际应用中常见的问题，以求对 D-W 检验达到更全面的认识。

一、D-W 检验的基本原理

对于线性回归模型：

$$Y = X\beta + \varepsilon \tag{1}$$

DOI:10.13546/j.cnki.tjyjc.2014.08.006

经济时间序列的ARIMA类模型构建

刘　明[a,b]

（兰州商学院 a.甘肃经济发展数量分析研究中心；b.统计学院，兰州 730020）

摘　要：时间序列模型在经济问题分析研究中具有重要作用。对经济时间序列建立ARIMA类模型包括平稳性检验、模型阶数识别、参数估计和模型检验等过程。建模过程中路径选择、季节效应处理、检验方法的综合应用及样本容量等几个问题是经济时间序列建模过程中需要重点关注的问题。文章在讨论上述问题的基础上设计出经济时间序列的建模流程。

关键词：经济时间序列；ARIMA；建模路经；季节效应；检验方法

中图分类号：O212　　**文献标识码**：A　　**文章编号**：1002-6487(2014)08-0029-03

0　引言

时间序列即按照时间的顺序所记录的事件发展变化的状态。在自然科学领域、社会科学领域乃至人们的日常生活都存在着对时间序列的观察和研究。在经济学及统计学科中，对于时间序列的建模一直是学者们的一个重点关注领域，已经成为计量经济学研究的一个重要的分支方向。从建模的角度看，时间序列分析方法包括两类，一类是频域分析方法，一类是时域分析方法。频域分析方法需要用到高深的数学工具，较为复杂，而且分析结果抽象，主要在物理学、气象学等自然科学中广泛应用，而在实际的经济问题研究中具有一定的使用局限性；时域分析方法依靠时间序列值之间的相关关系及其统计规律，并以此拟合出适当的数学模型（即时间序列模型）来加以描述，对于经济学研究而言，这种方法的可靠性较强，分析结果易于解释，是经济时间序列分析的主流方法，本文就是在频域分析框架下讨论时间序列的建模方法。通过时域分析方法所构建的时间序列模型在经济学领域已有广泛的应用，例如自回归移动平均（ARMA，ARIMA）模型、自回归条件异方差（ARCH）模型等。本文将重点以ARIMA（一般视ARMA模型是ARIMA模型的特例）类模型的构建过程为例，研究时间序列建模思路、建模依据、建模特点等问题，探讨经济时间序列的建模方法。显然，这里讨论的对象是单个经济时间序列变量，不涉及多元经济时间序列变量。

1　时间序列的ARIMA类模型构建过程

时间序列的建模属于典型的数据主导型建模方法，即主要根据样本数据所提供的信息来完成模型的构建。对于一组时间序列数据，主要建模步骤包括数据的平稳性及纯随机性检验、选择模型并确定其阶数、估计模型参数和模型检验几个部分。

1.1　平稳性及纯随机性检验

平稳时间序列和非平稳时间序列的建模方法是不同的，对于非平稳时间序列而言，主要存在两类：其一是具有时间趋势的时间序列，其二是随机游走序列，多数经济时间序列数据都属于前者。对于含有时间趋势的时间序列数据而言，一般有两种建模方式，第一种是构建趋势模型，即以时间为解释变量构建回归模型，在预测中的趋势外推预测法就是通过构建此类趋势模型完成的。第二种是构建自回归模型，当然这种自回归模型是非平稳的。在必要的情况下也可以构建含有时间趋势的自回归模型。对于随机游走序列的建模方式较为简单，只需构建该序列的一阶自回归模型即可，而且此时的自回归系数为1。

如果所面临的数据是平稳的时间序列，还需要进一步的进行纯随机性检验。因为对于一般的平稳时间序列来说，是可以建模的，但有一类平稳时间序列是不能直接建模的，那就是纯随机序列。纯随机序列不存在相关性，对于时域分析而言再没有可利用的信息，因而不能建模。换句话说，只要平稳时间序列存在自相关，就可以构建模型。

1.2　模型阶数识别

如果检验发现某一经济时间序列是平稳且非纯随机序列，可以考虑构建ARMA类模型。若要构建ARMA类模型，首先要识别自回归项和移动平均项的阶数，即模型的“定阶”。借助于计算机技术完成这一步并不困难，只需观察样本时间序列的自相关图和偏自相关图即可，其原理是针对于不同的模型（AR模型、MA模型及ARMA模型），其自相关系数和偏自相关系数有不同的变动特征，自相关图和偏自相关图也有各自的规律性，根据这些特征和规律，就可以对符合的模型形式及阶数加以选择和判断。还有一些其他的定阶方法，如F检验法、准则函数法等，这里不

基金项目：甘肃省高校人文社科重点研究基地甘肃经济发展数量分析研究中心资助

作者简介：刘　明（1981-），男，安徽霍邱人，博士研究生，讲师，研究方向：统计理论与方法、经济计量分析。

政补贴的支持和带动作用无疑是显见的。从这个意义上讲，财政支持必将成为农村金融生态环境优化的重要因素和力量。

㈡ 创新金融监管。一方面，通过监管优化涉农金融机构的法人治理结构，以此构建金融机构自我管理的良好基础，强化自律性。同时通过监管实现农村金融供给体系的有序、合理竞争局面，为金融机构建立外部市场约束机制。另一方面，变强制性监管为激励性监管，通过利益补贴的方式促使农村金融机构走集约型增长道路，以此克服监管领域中普遍存在的因信息不对称而产生的道德风险与逆向选择。

㈢ 整合现有农村组织资源。农村金融改革中，金融深化与农民组织化要同步进行，为了降低组织成本，提高组织效率，需要对农村现有组织资源进行必要整合，为农村金融的改革与发展提供一个广阔平台，建立一个以金融为核心的农村合作体系。

㈣ 大力开展农村金融教育。金融需求有效性的提高都将有赖于金融需求者的认知能力。金融教育、金融文化的普及是降低农村金融市场交易成本的重要因素，但目前为止仍没有得到重视。农村金融教育与农业技术培训同样重要，要以农村金融主力军——农村信用社众多的经营网点为基地，以产品、服务为载体，向农村各类经济活动主体输送金融基本理念与金融信息，通过开展各种形式的喜闻乐见的活动使金融成为农村居民生活的一部分。为充实该项职能，农村信用社软硬件条件均需得到改善，同时对于此类具有公共产品性质的服务，国家应给予农村信用社一定的补贴。

参考文献：

[1] 约瑟夫·斯蒂格利茨(Joseph E.Stiglitz).经济学[D].中国人民大学

[2] 韩俊.我国农村金融需求问题[J].中国农村信用合作，2008(1)

[3] 袁秀峰.农村信用社信贷市场定位再思考[J].金融与经济，2009 (1)

[4] 熊雪萍.农户金融行为与融资需求的实证分析[J].农业技术经济，2007 θ

甘肃省城乡收入差距库兹涅茨效应分析

◎兰州商学院 王永瑜 刘会勇

导 言

随着我国人均国内生产总值超过 3000 美元，收入分配问题越来越受到社会各界和经济学家的重视。主要源于两个方面原因：

首先是基于对我国经济实践的观察和思考。中国改革开放 30 多年来，经济增长举世瞩目。统计表明，在 1981 年—2009 年的 29 年中，有 24 年国内生产总值增长率超过 8%，平均增长率为 8.9%。据国家统计局初步测算，2010 年，我国国内生产总值达 39.8 万亿元，已经超过日本，居世界第二。但是，在国民经济总体高速增长的同时，由于中国改革开放中的体制性因素和经济发展的特点，城乡收入差距问题依然比较严重，并且这一问题已经成为制约我国经济与社会和谐发展的重要因素。

其次是基于对发达国家发展经验的理论思考。居民收入差距扩大是世界各国在工业化过程中普遍出现的阶段性问题。最早对其进行系统研究的是美国著名经济学家、1971 年诺贝尔经济学奖得主西蒙·库兹涅茨。他在 1955 年美国经济协会的演讲中，经过对 18 个国家经济增长与收入差距实证资料分析，得出收入分配的长期变动轨迹是“先恶化，后改进”，或用他自己的话说是“收入分配不平等的长期趋势可以假设为：在前工业文明向工业文明过渡的经济增长早期阶段迅速扩大，尔后是短暂的稳定，然后在增长的后期阶段逐渐缩小”，并且他通过比较一些国家的横截面资料，得出的结论是处于发展早期阶段的发展中国家比处于发展后期阶段的发达国家有更高的收入不平等，表现在图形上是一条先向上弯曲后向下弯曲的曲线，形似颠倒过来的 U 字，故人们将其称为“倒 U 曲

【统计理论与方法】

收入不平等变迁特征的探索性分析
——基于洛伦兹曲线的动态分解

黄恒君[a,b]

（兰州商学院 a. 统计学院；b. 甘肃省经济发展数量分析研究中心，甘肃 兰州 730020）

摘要：提出基于函数序列的收入不平等动态测度思路与方法：采用 B－样条拟合洛伦兹曲线序列；在生成函数型数据的基础上，对洛伦兹曲线序列进行函数型主成分分析。利用函数型数据建模，对 1990－2010 年中国城镇居民收入洛伦兹曲线序列变迁特征进行探索性数据分析，结果表明：采用人口五分法划分收入群体具有合理性；各收入群体对历年收入不平等程度变迁的贡献率分别为：低收入群体 2.30%，中等收入群体 80.33%，高收入群体 17.36%。

关键词：函数型主成分；洛伦兹曲线；收入差距

中图分类号：O212.1 文献标志码：A 文章编号：1007－3116(2012)10－0025－05

一、引 言

居民收入（以下简称收入）问题的研究历来是国内外经济学界关注的焦点。近 20 年来，收入差距在一定程度上迅速扩大，其引发的诸多经济社会问题倍受瞩目，相关研究广泛而深入。拟合函数（如收入分布、洛伦兹曲线）估算并分解不平等指标，是国内关于收入不平等测度研究的重要思路。具体研究方法主要包括：国内关于收入不平等变化的研究大多是遵循 Shorrocks 的理论[1]，按人口特征或要素分组等外生变量来分解收入差距，并以此来说明各分解因素的影响效果。例如，万广华等借助回归分析和 Shapley 分解系数和泰尔指数，将农民收入差距归因于各要素投入，认为资本投入已成为影响中国农村收入差距的最重要因素[2]；唐莉等通过对基尼系数的分解研究了城镇居民收入差距，认为中国城镇居民收入差距主要归因于地区差异[3]；庄健、张永光采用多项式函数形式拟合洛伦兹曲线，按照 20%、60%、20%收入群体比率，将基尼系数分解到各收入群体中[4]；王亚芬等在事先给定收入群体划分的基础上，通过计量经济模型，检验了低、中、高收入群体可支配收入与基尼系数之间的关系[5]。

上述关于收入不平等的研究，主要还是基于不平等指标（标量）的分解。事实上，收入差距变动的最新研究强调对整个收入分布进行分解，已有学者重视基于收入函数的研究[6]；迟巍等基于收入分布，通过分位数回归方法，将收入差距扩大的原因分解为劳动者的劳动力特点的变化以及对劳动力特点回报率的变化[7]；Di Nardo 等利用可加半参数模型对工资收入分布的影响因素进行分解[8]；Oaxaca 将线性变换引入收入差距扩大成因的研究[9]；Jenkins、Van Kerm，陈娟、孙敬水基于该线性变换，对不同年份的收入分布函数差异进行了位移和尺度分解[10－11]。然而，基于收入函数本身的动态变化研究并不多见。

本文对函数型数据分析在收入不平等测度中的应用作出尝试。为了研究近 20 年来收入差距变动特征，本文在拟合 1990－2010 年中国城镇居民洛伦兹曲线的基础上，将曲线视为样本，对历年收入差距变动规律进行探索性分析，按照收入差距变动特征提取并划分低、中、高收入群体，并测算各收入群体对收入不平等变动的贡献率。

收稿日期：2012－05－19

基金项目：国家社会科学基金资助项目《中国现行社会福利保障制度下城镇贫困人口的统计研究》(11BTJ002)

作者简介：黄恒君，男，浙江温州人，博士生，讲师，研究方向：调查技术与统计分析。

中国经济增长影响因素的变迁：2004～2011

刘　明
（兰州商学院，兰州　730020）

〔**摘　要**〕　在综合供给与需求角度下的经济增长的相关研究结论基础上，本文通过构建中国经济增长的面板数据模型研究发现，外来投资对中国经济增长的影响较弱，内部投资对经济增长贡献远高于外来投资；消费因素受到金融危机的影响较小，它对经济增长的拉动作用并没有降低，而出口因素受金融危机的影响较大；东、中部的劳动力处于富足状态，而西部经济需要更多的劳动力。面板数据模型的实证分析结果表明，金融危机对中国经济增长的影响是显著的，它改变了各影响因素对经济增长的促进或拉动作用，这种改变的幅度又随区域的不同而有所差异；中国经济增长结构不均衡，有进一步优化的空间；开拓以西部为主国内市场，不仅能使中国经济获得进一步增长的动力，更能为中国经济发展提供战略层面的安全保障。

〔**关键词**〕　金融危机　经济增长　增长结构　影响因素　面板数据模型
DOI：10.3969/j.issn.1004－910X.2013.11.019
〔**中图分类号**〕F120.3　〔**文献标识码**〕A

引　言

改革开放以后，中国经济经历了持续的高速增长，经济总量已跃居世界第二。从需求角度看，中国的经济增长主要依靠的是投资的增长和出口的扩张，这种增长模式可概括为“凯恩斯＋重商主义增长型”。从供给的角度看，中国改革开放后30多年经济增长的动力源于要素投入增加、技术引进和经济体制改革所带来的效率的提高。加入WTO后，中国实际上已经成为完全的市场经济国家，并进一步在全球经济增长中发挥着引领性作用。在经历了2003～2007年的快速增长之后，包括中国在内的各国经济因由美国次贷危机所引起的全球金融危机而受到重创，全球经济陷入衰退。中国政府在危机出现之后推出了一系列宏观刺激计划，包括著名的“四万亿投资”计划，这在很大程度上缓冲了欧美经济波动对中国经济产生的震动，促使中国率先摆脱金融危机的影响。但是在此之后，中国经济并未像所期待的那样强劲复苏，而是出现了反复，中国经济的高速增长态势也面临着考验。事实上，这主要是全球经济环境出现一些不良因素，使得全球经济形势突然变得扑朔迷离，中国经济外部环境的不确定性急剧上升。这些因素主要表现在如下几个方面。

一些经济大国采取了不负责任的宏观经济政策。当前，作为世界上最大的经济体和领头羊，美国的宏观经济政策主要包括：急速的扩张性财政支出与价格性货币政策的组合，核心政策措施包括执行7000亿美元的救市计划和大幅度的降低利率；结构性的减税与数量性货币政策的组合，主要措施包括布什政府和奥巴马政府的两轮减税和第一轮量化宽松货币政策；战略性财政支出扩张和数量性货币政策的组合，主要措施包括以基础设施和新能源作为投资重点和第二轮量化宽松货币政策。显然，从全球视角来看，这些政策的实施导致的最直接的后果是美元贬值，他国货币相对升值，进而改变国际贸易格局，造成了“美国亏本，全球买单”的不合理的经济现状。

欧债危机再次使全球经济陷入困境。继美国次贷危机之后，欧洲主权债务危机使本已陷入困境的全球经济雪上加霜。随着债务危机不断恶化，

收稿日期：2013－10－04
作者简介：刘明，兰州商学院统计学院讲师，兰州商学院甘肃经济发展数量分析研究中心兼职研究员，经济学硕士，中国人民大学统计学院博士生。研究方向：统计理论与方法、经济计量分析。

DOI:10.15885/j.cnki.cn11-4683/z.2011.03.015

中国农村居民消费状况分析
——基于西方消费理论的实证检验

刘 明
(兰州商学院统计学院,甘肃兰州 730020)

摘 要:该文根据近代西方经济学理论的几类消费理论,在近年来宏观经济数据的基础上构建能够反映出中国农村消费状况的经济模型,寻找影响中国农村居民消费的主要因素,分析中国农村居民的消费状况,探求解决促进农村居民消费需求增长问题的途径。

关键词: 消费理论; 农村居民消费; 收入
中图分类号:F036.3 文献标识码:A 文章编号:1671-2404(2011)44-0041-05

中共十七届五中全会提出,要坚持扩大内需战略、保持经济平稳较快发展,建立扩大消费需求的长效机制。扩大内需的重要内容之一就是启动和激活人口占70% 的农村市场。在宏观经济政策方面,中国政府在处理总需求问题上出现了三个转向: 一是由偏重外需向偏重内需转变,二是由偏重投资需求向偏重消费需求转变,三是由偏重城镇消费向偏重农民消费转变。这表明了政府对农村消费市场的重视,体现了农村消费增长对于国民经济的重要性。当前和今后一段时期扩大内需的重点在农村,难点也在农村,如何促进农村居民的消费需求增长已成为中国经济发展亟待解决的问题之一。

1 西方消费理论简述

西方消费理论按照发展的历史呈递关系依次主要存在这样几类假说: 绝对收入假说、相对收入假说、生命周期假说、持久收入假说和随机游走假说,它们构成了现代消费理论的核心。同时,这些也是本文研究的理论基础。

1.1 凯恩斯的绝对收入假说

绝对收入假说(*Absolute Income Hypothesis*) 也称为绝对收入理论,是凯恩斯在1936年出版的《就业、利息和货币通论》中提出的。该假说认为短期消费支出是实际收入的稳定函数,即消费取决于收入,消费与收入之间的关系可以用消费倾向来描述。随着收入的增加消费也将增加,但消费的增长低于收入的增长,消费增量在收入增量中所占的比重是递减的,即边际消费倾向递减。

1.2 杜森贝里的相对收入假说

相对收入假说(*Relative Income Hypothesis*) 是由美国经济学家*J*·杜森贝里(1949) 在《收入、储蓄和消费者行为理论》中提出来的。该假说的核心内容是,当期收入和过去的消费支出水平决定当期消费。杜森贝里认为存在两种效应影响消费水平: 一是示范效应,家庭消费决策主要参考其他同等收入水平家庭,即消费有模仿和攀比性; 二是棘轮效应,家庭消费即受本期绝对收入的影响,更受以前消费水平的影响,消费者易随收入的增加而增加消费,但不易随收入的减少而减少消费,以致产生有正截距的短期消费函数。按他的看法,消费与所得在长期维持固定比率,故长期消费函数是从原点出发的直线,但短期消费函数则为有正截距的曲线。

1.3 莫迪里安尼的生命周期假说

生命周期假说(*Life Cycle Hypothesis*) 是由美国经济学家*F*·莫迪利安尼和*R*·布伦贝格、*A*·安东共同提出来的。该假说的前提是: 首先假定消费者是理性的,能以合理的方式使用自己的收入,进行消费; 其次,消费者行为的唯一目标是实现效用最大化。这样,理性的消费者将根据效用最大化的原则使用一生的收入,安排一生的消费与储蓄,使一生中的收入等于消费。该理论认为,每个家庭都是根据一生的全部预期收入来安排自己的消费支出的,即每个家庭在每一时点上的消费和储蓄决策都反映了该家庭希望在其生命周期各个阶段达到消费的理想分布,以

收稿日期: 2010-11-29

作者简介: 刘明,兰州商学院讲师,经济学硕士,主要从事经济计量分析等方面的研究。E-mail: liumingpzh@163.com

《经济问题探索》2016 年第 4 期

中国制造业空间分布的异质性*

——基于 GWR 与分位数回归的分析

刘　明

(兰州财经大学，兰州 730020)

摘　要：Moran's I 指数和 LISA 指数的分析表明，中国制造业各主要指标在省域层面上都呈现出空间相关性，且相关密切程度和相关范围在逐步扩大；制造业空间地理加权生产函数表明，资本对制造业产出影响程度较大的地区主要集中在华北、东北，资本对产出影响程度较小的地区主要集中在华东和华南，劳动要素对制造业产出影响程度较大的区域集中在华东、华南和西南等人口密度普遍较大的地区；制造业分位数回归生产函数表明，在制造业发展水平较低的区域，资本要素对产出的影响远高于劳动要素而起着决定性作用，随着发展水平的逐步提高，资本要素对产出的影响在逐渐的减小而劳动要素对产出的影响在逐渐加大，在制造业发展水平较高的地区，劳动要素对制造业产出的影响反而高于资本要素。研究还发现，中国制造业在中等发达区域的规模经济效应明显，因此有向中部地区转移的内在要求。

关键词：制造业；空间异质性；空间相关；GWR；分位数回归

制造业是中国经济发展的基础性产业，是实体经济的核心。由经济发展的历史来看，中国若实现由“经济大国”向“经济强国”的迈进，必须首先实现“制造业大国”向“制造业强国”的迈进。影响中国制造业发展的因素有很多，区域发展不协调是其中之一，区域发展的不均衡性使中国制造业在各省域的发展状况、发展模式均有较大差异，而且在不同的区域出现出块状发展的特征，例如长三角区域、西北区域等，这些区域的内部制造业发展有一定的相似性，而在区域间又存在很大的差异性。因此，有必要从空间统计学的角度对这类差异性进行分析研究，即对中国制造业进行空间异质性研究。这种研究主要从制造业相关指标数据入手，根据数据信息对空间异质性进行分析和研究，探索中国制造业的空间异质性特征。

空间异质性是指空间中各样本数据由于所处的空间或地理位置的不同而表现出来的差异性。对于空间异质性的分析角度有很多，常见的有这样的两个角度，一是单个变量的依赖关系，二是多个变量的依赖关系。单个变量依赖关系的角度通常是从空间相关性研究从发，分析相关性在不同区域内的变化，以此展示空间异质性，这类研究通常是在全局相关性研究的基础上，进一步借助于局部相关性指数（Local Indicators of Spatial Association，LISA）而展开。空间全局相关性的探测与检验方法一般依赖于Moran's I 统计量，这是探测和检验变量空间相关性的最常用的统计量。多个变量的依赖关系的角度通常是由多个变量构建回归模型，通过分析模型在不同区域内的变化来分析展示空间异质性，这种方法常借助于地理加权回归模型（Geographically Weighted Regression，GWR）。因此，本文将在对中国制造业进行空间相关性分析的基础上进一步展开空间异质性的分析。涉及的统计分析方法主要有空间相关系数、局部空间相关系数、地理加权回归模型以及分位数回归模型①。数据方面，本文使用了中国大陆地区制造业的 31 个省、市、区的 2008 年至 2013 年的数据，涉及的指标包括总产值、资产总量、从业人数、利润总额、单位劳动产

作者简介：刘明（1981 -），男，兰州财经大学统计学院副教授，兰州财经大学甘肃经济发展数量分析研究中心副主任，经济学博士，研究方向：统计理论与方法、经济计量分析、空间数据建模。

*　基金项目：国家社科基金重大项目“经济社会公共数据的空间统计样本数据开发及应用研究”（11&ZD157）；国家统计局统计科学研究重点项目“空间数据建模技术及其在我国居民消费分析中的应用”（2013LZ11）；甘肃省人文社科重点研究基地“甘肃经济发展数量分析研究中心”资助。

① 为节约篇幅，本文不再对这些统计方法加以阐述，读者可参见相关统计类或计量经济学类教材及论著。

DOI:10.13546/j.cnki.tjyjc.2012.10.013

Logistic模型预测的新思路

刘 明[1,2]

（1.兰州商学院 统计学院，兰州 730020;2.甘肃经济发展数量分析研究中心，兰州 730020）

摘 要：预测是Logistic模型的一个重要功能，文章在研究运用对数回归模型进行预测的问题基础上，进一步对Logistic模型在预测过程的由于模型随机项的非正态性、异方差性而引起的问题进行分析研究，研究发现，传统的预测方法会使得预测结果的精度不高，在考虑到模型随机项的存在及其具体特征后，通过分析推导，构造出新的预测思路和方法。

关键词：对数回归模型；Logistic模型；预测

中图分类号：O212；F224.0 **文献标识码**：A **文章编号**：1002-6487(2012)10-0082-02

0 问题的提出

Logistic模型是一种用于分析定性变量的回归模型，尤其是二值的被解释变量，可以计算和预测其属于某一取值（类别）的概率（当然，和其他定性数据分析方法一样，Logistic模型也可以用来分析数值变量）。Logistic模型在应用过程中需要对被解释变量进行Logistic变换，这一过程需要将数据取自然底数对数，因此它属于对数回归模型。Logistic模型的构造和参数估计都依托于传统的回归分析方法，在对具体问题进行分析和预测时，也和传统的回归分析相似，因此在经济学、管理学中具有广泛的应用。笔者从事统计理论和计量经济学等方面的教学和科研工作，在工作中发现，人们利用Logistic模型进行预测的传统方法存在着不够严谨的思维方式，以至于预测结果的精度不高。究其原因，是对Logistic模型的本质没有准确的把握，没有弄清模型的结构和预测的原理。首先，Logistic模型是一种对数回归模型，被解释变量和模型随机项之间不是线性关系；其次，Logistic模型在模型设计过程中形成的随机项可能不服从传统回归模型的经典假定。因此不能直接运用传统回归模型的预测思路来完成Logistic模型的预测工作。针对上述事实，笔者拟从对数回归模型的预测问题入手，指出人们在预测过程中的缺陷所在，在对数回归模型的预测思路和预测方法的基础上，进一步探讨Logistic模型的预测问题。

1 对数回归模型的预测问题

传统的线性回归模型的预测思路是，根据所估计的样本回归模型，由预测点上解释变量的取值，计算出被解释变量的均值的估计值，以该估计值作为被解释变量的预测值。它实际上是取被解释变量关于解释变量的条件均值。令含有 k 个解释变量的总体回归模型和样本回归模型为：

$$y_i=\beta_0+\beta_1x_{1i}+\ldots+\beta_kx_{ki}+\mu_i \quad (1)$$

$$y_i=\hat{\beta}_0+\hat{\beta}_1x_{1i}+\ldots+\hat{\beta}_kx_{ki}+\hat{\mu}_i \quad (2)$$

μ 是服从经典假定的随机项。假设在预测点上解释变量的取值依次为 x_{1f}、x_{2f}、…、x_{kf}，则被解释变量 y 在该预测点上的预测值即为：

$$\hat{y}_f=\hat{\beta}_0+\hat{\beta}_1x_{1f}+\ldots+\hat{\beta}_kx_{kf} \quad (3)$$

传统线性回归模型的预测值的计算依据是其总体回归模型，因为在预测点上的总体回归模型为：

$$y_f=\beta_0+\beta_1x_{1f}+\ldots+\beta_kx_{kf}+\mu_f$$

随机项 μ 的均值等于零，因此可得 y_f 关于解释变量的条件均值为：

$E(y_f|x_{1f},\ldots,x_{kf})=\beta_0+\beta_1x_{1f}+\ldots+\beta_kx_{kf}$ 再根据模型参数估计结果即得出预测值计算式(3)。显然，预测过程中考虑了被解释变量 y 与随机项 μ 的线性关系，而 μ 的均值为零，因此随机项对预测结果的影响不作考虑。

现在考察如下形式的对数回归模型：

$$\log(h_i)=a_0+a_1x_{1i}+\ldots+a_kx_{ki}+\eta_i \quad (4)$$

$$\log(h_i)=\hat{a}_0+\hat{a}_1x_{1i}+\ldots+\hat{a}_kx_{ki}+\hat{\eta}_i \quad (5)$$

式(4)为总体模型，式(5)为样本模型，η 是服从经典假定的随机项，$\log(\cdot)$ 是自然底数对数函数。设预测点上解释变量的取值依次为 x_{1f}、x_{2f}、…、x_{kf}，依照传统回归模型的预测思路，将预测点上解释变量的取值代入式(5)：

$$\log(\hat{h}_f)=\hat{a}_0+\hat{a}_1x_{1f}+\ldots+\hat{a}_kx_{kf}$$

将两边取指数函数值即可得到 y 的预测值：

$$\hat{h}_f=e^{\hat{a}_0+\hat{a}_1x_{1f}+\ldots+\hat{a}_kx_{kf}}$$

$$=\exp\{\hat{a}_0+\hat{a}_1x_{1f}+\ldots+\hat{a}_kx_{kf}\} \quad (6)$$

这即根据传统线性回归模型的预测思路——不考虑随机项的影响而得出的对数回归模型的预测结果。

作者简介：刘 明（1981-），男，安徽霍邱人，硕士，讲师，研究方向：统计理论与方法、经济计量分析。

单位名录库地理信息建立途径与“互联网 +”维护管理的探讨*

陶 然　黄恒君

内容提要:“互联网 +”模式下,基于地理信息的统计数据开发利用对提高政府统计能力带来的影响不可忽视。结合美国 MAF/TIGER 系统建立的实践经验,通过分析名录库单位地址与电子地图中建筑物的匹配关系,本文提出了利用经济普查获取的单位位置坐标,通过 5 年一次的全面更新为名录库建立地理识别信息,探讨未来在非普查年份如何获取位置坐标以用于单位名录库地理信息的维护管理,重点讨论了通过互联网资源采集信息点(POI)数据和反向地址编译获取位置坐标的技术手段。统计系统基本单位名录地理信息的建立与开发将进一步拓展我国基本单位名录库的应用领域。

关键词:单位名录库;地理信息;互联网;位置坐标
中图分类号:C812　　文献标识码:A　　文章编号:1002 - 4565(2016)02 - 0010 - 08

Geographic Information of Business Register: Establishing and Updating Approach in the Internet Plus Mode

Tao Ran & Huang Hengjun

Abstract: In the Internet Plus mode, the impacts of statistical data exploitation on improving the ability of government statistics based on geographic information cannot be ignored. Combined with practical experience of establishing MAF / TIGER system in the United States, and matching relationship analysis between unit address of business register and the building in digital map, we put forward to establish corresponding geographic identification information with register, discuss how to obtain the position coordinates resources in a non-census year for the maintenance and management of geographic information in the future, focus on the collection of information through the point of interest data and reverse address compiler. Establishment and development of geographic information will further expand the application field of business register.

Key words: Business Register; Geographic Information; Internet; Position Coordinates

一、引言

统计地理信息系统(统计 GIS)是基于电子地图实现各类调查对象和统计信息集成、定位、展现、汇总、分析、服务的综合型信息系统,是利用现代科学技术手段展示统计信息的有效平台①。作为社会经济现象的综合反映,统计数据普遍具有空间属性。大数据时代,传统的统计数据图表和统计分析方法不能有效地展现海量统计数据的空间特征,而通过统计地理信息系统将统计数据与地理信息整合,能够充分挖掘和展示统计数据所隐含的空间分布特点和规律。

由政府统计部门主导的周期性普查是定期搜集全面反映社会经济现象的统计数据的方式,在世界各国周期性普查实践中,地理信息技术作为有效的数据采集、分析与展示工具,通过不断与普查数据生

* 本文是国家自然科学基金资助项目“基于涵盖误差的我国周期性普查数据质量评估方法:理论与应用研究”(71301033)、国家社科基金资助项目“基于大数据整合的空气质量测度方法研究”(14CTJ009)、全国统计科研计划项目“基于普查涵盖误差测量技术的基本单位名录库维护与更新研究”(2011LX003)的阶段性成果。

① 徐一帆在 2012 年全国普查中心系统工作会议上的讲话,2012 年 1 月 6 日。

【统计理论与方法】

线性回归模型设定的两个常见错误分析

刘 明
(兰州商学院 统计学院, 甘肃 兰州 730020)

摘要:删除截距项和遗漏解释变量是线性回归模型估计中的两个常见错误,删除截距项错误发生的原因是检验过程中发现其不显著而将其剔除,这会造成模型参数估计和假设检验的失真;遗漏解释变量的错误发生原因是人们错误认为只要变量存在相关性且存在因果联系就可以进行回归分析,以至于不考虑其它重要的解释变量,此时建立的模型不能用于经济结构分析和政策评价,最多只能用于预测目的。

关键词:设定错误;截距项;解释变量
中图分类号:0212 文献标志码:A 文章编号:1007-3116(2011)08-0011-04

一、问题的提出

线性回归模型是最基本的计量经济学模型,也是研究经济变量关系最常用的模型,它是经典计量经济学的主体内容,经典计量经济学就是围绕线性回归模型的设定、估计、检验和应用展开的。线性回归模型的参数估计、假设检验(包括统计检验和计量经济学检验)有着一套较为完备的统计学方法体系,只要对这一体系有所把握,在实际应用中就不会出现失误。根据研究对象和研究目的来构造回归模型并加以应用,事实证明,模型是否得到正确的应用往往取决于是否构造了一个良好的、正确的模型。对于线性回归模型的设定,是最容易出现错误、而且是最难以发现错误的环节,模型设定的正确与否,直接关系到建模的成败,一个设定错误,会使整个研究过程和研究结论都变得毫无意义和价值,因此需要仔细的斟酌研究。线性回归模型设定的常见错误主要包括:增加错误的解释变量,错误的模型数学关系式,删除截距项,遗漏重要的解释变量。增加了错误的解释变量可以通过 t 检验、F 检验来诊断并发现该错误的变量,对于变量间数学关系式的选择,可以通过研究分析变量间的散点图来发现它们之间的关系,并通过拟合误差(例如均方误差、平均绝对误差等)来比较分析,选择正确的数学关系式的模型。本文的主要任务就是研究讨论删除截距项、遗漏重要解释变量这两类设定错误,分析它们的发生原因及后果,提出相应的解决办法,并讨论例外的情形。

二、删除截距项的线性回归模型

构建线性回归模型时可能出现这类情况:截距项的显著性 t 检验未能通过,即截距项的 t 检验结果支持其等于零的假定。此时截距项因为不显著而可能被删除。线性回归模型中的截距项通常不做经济意义解释,但这并不意味着截距项可有可无。如果在总体回归模型中应该包含截距项,删除后则会产生一系列不良后果。为简单起见,这里以一元线性回归模型为例对删除截距项所产生的后果进行讨论。

设正确的总体线性回归模型为:

$$y_i = \beta_0 + \beta_1 x_i + \mu_i \qquad (*)$$

μ_i 为随机干扰项,其满足高斯假定,设其方差为 σ^2。

未删除截距项的样本回归模型表示为:

$$y_i = \hat{\beta}_0 + \hat{\beta}_1 x_i + \hat{\mu}_i$$

删除截距项的样本回归模型表示为:

$$y_i = \tilde{\beta}_1 x_i + e_i$$

e_i 为样本模型的残差。设样本容量为 n,则运用普通最小二乘法可得模型参数估计量:

收稿日期:2010-12-05
作者简介:刘明,男,安徽霍邱人,讲师,经济学硕士,研究方向:经济计量分析。

经济学研究中的线性回归模型设定的再思考

刘　明[a,b]

（兰州商学院 a.甘肃经济发展数量分析研究中心；b.统计学院，兰州 730020）

摘　要：文章在作者原有研究的基础上提出了应用于经济学研究中线性回归模型设定的三个一般性原则，并提出了理论驱动型和数据驱动型两类线性回归模型设定方法，总结出了单方程线性回归模型设定的基本过程。

关键词：线性回归模型；模型设定；理论驱动型设定法；数据驱动型设定法

中图分类号：O212　　**文献标识码**：A　　**文章编号**：1002-6487(2015)07-0008-03

0　引言

线性回归模型的设定是回归分析方法和实际社会经济问题研究的交合点。很多人能够熟练地对线性回归模型进行参数估计、检验，也深谙针对某特定问题建立的线性回归模型所表述的社会经济意义，但是却因不能熟练地把握模型设定技巧，致使模型对实际问题没有足够的解释能力，甚至发生谬误。客观地说，把握线性回归模型的设定有一定的难度，他不仅要求对线性回归模型数理方法非常熟练，更要求对所研究的问题有深刻全面的认识。因而可以认为，线性回归模型设定问题既有科学性，又不乏艺术性，该领域的研究是科学性和艺术性的统一，即一方面要根据客观实际，运用严谨的数学思维设定模型的函数形式，另一方面，模型设定效果的优劣也依赖于研究者对客观问题的认知与把握程度。就方法而言，线性回归模型属于自然科学范畴，我们用线性回归模型来研究经济学问题，不可能完全按照自然科学的研究模式来进行，因为经济学是一门社会科学，虽然自然科学研究方法和研究模式为社会科学研究提供了重要的借鉴，但不能完全替代。线性回归模型设定问题研究不是一种完全的自然科学研究模式，因为这需要结合实际的经济背景才能完成，因而又涉及社会科学领域。正因为如此，模型设定问题研究在强调科学性的同时，又体现了一定的艺术性。因而，本文的思路是，以科学性为指导，构造线性回归模型设定方法与路径，使用辩证的方式对模型设定问题进行研究。本文将以经济学应用为背景，探讨和分析线性回归模型设定过程中具体方法，以实现实证分析中线性回归模型的正确设定。为了使研究更具针对性，本文所讨论的线性回归模型均指单方程形式的线性回归模型。

1　线性回归模型设定基本方法

对于线性回归模型的设定的一般性方法，笔者曾进行了一些探讨[1]。一般而言，回归模型的设定不能仅依据相关关系和因果关系，更要充分考虑影响研究对象的一般性因素，做到不遗漏解释变量[1]。对于一般的计量经济学回归模型的设定，其基本思路是，首先要依赖于经济理论和所研究问题的经济背景，理清经济系统中的变量，理顺这些变量的依赖关系和从属关系；接着依据理论并结合实际数据，确定变量间的数学关系模式，以保证所设定的模型形式符合实际经济关系；最后，利用统计检验中的设定检验方法（例如RESET检验）对完成参数估计之后的回归模型进行检验，以确定原模型设定是否存在偏误。在实际研究中，当面临经济学问题时，一般目标是设定一个能正确反映经济关系的、科学的、符合实际研究需要的计量经济学回归模型，但计量经济学回归模型之外的其他统计学回归模型对解决经济学问题的作用也不容忽视，例如用于预测目的的趋势外推模型等；当面对的是一些特殊的社会经济现象或出于一些特定的研究目的时，非经济学性质的一般统计回归模型也有着广泛的应用，例如身高与体重关系的回归模型等。因此，在设定回归模型尤其是计量经济学回归模型时，在遵循一般性的设定思路的同时，也要考虑到一些特殊情况的处理。在线性回归模型设定过程中，尤其是以经济学为应用背景的回归模型的设定，其各类统计检验方法在模型设定过程中能及时反馈模型设定效果，直接或间接地对模型设定的正确与否进行检验。模型的统计检验结果对于诊断模型设定偏误问题非常必要，甚至至关重要，这些检验结论可以帮助研究者重新审视所设定的模型，为模型的正确设定提供有力帮助。当运用统计检验发现模型存在一些问题时，应尽可能地避免使用一些技术性手段对参数估计方法及估计结果进行修正，而应首先考虑探查模型的设定偏误问题。

笔者综合已有的研究结论和实际研究经验[1]，提出线性回归模型设定的三个一般性原则，在实际研究中遵循这些原则，可以使模型设定工作有的放矢，找到研究的起始

基金项目：国家统计局统计科学研究计划重点项目（2013LZ11）

作者简介：刘　明(1981-)，男，安徽霍邱人，硕士，讲师，研究方向：统计理论与方法、经济计量分析。

一类计量经济学模型设定偏误诊断思路及展示

刘　明[1],黄彦彦[1,2]
(1.兰州商学院 甘肃经济发展数量分析研究中心,兰州 730020;2.中国社会科学院 研究生院,北京 102488)

摘　要:在分析计量经济学模型的几类基础统计检验功能的基础上,文章讨论并研究了这些检验方法用于诊断计量经济学模型设定偏误的机理及可行性,并进一步结合实际经济问题的研究过程,展示在模型构建过程中统计检验方法关于模型设定偏误诊断的作用与功能。理论分析和实证研究结果均显示,这些统计检验方法可以有效用于计量经济学模型设定偏误的诊断。

关键词:计量经济学模型;统计检验;模型设定偏误诊断
中图分类号:O212;F222　　**文献标识码**:A　　**文章编号**:1002-6487(2015)04-0004-05

1　研究背景与文献回顾

回归分析方法是最为重要的统计分析方法之一,构建线性回归模型是回归分析中最基础也是最核心的内容。随着学科领域间的交互融合与发展,构建计量经济学回归模型也是计量经济学学科的最为基础的内容。在实际应用中建立计量经济学回归模型的第一步就是模型设定,即如何设定出一个正确的计量经济学回归模型,这是能否将模型成功应用、能否分析和解决实际经济社会问题的关键。如果研究对象是检验某一经济理论或假说,那么计量经济学模型的设定问题将很简单:根据理论或假说的内容要求即可完成;如果研究对象是某一实际经济问题,那么计量经济学模型的设定将会变得复杂,因为实际经济问题中难以捕捉的信息及不确定因素较多,避开模型设定偏误而构造出一个正确的模型并非易事。本文将以经济学为背景,探讨和分析回归模型的几类统计检验方法的内在信息,并进一步讨论如何利用这些统计检验方法来诊断模型出现的设定偏误问题,以实现计量经济学模型的正确设定。首先指出,本文将讨论的回归模型均指单方程形式的线性计量经济学回归模型,文中将其简称为回归模型或计量经济学模型。

关于计量经济学模型设定方面的研究,国外主要集中于两个方面:其一是对模型设定效果的评判要求;其二是模型设定正确与否的统计检验。在模型设定效果评判方面,Intriligator认为,构建回归模型有效的方法是找到对被解释变量有直接影响的、且不能被模型中已有变量所替代的解释变量[1]。Intriligator的观点体现了两个方面的含义,一是模型要简单,只取对被解释变量有直接影响的变量为解释变量;二是避开变量所表述的经济含义的重合,从模型的角度来说,即避开多重共线性的影响。Hendry和Richard认为,好的计量经济学模型应满足可获得有效数据、弱外生性解释变量、参数估计结果稳定、纯随机残差、原有条件下模型不能再改进等条件[2],这些都是模型避开设定偏误的内在要求。在模型设定偏误问题的检验方面,研究成果相对较多。Ramsey提出了一类检验模型是否存在设定偏误的方法,称为RESET(regression specification error test)检验[3],这是一种模型受约束条件的检验,可通过F检验或LM检验来完成。RESET检验可以用来判断模型是否存在设定偏误,但不能有效地提出正确的模型设定方式。后来Davidson、MacKinnon给出了和RESET检验相似的另一种检验形式[4]。Wallace对传统的只根据模型建立之后的检验结果来反映模型设定效果的做法进行了批判,他认为需要对所设定的模型进行预检验,以及时发现不足[5]。Hausman认为在模型设定正确的情形下参数估计结果应具有相同或相近的统计性质,他通过考察不同假定条件下参数估计值的有效性设计了一种检验方法,即Hausman检验[6],这种检验方法后来进一步发展为联立方程模型的设定检验,用于检验方程的联立性。MacKinnon、White及Davidson提出了一种被称为MWD检验的方法,用来判断设定回归模型时应选择线性模型还是线性对数模型[7]。Arminger与Schoenberg提出了一种协方差结构模型的设定检验,主要用于检验由于模型随机项和内生变量的相关性而造成的设定偏误[8]。国外的研究主要体现在模型设定的事后检验方面,而国内关于计量经济学模型设定研究主要集中于如何实现模型的正确设定方面。李子奈从计量经济学应用研究中总体回归模型设定的任务和目标出发,通过对总体模型设定的研究目的导向、经济学理论导向、数据关系导向的分析与评价,讨论了模型设定的唯一性、一般性、现实性和统计检验必要性等原则,提出总体回归模型设定的"经济主体动力学关系导向"原则和框架[9]。葛新权讨论了时间序列回归模型的设定问题,他认为设定此

基金项目:甘肃省高校人文社科重点研究基地甘肃经济发展数量分析研究中心资助项目(SLYB201205)
作者简介:刘　明(1981-),安徽霍邱人,博士研究生,讲师,研究方向:统计理论与方法、经济计量分析。

虚拟变量回归模型的应用:方差分析及异常值检验

刘　明
（兰州商学院 甘肃经济发展数量分析研究中心,兰州 730020）

摘　要:虚拟变量的作用不仅局限于量化品质型数据,在不同的设置环境中有着不同的重要应用。文章简述了虚拟变量模型在结构变化分析、分段回归、交互影响等方面的基本应用,并进一步讨论了虚拟变量模型在方差分析及异常值检验中的使用方法和该方法的显著功效,通过比较分析得出了"利用虚拟变量模型进行方差分析和异常值检验更优"的结论。

关键词:虚拟变量模型;方差分析;异常值检验
中图分类号:O212　　**文献标识码:**A　　**文章编号:**1002-6487(2014)16-0010-04

0　引言

虚拟变量的基本功能是将品质型数据转化为数值型数据,以便于将一些不能直接量化的因素引入到统计模型中加以研究。虚拟变量回归模型就是包含有虚拟变量的回归模型,是经典线性回归模型的一类重要形式。虚拟变量回归模型是一种将定性问题进行量化分析的重要工具,在实际应用中较为广泛。在线性回归模型中,解释变量和被解释变量都可以存在虚拟变量,本文考察的是虚拟变量充当解释变量的情形,在此情形下可以构造出不同结构的虚拟变量模型,它们在实际问题研究中有着不同的应用。本文是简述虚拟变量模型的常见应用,包括结构变化分析、分段回归分析和交互相应分析。讨论虚拟变量回归模型的方差分析的功能,对于这一功能也有文献进行过讨论[1,2],他们通过对方差分析的F检验和虚拟变量回归的F检验进行了统一性论证,本文中的这一论证结果和文献[1]、[2]是一致的,但本文重点从各自的统计思想出发阐述了方差分析的F检验和虚拟变量回归的F检验的统一性,这有利于从统计思想的角度对其进行认识。提出了利用虚拟变量进行方差分析的一个新的功能——部分均值相等检验,对虚拟变量回归模型的方差分析和传统方差分析的统计思想进行讨论,并进一步比较二者在实际问题分析中的功效。讨论了如何运用虚拟变量回归模型检验异常值,讨论了该方法的检验原理并给出了具体检验方法,同时与传统异常值检验方法如Dixon检验、Grubbs检验等进行了比较分析。

1　虚拟变量模型的基本应用

虚拟变量模型的基本应用主要在三个方面:一是对事物的结构变化进行分析,二是构造分段回归模型,三是刻画属性变量间的交互效应。

1.1　结构变化分析

结构变化分析是研究事物是否存在结构性的变化,其实质是检验所设定的模型在样本范围内是否为同一线性回归模型。模型存在的形式无碍乎是这样三种:截距变化而斜率不变;斜率变化而截距不变;截距斜率同时改变。此三种情形对应着三种虚拟变量回归形式:平行回归、交汇回归、相异回归。

考虑被解释变量 Y_t 和解释变量 X_t,设计虚拟变量 $D=\begin{cases}1 & \text{存在某属性}\\ 0 & \text{不存在某属性}\end{cases}$,平行回归模型结构形式如下:

$$Y_t=\alpha_1+\alpha_2 D+\beta_1 X_t+\mu_t$$

平行回归模型中虚拟变量的存在改变了模型的截距,而模型的斜率并无变化。

交汇回归模型的形式为:

$$Y_t=\alpha_1+\beta_1 X_t+\beta_2(DX_t)+\mu_t$$

该模型中多了一个由原解释变量和虚拟变量之乘积而形成的一个新的解释变量。

相异回归模型综合平行回归模型和交汇回归模型,其形式为:

$$Y_t=\alpha_1+\alpha_2 D+\beta_1 X_t+\beta_2(DX_t)+\mu_t$$

它表明,随着虚拟变量取值的不同,该模型的截距和斜率都会发生变化。

上述三类结构模型是虚拟变量模型最基本的形式。在进行结构变化分析时通常的做法是,首先根据样本数据估计出相异回归模型,利用相异回归模型的参数检验结论即可判断应该使用哪一类模型。具体说,可以根据参数显著性检验的t检验方法来检验相异回归模型中参数 α_2、β_2 是否为零来进行模型选择和结构变化分析:若 α_2 等于零而 β_2 不等于零,则可使用交汇回归模型进行考察;若 α_2

基金项目:甘肃省高校人文社科重点研究基地甘肃经济发展数量分析研究中心资助项目
作者简介:刘　明(1981-),男,安徽霍邱人,讲师,研究方向:统计理论与方法、经济计量分析。

【统计理论与方法】

一种收入分布函数序列的拟合方法及扩展应用

黄恒君,刘黎明
(首都经济贸易大学 统计学院, 北京 100026)

摘要:针对收入分布函数形式选择问题,提出具有"自适应"能力的收入分布序列拟合思路,给出基于 B-样条的收入分布函数形式,并对收入分布参数进行最小二乘估计。拟合了中国历年城镇居民收入分布序列;导出中国 1996-2009 年洛伦兹曲线和基尼系数;从函数角度刻画了城镇居民收入水平不断提高的同时,收入差距扩大的动态趋势;验证了城镇居民收入差距的变动轨迹体现着"阶梯形"扩大的特征。

关键词:收入分布拟合; 样条; 洛伦兹曲线; 函数型数据分析
中图分类号:O212.1 文献标志码:A 文章编号:1007-3116(2011)12-0014-05

一、引 言

居民收入(以下简称收入)分配问题是国内外经济学界研究的热点问题。近年来,随着中国经济的高速增长和体制改革的进一步深入,收入分配领域的矛盾日益凸现。正确认识当前居民收入分配问题,首先要探讨居民收入的测度问题。

收入分布曲线拟合是收入测度研究中的重要内容。目前,收入分布拟合主要有参数估计和非参数估计两种方法。参数估计是函数驱动型的,只要已知分布能逼近真实收入分布,收入研究中的定量问题大多就能得到顺利解决(如导出密度函数、洛伦兹曲线等)。该方法有着完善分布族,Mc Donald J B 对各种收入分布之间的关系进行了概括,包括实证分析中最常用的对数正态分布、伽马分布和帕累托分布[1]。实证研究表明:帕雷托分布可以较好拟合高收入组的收入分布[2][3]120-122;居民收入分布的中间部分接近对数正态分布[4-5];伽马分布在拟合效果和参数估计简单角度进行折衷,在整体拟合上更为适合[6]。也有人认为 Logistic 分布比伽马分布、对数正态分布和帕累托分布具有更好的拟合效果[7]。

收稿日期:2011-07-22
基金项目:国家社会科学基金项目《中国现行社会福利保障制度下城镇贫困人口的统计研究》(11BTJ002);首都经济贸易大学博士研究生科技创新项目《社会保障与税收对中等收入群体的影响研究》(CUEB2010533)
作者简介:黄恒君,男,浙江温州人,博士生,讲师,研究方向:调查技术与统计分析;
刘黎明,女,山东济南人,教授,博士生导师,研究方向:应用数理统计。

The Challenge Encountered by Traditional Statistics

LIU Chao[1a,b], WU Xi-zhi[2]
(a. School of Mathematics and Systems Science; b. LMIB of the Ministry of Education,
1. Beihang University, Beijing 100191, China; 2. School of Statistics, Renmin University of China, Beijing 100872, China)

Abstract: Via Breiman's concept of two kinds of statistical cultures, this paper aims at the black-box characteristicsof statistics, analyzing the challenges and crises encountered by traditional statisticsand the question of where is the future of statistics. Research results show that statistics must come back to its tradition that is to face real tasks dealing with data and to establish relevant theorems. Only in this way, we can deal with the challenges in the new era.

Key words: statistics; mathematics; data modeling; algorithmic modeling (责任编辑:李 勤)

大数据视角下名录库更新维护*

——基于互联网异源异构数据整合的探讨

傅德印　黄恒君　陶　然

内容提要: 统计系统基本单位名录库是统计数据质量的基石,现有数据源在成本、时效性、数据提供者负担方面存在劣势。为此,本文提出一种互联网大数据整合视角下的名录库更新维护思路: 从参与者行为、数据质量角度论证了将异源异构互联网数据作为名录库更新数据源的优势,讨论了名录库基本信息、属性信息及地理定位信息获取的技术手段,并给出应用实例。

关键词: 大数据; 名录库; 政府统计; 数据质量
中图分类号: C811　　文献标识码: A　　文章编号: 1002 - 4565(2015) 01 - 0005 - 06

Updating of Basic Unit Database: Internet Big Data Integration Approach

Fu Deyin　Huang Hengjun　Tao Ran

Abstract: Statistical Basic Unit Database is the cornerstone of data quality. There are disadvantages of existing data sources on cost, timeliness and burden of data providers. This paper proposes a updating approach of Basic Unit Database based on internet data integration, discusses the advantage of taking the internet data as an data source for updating and provides technology details of obtaining basic, attributive and geographic information of Basic Unit Database. An application is also mentioned.

Key words: Big Data; Basic Unit Database; Official Statistics; Data Quality

一、引言

随着传感器等技术的兴起,社交媒体等信息发布方式的涌现,数据正以前所未有的形式、速度、广度不断累积和增长。大数据现象引起国内外学术团体、政府机构及企业部门的浓厚兴趣。

就政府统计而言,大数据现象不可避免地对现有统计工作流程产生影响,目前已有了初步的探讨: 联合国统计署对政府统计中的大数据界定、数据来源、数据整合等问题提出了初步框架[1]。荷兰统计局已在价格统计、旅游统计等传统统计工作流程中应用互联网数据[2]。在我国,国家统计局也对大数据现象给予充分关注和积极应对[3]。

由于大数据处理方式与传统统计工作流程差异显著,大数据应用于政府统计工作应当包括三个递进的层次: 第一,利用异源异构大数据源,辅助和补充现有统计工作流程; 第二,利用大数据思维与方法,改造现有统计工作流程; 第三,创造全新的统计方法和指标。

我们认为,将大数据源融入到当前统计工作流程,辅助解决统计工作中亟待处理的问题,可以作为政府统计大数据应用迈出的第一步。由于基本单位名录库建设是政府统计工作的核心与基石,也是我国统计"四大工程"之一,本文基于文献[4]就名录库更新方面的大数据具体应用做出讨论,提出对大数据用于名录库建设方面的设想。

具体来讲,由于互联网成为大量信息的主流载体,本文将异源异构互联网信息作为数据源,试图探讨一种大数据整合视角下的名录库更新思路,作为

* 本文获得教育部人文社会科学重点研究基地重大项目"政府统计数据质量保证体系研究"(12JJD790010) 、全国统计科学研究重点项目"海量异源异构数据的采集、存储和分析方案研究"(2013LZ44) 、全国统计科学研究重点项目"基于普查涵盖误差测量技术的基本单位名录库维护与更新研究"(2011LX003) 资助。

海量半结构化数据采集、存储及分析*

——基于实时空气质量数据处理的实践

黄恒君　漆　威

内容提要：大数据现象及处理引起了社会各界的关注。本文以大数据宏观层面理论为依据，试图从微观层面讨论一类大数据的具体处理，归纳提出一种基于开源架构的海量半结构化数据采集、存储及分析自动化解决方案，并分析解决方案的开放性、融合性和经济性的特点，指出解决方案的可拓展方面。同时，结合海量空气质量实时数据，分析解决方案的具体开发细节，给出解决方案运行的经验做法，讨论分析过程的大数据压缩机制。

关键词：大数据；数据挖掘；空气质量；函数型
中图分类号：C812　　文献标识码：A　　文章编号：1002－4565(2014)05－0010－07

Massive Semi-Structured Data: Collection, Storage and Analysis

——Based on the Practice of Real-time Air Quality Data Processing

Huang Hengjun & Qi Wei

Abstract: Big data phenomenon and processing has aroused attention from all sectors of the community. Based on macro-level discussion of big data, this paper tries to treat a type of big data in case-level. An automation solution of massive semi-structured data collection, storage and analysis was proposed under open source framework. The features of our solution, which include openness, integration and economy, were discussed. The extension of the solution was also pointed out. Meanwhile, based on our massive real-time air quality data, this paper give out the specific development details, running experience and practice, and big data compression schemes also been discussed.

Key words: Big Data; Data Mining; Air Quality; Functional

一、引言

随着云计算、传感器等技术的兴起，微博、社交网络等信息发布方式的涌现，数据正以前所未有的形式、速度、广度不断累积和增长。大数据现象引起国内外广泛关注。

就统计机构而言，大数据的处理需求可能更为迫切——大数据现象影响到当前统计工作流程。面对大数据洪流，如何进行相应采集存储、开发利用，成为社会各界，特别是统计机构关注的焦点。

就笔者搜集到的资料看，目前，关于大数据现象及人们的反应[2]、大数据的界定与管理[3]、大数据应用与研究所面临的问题与挑战[4]、大数据处理的可能关键技术[5]等，已有学者进行了探讨。国家统计局也正对大数据做出积极的应对[1]。笔者认为，作为新兴事物的大数据，这些宏观层面的讨论是十分必要且对大数据研究具有重要的指导价值。进一步地，若以宏观层面理论探讨为指导，从微观层面给出一个大数据的具体解决方案，则对大数据的理解和研究也是不无裨益的。

为此，本文以笔者编程开发的实时发布空气质量大数据解决方案①作为实例，归纳提出一个针对半结构化数据源，具有连续发展特征大数据的采集、存储和分析解决方案。有别于发展高效运算的大规

* 本文获得教育部人文社会科学重点研究基地重大项目“政府统计数据质量保证体系研究”(12JJD790010)；全国统计科学研究重点项目“海量异源异构数据的采集、存储和分析方案研究”(2013LZ44)；全国统计科学研究重点项目“基于普查涵盖误差测量技术的基本单位名录库维护与更新研究”(2011LX003)资助。

① 截止 2013 年 8 月 6 日，该解决方案已在笔者个人服务器上稳定运行超过 3 个月。

大数据在政府统计中的角色定位及应用路径探讨*

漆 威 黄恒君

内容摘要：大数据为政府统计生产带来了重大发展机遇和挑战。本文从政府统计应用视角，对大数据在政府统计生产中的可用性进行分析。首先，从大数据生产流程和产业链视角评价大数据在政府统计中的角色定位；其次，从统计生产的技术视角探讨大数据在政府统计开发应用中的特点；最后，给出大数据在政府统计中的开发应用路径建议：以传统统计工作流程为主，将大数据源融入其中，注重大数据产品的可比性，注重大数据源的累积与开拓。

关键词：大数据；政府统计；数据源

中图分类号：C829　　文献标识码：A　　文章编号：1004-7794(2016)04-0060-05

DOI：10.13778/j.cnki.11-3705/c.2016.04.012

一、引言

大数据的出现深刻影响着社会发展的方方面面，大数据已经在一些国家成为重要的战略资源。《中共中央关于制定国民经济和社会发展第十三个五年规划的建议》中两处着笔提到大数据，一是“实施国家大数据战略，推进数据资源开放共享”，把大数据及其产业的发展提升到了国家战略层面，这是从未有过的高度；二是强调“运用大数据技术，提高经济运行信息及时性和准确性”，这则是直面大数据的统计生产功能。同时，也提出一个讨论命题：大数据及其产业与统计生产的关系。

来势凶猛的大数据现象对各行各业均是一种机遇，也是一种挑战。对于作为重要数据生产者的政府统计机构，大数据时代的影响、机遇与挑战则是首当其冲和前所未有的。随着数据用户对统计产品的要求越来越高，而数据提供者的配合程度又越来越低，加之非政府统计团体的统计能力越来越强，并开始使用传统调查以外的大数据手段搜集数据，发布与政府统计机构相似的统计产品。在这样的背景下，全世界政府统计机构都将面临新的路径抉择：是坚持传统统计调查，全方位拥抱大数据，还是实现两者的有机融合？这一切最终取决于数据用户对统计产品的需求和对数据质量的要求，以及非政府团体的统计能力所引发的挑战。

当然，无论哪种方式适用于政府统计生产，我们都有必要从不同的角度剖析大数据现象在政府统计工作的意义、价值与角色定位。对此，目前已经有了一些讨论，诸如国家统计局对大数据统计应用的方法、技术、案例等的系统性探索[1]，政府统计应对大数据挑战的稳扎稳打策略讨论[2]，政府统计视角下大数据的概念、数据源、制度建设等的探讨[3]，以及大数据在政府统计中应用路径的前瞻[4]等。

在已有研究的基础上，本文同样是以政府统计应用为出发点，探讨大数据在政府统计生产中的可用性。不同的是，本文从大数据内涵、业务流程和大数据产业链的视角，评价大数据在政府统计生产中的角色定位；梳理出大数据在政府统计应用中的典型特征和侧重点，进而给出大数据在政府统计中的

*基金项目：教育部人文社会科学重点研究基地重大项目“政府统计数据质量保证体系研究”(12JJD790010)；陇原青年创新人才扶持计划项目“基于大数据整合的‘废旧数据’应用研究”(14GSD95)；甘肃省财政厅高校基本科研业务费项目“大数据整合下的统计调查技术及其经济应用研究”(GZ14007)。