**对外经济贸易大学本科教学案例**

 **结项材料**

教学案例名称：基于网络抓取数据的宏观经济指标预测案例分析

编写负责人姓名： 唐晓彬 彬

联系电话： 18811515185 5

院（部）名称： 统计学院 院 院

**案例1**

**基于互联网大数据背景下的CPI预测研究**

**摘 要：**本案例预测CPI的方法能克服传统CPI统计方法的滞后性。基于网络搜索数据，以宏观层级、微观层级和金融层级为研究视角，选取与CPI相关的网络搜索词，获取其百度指数，运用PCA降维并合成宏观类综合指数、微观类综合指数和金融类综合指数。并以该三类综合指数为基础，建立CPI预测模型并拟合出CPI的预测值。结果显示，基于网络搜索数据拟合出的CPI预测值与真实的CPI的走势相吻合，样本内和样本外预测的平均绝对误差均较小，预测效果较好。

**关键词：**大数据时代；PCA降维方法；网络搜索指数；CPI预测

**一、案例背景**

居民消费价格指数（CPI）反映了居民家庭一般所购买的消费品和服务的价格水平变动情况。CPI不仅与人民群众的生活关系密切，而且在整个宏观经济体系中占有重要地位。它是国民经济核算体系中的重要指标，对经济分析、决策、价格总水平监测和宏观调控的具有重要的参考价值。目前，CPI的数值依然是通过相关的统计调查来获取的，一般在次月中旬发布。而国家在制定一项宏观政策之前，往往需要充分了解当前的物价水平，由于我国CPI发布的滞后性是制约宏观调控的重要因素。因此，准确地预测CPI具有重大的现实意义。

国内外学者们关于CPI的预测的研究已经比较成熟。前期的研究主要集中于定性预测方法和定量预测方法。定性预测主要是在经济理论的基础上通过找出影响CPI的因素来预测它的未来走势（王学庆[1]），但该预测方法存在以下两点缺陷：（1）一些经济理论通常是在一种比较理想的假设条件下才能够成立，而现实的经济社会情况错综复杂，往往和理论有一定的差距；（2）即使现实和理论大体相符，但定性预测只能提供未来CPI的大致走势，很难做到对CPI的精确预测。定量预测要明显优越于定性预测，其定量预测多采用时间序列（张鸣芳[2]）、灰色预测（陆明希等[3]）、组合预测（贺勇[4]）、神经网络（王宇[5]）等方法，其原理都是用已经发生的数据预测未来。然而，该预测方法也有其局限性，表现为历史数据只是反映当时的情况，而宏观经济时常会发生结构性突变，此时，这种预测方法的准确性将会大打折扣。

随着互联网大数据时代的来临，市场主体们获取信息的主要渠道已经由线下转移到线上。在众多网络服务中，网络搜索已经成为最主要的信息入口。人们在进行网络搜索的同时，网站也记录了人们的搜索行为，因此，消费者和生产者在商品市场的行为会在网络搜索数据上得到反映，网络搜索数据一定程度上反映了商品或服务的供求关系，由供求关系决定的价格应该与网络搜索数据存在某种相关关系。由于这种价格传导机制具有时滞性，因此网络搜索数据的变化也会领先于价格的变化，从而可以考虑利用网络搜索相关数据来预测CPI。

**二、理论模型构建**

基于网络搜索数据对社会经济活动的预测已经在国内外取得了一定的研究成果。例如，Ginsberg[6]构建了基于Google搜索数据的流感监测模型，能够提前1至2周预测流感的发病趋势，然而在2013年的预测中，这个模型在预测流感发病趋势时却出现重大偏差。国内学者张崇[7]基于谷歌趋势数据从宏观，微观两个角度出发预测出了CPI走势，预测的平均绝对误差较小，预测效果较好。但是，他在利用时差相关法合成指数的时候存在如下问题：（1）并未考虑到各关键词的搜索指数存在相关性；（2）在合成指数时，以和CPI的相关系数作为各关键词的权重，科学依据不充分。为此，本案例参考孙毅等[8]的做法采用主成分分析方法对高维数据进行降维，合成几个不相关的主成分，并以方差贡献率作为权重合成分类指数。然后多次尝试模拟出最优模型。

**三、计量模型设定及检验**

**1. 百度搜索指标的选取**

**1.1 百度指数的介绍**

关于网络搜索指数，大部分学者均选取谷歌搜索指数作为网络搜索指数的代表例如Ginsberg[6]、Doomik[9]，但自从2010年上半年后，谷歌搜索对中国大陆用户停止使用，而百度一跃成为了内地第一搜索引擎。因此，我们认为百度指数更能反映近几年中国网民的兴趣关注点。百度搜索指数是以网民在百度的搜索量为数据基础，以关键词为统计对象，科学分析并计算出各个关键词在百度网页搜索中搜索频次的加权和。百度指数官网通常提供三种关键词的搜索指数：日数据、最近七天或最近30天的平均搜索指数。因日度数据的获得能追溯到2011年1月1日，另外两种数据只能反映在当前日期上的近一周或近一个月的趋势，样本长度太短。所以本案例采用了从2011年1月开始的关键词百度指数的日度数据。根据使用百度搜索来源的不同，搜索指数分为整体趋势、PC搜索指数和移动搜索指数。整体搜索指数是由PC搜索指数和移动搜索指数这两部分构成，所以案例使用的数据是各个关键词从2011年1月1日开始每日的百度指数中的总体搜索指数，数据均在百度指数官方网站查询获得。

**1.2 关键词的选取及数据的处理**

物价指数的变动可以由人们对相关信息的网络搜索数据的变动来反映。宏观形势的变动会对CPI产生影响；微观层面供求关系的变动是物价指数变动的直接原因；由于金融日益成为国民经济中非常重要的部分，在本案例中，把金融从宏观角度中分离出来，作为一个独立的类别，以期更细致更精准地反映和预测物价。为此，案例从宏观层面、微观层面和金融层面三大类选取网络搜索关键词。

首先，根据CPI的影响因素与CPI的构成选取初始关键词，例如，宏观层面上选取：GDP增长、就业、CPI等；微观层面选取诸如：猪肉价格，房价，家电下乡，蔬菜等；金融层面：货币，货币政策，利率，汇率，股票市场等；然后，根据百度的相关搜索推荐功能将每一个初始关键词扩展一些相关的词。比如从初始关键词“CPI”扩展成“CPI是什么”、“CPI指数”等，然后剔除掉语义重复的词，最终剩下89个关键词，详见表1。最后，我们将根据上述关键词搜索到的数据进行处理，具体处理方法如下：（1）因官方统计的物价指数是月度指标，为此，我们将每个关键词的日度数据加总平均为月度数据；（2）由于网络搜索数据具有短期波动大的特征，我们利用三项移动平滑法消除短期波动，突显长期趋势；（3）由于同比数据能够剔除变动的季节效应，我们选取官方公布的CPI同比数据作为研究的对比对象。为了与CPI同比数据具有可比性，我们将网络搜索指数的月平均数据转化为月度同比数据，本文月度同比网络搜索指数的样本区间从2012年1月开始。

表1本文选取的关键词

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 宏观类 | CPI | GDP | PPI | 财政赤字 | 物价上涨 | 中国投资 |
| 宏观调控 | 经济增长 | 就业难 | 就业形势 | 通货膨胀 | 税收 |
| 工资 | 物价 | GDP增长率 | 大学生就业 | 居民消费价格指数 | 　 |
| 微观类 | 晒工资 | ktv | 保健 | 冰箱价格 | 菜价 | 车价 |
| 春季旅游 | 房价 | 房价上涨 | 瘦身减肥 | 房价走势 | 花生价格 |
| 酒店 | 酒店预订 | 看病难 | 粮食 | 旅游 | 旅游景点 |
| 食用油价格 | 房价收入比 | 铁路货运价格 | 携程旅游网 | 液晶电视价格 | 猪肉价格走势图 |
| 团购 | 家教 | 教育 | 教育支出 | 今日菜价 | 旅游网 |
| 猪肉 | 猪肉价格 | 医疗费 | 装修 | 车价查询 | 家电下乡 |
| 冰箱价格 | 笔记本价格 | 工业用电价格 | 医疗费报销 | 　 | 　 |
| 金融类 | 存款利率 | 贷款利率 | 货币供给 | 广义货币 | 汇率 | 汇率查询 |
| 汇率换算 | 利率 | 货币政策 | 利率调整 | 利息 | 美元汇率 |
| 石油期货 | 资产 | 央行 | 基础货币 | 今日汇率 | 金融 |
| 信贷 | 投资理财 | 央行加息 | 银行利率 | 债券基金 | 棉花价格 |
| 金融危机 | 商品期货 | 准备金 | 棉花期货 | 石油价格 | 利率市场化 |
| 房地产价格 | 宽松货币政策 | 　 | 　 | 　 | 　 |

**2 CPI的预测分析**

**2.1 PCA数据降维**

采取主成分分析方法将三大类共89个关键词进行降维，形成三大类综合指数，即宏观类综合指数、微观类综合指数和金融类综合指数。采用该方法不仅能够消除各关键词之间的相关性，而且能够在综合大类指数时赋予权重提供理论依据。下面以宏观类为代表做主成分分析，得到如下的主成分方差分析提取表：

表2 主成分方差分解提取表

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 成分 | 特征值 | 方差所占百分比 | 累计方差 | 累计方差所占百分比 |
| 1 | 4.7946 | 0.282 | 4.7946 | 0.282 |
| 2 | 2.6753 | 0.1574 | 7.4699 | 0.4394 |
| 3 | 2.399 | 0.1411 | 9.8688 | 0.5805 |
| 4 | 1.8396 | 0.1082 | 11.7085 | 0.6887 |
| 5 | 1.0025 | 0.059 | 12.711 | 0.7477 |
| 6 | 0.8804 | 0.0518 | 13.5914 | 0.7995 |
| 7 | 0.6866 | 0.0404 | 14.2779 | 0.8399 |
| 8 | 0.6421 | 0.0378 | 14.92 | 0.8776 |
| 9 | 0.5666 | 0.0333 | 15.4866 | 0.911 |
| 10 | 0.3656 | 0.0215 | 15.8522 | 0.9325 |
| 11 | 0.2923 | 0.0172 | 16.1445 | 0.9497 |
| 12 | 0.2288 | 0.0135 | 16.3733 | 0.9631 |
| 13 | 0.1878 | 0.011 | 16.5611 | 0.9742 |
| 14 | 0.1681 | 0.0099 | 16.7292 | 0.9841 |
| 15 | 0.1269 | 0.0075 | 16.856 | 0.9915 |
| 16 | 0.0932 | 0.0055 | 16.9492 | 0.997 |
| 17 | 0.0508 | 0.003 | 17 | 1 |

由上表可知，根据需提取85%的信息原则，宏观类需提取前八个主成分。其因子载荷如下表3所示：

表3 前八个主成分的因子载荷

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 　 | PC 1 | PC 2 | PC 3 | PC 4 | PC 5 | PC 6 | PC 7 | PC 8 |
| CPI | 0.35 | -0.13 | -0.11 | 0.11 | -0.38 | 0.21 | -0.21 | -0.12 |
| GDP | 0.25 | 0.08 | 0.23 | 0.26 | -0.17 | 0.03 | -0.27 | -0.23 |
| GDP增长率 | -0.01 | 0.07 | 0.51 | -0.31 | -0.01 | 0.22 | -0.05 | 0.21 |
| 财政赤字 | 0.2 | 0.04 | 0.02 | -0.19 | 0.68 | 0.2 | 0.04 | -0.57 |
| 大学生就业 | 0.25 | 0.43 | -0.1 | -0.02 | -0.03 | 0.05 | -0.19 | -0.03 |
| 工资 | 0.11 | 0.1 | 0.26 | 0.54 | 0.21 | -0.23 | 0.25 | -0.08 |
| 宏观调控 | -0.16 | 0.13 | 0.28 | 0.48 | 0.21 | -0.09 | -0.29 | 0.15 |
| 经济增长 | 0.24 | -0.03 | 0.45 | -0.1 | 0.16 | -0.04 | 0.2 | 0.38 |
| 就业难 | 0.03 | 0.54 | -0.2 | 0.06 | -0.03 | -0.08 | -0.12 | 0.06 |
| 就业形势 | -0.01 | 0.54 | -0.04 | -0.18 | -0.06 | -0.03 | 0.15 | 0.14 |
| 居民消费价格指数 | 0.33 | 0 | 0.11 | -0.11 | 0.1 | 0.16 | -0.55 | 0.15 |
| 进出口贸易 | -0.04 | 0.06 | -0.11 | 0.39 | -0.01 | 0.82 | 0.29 | 0.14 |
| 通货膨胀 | 0.39 | 0 | -0.08 | -0.04 | -0.02 | -0.07 | 0.32 | 0.26 |
| 物价 | 0.32 | -0.28 | 0.02 | 0.16 | -0.19 | -0.21 | 0.08 | -0.09 |
| 物价上涨 | 0.36 | -0.18 | -0.16 | -0.06 | 0.2 | 0.07 | 0.03 | 0.25 |
| 中国投资 | 0.37 | 0.21 | -0.11 | 0.03 | -0.03 | -0.16 | 0.26 | -0.1 |
| PPI | 0 | 0.13 | 0.45 | -0.15 | -0.4 | 0.12 | 0.24 | -0.42 |

然后，我们分别算出各主成分的得分后，并按照方差贡献率进行加权，最后得到宏观类综合指数（XH）。

同理可得，微观类因素提取了四个主成分综合信息，得到微观类综合指数（XW）；金融类因素提取了八个主成分，构建了金融类综合指数(XJ)，分别为：

根据上式（1）~（3）分别得出宏观类指数、微观类指数和金融类指数。

**2.2 CPI预测模型**

进一步地，我们基于上述三大类搜索指数建立CPI的拟合模型。由于环比CPI未消除季节性因素，因此，本文的被解释变量选月度同比CPI；解释变量为宏观类综合指数XH、微观类综合指数XW、金融类综合指数XJ。由于存在价格粘性现象，解释变量加入前一期CPI。多次尝试拟合模型后，我们发现宏观类综合指数（XH）的前第15期、微观类综合指数（XW）的前第4期、金融类综合指数（XJ）的前第3期和CPI的前第10期对当期CPI具有显著的影响，且增加了模型的解释能力。最后，拟合出的CPI预测模型为：

为0.8752，调整为0.8521，统计量为37.88，值为1.78，各变量的值和显著程度如下表：

 表4估计结果

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 　 | 系数 | 标准差 | t值 | p值 |
| XH(-15) | -0.0046 | 0.001115 | -4.1571 | 0.0003 |
| XJ(-3) | -0.00079 | 0.000224 | -3.5224 | 0.0015 |
| XW(-4) | -0.0098 | 0.003271 | -3.0076 | 0.0056 |
| CPI(-1) | 0.5624 | 0.089699 | 6.2701 | 0 |
| CPI(-10) | -0.2483 | 0.092244 | -2.692 | 0.012 |

由上表可知，各个解释变量均显著，且调整达到0.8521，拟合效果显著。从DW值来看，DW值接近于2，表明残差项不存在自相关性，我们对残差进行单位根检验，其p值为0.0017，表明在1%的显著水平下，残差为平稳性时间序列。

我们分别对CPI进行样本内预测与样本外预测。通过样本内预测出2012年1月到2015年12月的CPI与真实的CPI如下图所示：

图1 CPI的拟合值和真实值

从上图可以看出，我们拟合出的CPI与真实的CPI的走势基本吻合，且样本内预测的均方根误差为0.25，平均绝对误差为0.1941，均小于张崇（2012）[7]的均方根误差（0.62）和平均绝对误差（0.48），我们的预测效果要优于张崇对CPI的预测。而且，我们建立的CPI拟合模型自变量中的搜索项分别是第 t-15和 t-3、t-4 期，同比CPI的第t-1 和 t-10 期，所以在得知当月的CPI时，就可以预测出下个月的CPI，比国家统计局官方网站公布的CPI提前了1个月，因此，本文提出的CPI预测方法克服了传统方法的滞后性。

进一步地，对样本外进行了预测，其预测结果如下表5所示：

表5样本外预测结果

|  |  |  |  |
| --- | --- | --- | --- |
| 　 | 2016年1月 | 2016年2月 | 2016年3月 |
| CPI真实值 | 101.75 | 102.28 | 102.3 |
| CPI拟合值 | 102.11725 | 102.29217 | 102.34649 |

从上表5可以看出，2016年1月的真实CPI为101.75，我们预测出CPI值为102.11725

2016年2月的真实CPI为102.28，预测出CPI值为102.29217；2016年3月的真实CPI为102.3，我们预测出CPI值为102.34649。预测出的CPI与真实的CP非常接近，其该预测的平均绝对误差为0.1419，其值也较小，表明该预测模型样本外预测效果较好。

**四、小结**

对本案例的学习，需要强调以下几点：

1.基于网络搜索数据来预测宏观经济指标，是大数据时代的一种重要预测方法。网络数据大量且实时产生，具有很高的可获得性。在大数据时代，相对传统调查数据，网络搜索数据能更灵敏更精准地反应宏观政策经济走势。借助网络搜索数据，本案例对我国CPI进行了预测研究，研究结果表明基于网络搜索数据能够准确的预测CPI的变动趋势。案例为传统CPI的预测方法提供了传统有效补充，其主要体现为：采用主成分分析方法来提取综合指数，解决了以往学者研究的关键词之间存在相关性问题，并能够科学合理地赋予权重。实证效果预测显示，预测偏差很小。

2.案例的不足之处在于选取关键词时，未能很好地建立一种科学有效的选词机制。一套优良的选词机制，不仅能够选出对所研究问题高度灵敏的词，而且还应建立一种关键词实时更新机制，使其能快速准确反映最新的，这是今后研究的重点。

**参考文献：**

[1]王学庆，郭丽岩．2009年CPI分析及2010年预测［J］．中国物价，2009（10）：3-6．

[2]张鸣芳，项燕霞，齐东军．居民消费价格指数季节调整实证研究［J］．财经研究，2004( 3)：133-144．

[3]陆明希，严广乐．基于神经网络灰色Verhulst算法的CPI预测模型［J］．统计与决策，2009（17）：52-53．

[4]贺勇，诸克军，黄毓芝．基于 GEP 的经济时间序列组合预测方法研究［J］．科技管理研究，2008( 5) : 180-182．

[5]王宇，李旭．基于BP神经网络的我国 CPI 预测与对策［J］．计算机科学，2009( 10) : 256-264．

[6] Ginsberg J，Mohebbi M H，Patel R S，et al． Detecting influenza epidemics using search engine Query data［J］． Nature，2009，457，1012-1014．

[7]张崇，吕本富，彭赓，刘颖．网络搜索与CPI的相关性研究［J］．管理科学学报，2012(7): 50-58．

[8]孙毅，戴维，董纪昌，吕本富．基于主成分分析的网络搜索数据合成方法研究［J］．数学的实践与认识，2014，44（21）：:121-128．

[9] Doornik J A． Improving the Timeliness Of Data On Influenza-Like Illnesses Using Google Search Data［R］． Technical Report，University of Oxford，2009．

**Research based on the background of big data in the CPI prediction**

**Abstract：In this paper, from the perspective of web search, we select web-search key words related to CPI from macroscopic,microscopic and financial aspects, acquire its Baidu Index, and use PCA to implement dimension reduction and compound macro comprehensive index, the micro comprehensive index and financial comprehensive index.Then based on these three indexes, we establish CPI forecasting model and obtain its fitted value. The result shows that the fitted value on account of web search corresponds to the real CPI very well. Mean absolute error of both forecasting in sample and forecasting out of sample are pretty slight. In addition, the forecasting method in this paper can overcome the hysteresis of traditional forecasting methods.**

**Key words: big data era； PCA data dimension reduction；web-searching index；CPI forecasting**

**案例使用说明**

**1.教学目的与用途**

在互联网大数据时代蓬勃发展的今天，指导学生学习如何运用网络搜索指数，帮助学生掌握PCA降维的方法，并在此基础上学会建立有效的预测模型。本案例适用于计量经济学、多元统计分析等统计课程，不同基础的学生均可选择部分或全部内容来学习。

**2.启发思考题**

（1）由于需要统计调查来获取，CPI数值的公布具有滞后性。那么随着互联网大数据时代的来临，可否通过网络搜索数据来对CPI的值做出预测呢？

（2）如果可以通过网络搜索数据对CPI的值做出预测，选取哪些数据指标能更具有代表性？它们之间的权重又该如何确定？

（3）如何评估CPI预测模型的预测效果？

**3.分析思路**

CPI的数值是通过相关的统计调查来获取的，一般在次月中旬发布，具有滞后性。而由供求关系决定的价格与网络搜索数据存在某种相关关系，由于这种价格传导机制具有时滞性，因此网络搜索数据的变化也会领先于价格的变化，从而可以考虑利用网络搜索相关数据来预测CPI。本案例采取主成分分析方法，将三大类共89个关键词进行降维，形成三大类综合指数，即宏观类综合指数、微观类综合指数和金融类综合指数，以方差贡献率作为权重合成分类指数，然后多次尝试模拟出最优模型，并在此基础上对模型的预测效果展开评估。

**4.理论依据与分析**

基于网络搜索数据对社会经济活动的预测已经取得了一定的研究成果，我们可根据指数数据分角度预测CPI的走势。因此，本案例采用主成分分析方法对高维数据进行降维，合成几个不相关的主成分，并以方差贡献率作为权重合成分类指数，然后通过多次试验模拟出最优模型。

主成分分析是一种利用降维思想，在损失很少信息的前提下，把多个指标转化为几个综合指标的多元统计方法。它把原来的n个特征用数目更少的m个特征取代，新特征是旧特征的线性组合，这些线性组合最大化样本方差，尽量使新的m个特征互不相关。从旧特征到新特征的映射捕获数据中的固有的变异性。即通过对原始变量相关矩阵或协方差矩阵内部结构关系的研究，利用原始变量的线性组合形成几个综合指标，在保留原始变量主要信息的前提下起到降维与简化问题的作用，使得在研究复杂问题时更容易抓住主要矛盾。

**5.关键要点**

通过教学，希望学生能够：增强大数据时代下搜索和应用网络数据的能力；掌握运用主成分分析来提取综合指数的方法，并能够合理地赋予权重；了解CPI的传统预测方法，并在此基础上学习构建有效的预测模型。

**6.建议课堂计划**

建议结合宏观的经济背景分三部分展开教学。首先使学生了解网络搜索指数如百度指数的运作方式，学习按类选取网络搜索关键词的操作方法。其次介绍主成分分析的理论与基本思想，体现主成分分析在降维过程中的应用。最后予以案例分析，引领学生在掌握基础知识后学习构建有效的预测模型。

**案例2**

**大数据背景下的CPI实时预测研究**

**摘 要：本文基于2011年1月至2016年8月编制CPI的93类商品或服务的百度指数，基于Elastic-Net惩罚因子的分布滞后模型，构建了CPI实时预测模型，实证检验了网络搜索数据对全国CPI、城镇CPI和农村CPI的实时预测功能。研究结果表明，基于网络搜索数据拟合出的CPI预测值与真实的CPI的走势相吻合，预测效果较好，该方法克服了传统CPI统计计算方法上的时间滞后性。**

 **关键词：大数据；CPI；实时预测**

**一、案例背景**

一直以来，CPI的预测是学术界研究的一个热点问题。关于CPI的预测，多数学者主要集中于对CPI的当期预测的研究，该类CPI预测方法主要有：时间序列预测方法、灰色预测方法、组合预测方法和神经网络预测方法。但上述各类预测方法具有一定的局限性，表现在采用的数据均为当期特定数据（current-vintage data），即研究者在做研究时所能得到的最终修订的数据，然而该数据并非真实宏观经济最初情况的表象，而是经历了多次的常规修订和全面修订，其中必然包含噪音信息，从而导致预测结果的偏差较大。

准确地预测CPI必须搜集当前经济状态的实时信息。实时数据（real-time vintage data）是初次发布信息的数据，由于具有即时产生和可获得性的特征，能很好地避免未来时间对过去某一特定时点数据的干扰问题。在当前的大数据时代下，海量的实时数据的即时获取已经成为现实，而网络搜索数据具有大数据特征的数据类型。因此,在本案例中，尝试利用网络搜索数据对CPI进行实时预测。

**二、理论模型构建**

本案例应用了Elastic-net估计方法原理。在大数据背景下，数据往往呈现高维度、样本长度短的特点，即大p小n。如果直接对其进行建模，会大大降低估计的精度。此外，解释变量当期和其滞后期往往存在高度相关关系，模型产生的多重共线性会进一步降低预测的准确性。因此，为了实现对该类型的数据进行有效地拟合，通常做法是对变量进行降维或者进行有效的变量选择。

最常用的降维方法就是因子分析。该种方法对高维信息进行降维，提取出若干个具有代表性的因子，但缺点是该因子的经济含义不明确。而Elastic-Net（Zou H and Hastie,2005）是一种有效的变量选择方法，该方法不仅避免了逐步回归进行变量选择时产生的模型稳定性低的特点，而且避免了Lasso方法只能得到过于稀疏模型的缺陷。下面对Elastic-Net估计方法的原理进行详细地介绍。

关于线性回归模型： （1）

其中，，，

，，。

定义：

 （2）

可得：

 （3）

理论上，这种基于岭回归和Lasso回归的凸组合可以得到模型的稀疏解，但实践效果并非十分理想。Hui Zou and Hastie Trevor（2005）提出了对该方法进行尺度化 处理的做法，即使用下列式（4）解决了稀疏解的问题。

 （4）

Elastic-Net参数估计过程如下：首先给出的参数空间，对每一个，结合LARS-EN算法[[1]](#footnote-2)产生elastic net回归的所有解决路径，然后根据K重交叉验证法，决定相应的另一个参数。最终，再根据均方误差最小的及相应的的值确定模型的。

**三、计量模型设定及检验**

**1.关键词选取**

国家统计局将居民日常消费所涉及的基本商品和服务分为八大类，分别为食品、烟酒及用品、衣着、家庭设备用品及维修服务、医疗保健和个人用品、交通和通信、娱乐教育文化用品及服务以及居住。每一大类下又分有若干个中类，每一中类下又有若干个小类，共计262个小类。由于某些商品的名称比较生僻，或与人们的日常表达习惯不同，如果直接以这262类商品或服务的名称作为键入的关键词，在百度指数官网查询，这种做法不合理。考虑到该问题，本文对每一个关键词，均结合着人们日常生活对该种商品或服务的表达、搜索习惯，进行主观的思考甄别，多次尝试，选定最合适的关键词，剔除掉百度指数过低的关键词，最终我们选定了93个关键词，如下表1所示：

表1 选出的关键词

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 食品类 | 大米价格 | 面粉价格 | 黄豆价格 | 白糖价格 | 猪肉价格 | 牛肉价格 |
| 羊肉价格 | 鸡肉价格 | 茶叶价格 | 蔬菜价格 | 红薯价格 | 食盐价格 |
| 食用油哪种好 | 面包店加盟 | 牛奶品牌排行榜 | 淡水鱼价格 | 奶粉排行榜10强 | 巧克力品牌 |
| 坚果 | 零食 | 酸奶 | 饮料 | 水果 |  |
| 烟酒及用品 | 香烟价格表 | 葡萄酒品牌 | 啤酒品牌 | 白酒 |  |  |
| 衣着 | 衬衫 | T恤 | 裤子 | 套装 | 连衣裙 | 外套 |
| 休闲裤 | 内衣 | 童装 | 男鞋 | 女鞋 | 女包 |
| 男包 |  |  |  |  |  |
| 家庭设备用品及维修服务 | 电风扇什么牌子好 | 冰箱哪个牌子好 | 空调排行榜 | 微波炉哪个牌子好 | 家具品牌 | 室内装饰 |
| 毛毯批发 | 被子品牌 | 床上用品 | 餐具 | 厨具 |  |
| 医疗保健个人用品 | 医疗器械 | 中药材 | 中成药 | 保健品 | 医疗费用 | 首饰 |
| 化妆品排行榜 | 护肤品排行榜 | 手表品牌排行榜 | 美容院 |  |  |
| 交通通信 | 手机 | 汽车 | 二手车 | 汽车之家 | 二手车交易市场 | 手机排行榜 |
| 汽油价格 | 柴油批发价 | 车险哪家好 | 话费充值 | 快递查询 |  |
| 娱乐教育文化用品及服务 | 液晶电视机价格 | 摄像机价格 | 数码相机哪个牌子好 | 笔记本电脑排名 | 户外运动品牌 | 网游排行榜 |
| 电视剧排行榜 | 儿童玩具 | 宾馆预订 | 景点门票 | 家用音响 | 乐器培训 |
| 健身房 | 旅游 | 文具 |  |  |  |
|  |  |  |  |  |  |  |
| 居住 | 液化石油气价格 | 天然气价格 | 装修材料 | 房租 | 水电费 | 　 |

我们对该93个关键词的日百度指数进行处理，其处理原因及方法如下：（1）因官方统计的物价指数是月度指标，为此，我们将每个关键词的日数据加总并取平均作为其月度数据；（2）由于网络搜索数据具有短期波动大的特征，我们利用三项移动平滑法消除短期波动，突显长期趋势；（3）由于同比数据能够剔除变动的季节效应，我们选取官方公布的同比CPI数据作为研究的被解释变量，其样本区间为2012年1月到2016年8月。为了与同比CPI数据具有可比性，我们将网络搜索指数的月平均数据转化为月度同比数据。所以，月度同比网络搜索指数的样本区间与同比CPI一致。

**2.模型设定**

根据上述Elastic Net模型的原理，我们选取关键词的百度指数当期及滞后期作为解释变量，同比CPI月度数据为被解释变量。本文建立如下CPI的实时预测模型：

 （9）

公式中的表示第个关键词的第阶滞后项的百度指数。由于样本容量有限，设定p为2。故原始进入模型的解释变量共有279（933）个。

**3.实证结果**

我们按照Hui Zou and Hastie Trevor（2005）的做法，的参数空间设定为内按等比数列取的100个值。结合R软件，采用K重交叉验证方法时，通常做法是令K=10。对每个参数组合构造Elastic-Net估计量,计算出最小的均方误差（MSE），选出最小均方误差下相应的的值。

为了研究网络搜索指数对不同地区区域的预测能力，本文分别就全国同比CPI、城镇同比CPI(SCPI)和农村同比CPI(RCPI)与93个关键词当前期及滞后一期，二期的百度指数进行实证检验。经过实证研究，结果表明，当为0.00008317638、0.00000144544、0.000174时，CPI、SCPI、RCPI的均方预测误差最小。所以，在这些特定的值下拟合出来的模型即为所需要的最优模型。在预测全国CPI的模型中，大部分的变量均被选中（剔除的变量见表2），这说明这些变量的当期或滞后期对CPI的变动具有显著影响。

表2 全国CPI实时预测模型剔除掉的变量[[2]](#footnote-4)

|  |  |
| --- | --- |
| 当期变量 | 黄豆价格，家具品牌，裤子，连衣裙，面包店加盟，奶粉排行榜10强，牛奶品牌排行榜，牛肉价格，啤酒品牌，葡萄酒品牌，食用油哪种好，蔬菜价格，套装，香烟价格表，饮料，被子品牌，车险哪家好，电视剧排行榜，儿童玩具，二手车，健身房，景点门票，汽车之家，手表品牌排行榜，手机排行榜，首饰，文具 |
| 滞后一期变量 | 茶叶价格，淡水鱼价格，电风扇什么牌子好，红薯价格，鸡蛋价格，家具品牌，男鞋，女鞋，巧克力品牌，微波炉哪个牌子好，休闲裤，羊肉价格，保健品，床上用品，电视剧排行榜，护肤品排行榜，快递查询，美容院，汽车之家，汽油价格，手表品牌排行榜，手机排行榜，水电费，文具，中成药，中药材 |
| 滞后二期变量 | 白酒，裤子，连衣裙，面包店加盟，葡萄酒品牌，水果，酸奶，童装，香烟价格表，保健品，笔记本电脑排名，电视剧排行榜，房租，家用音响，健身房，旅游，毛毯批发，室内装饰，手机排行榜，网游排行榜，文具，液化石油气价格，液晶电视机价格 |

进一步地，分别建立SCPI和RCPI的实时预测模型，筛选出来的变量个数要远远少于全国CPI模型。具体筛选出来的变量详见表3。

表3城镇CPI和农村CPI模型筛选出来的变量

|  |  |  |
| --- | --- | --- |
|  | 城镇CPI | 农村CPI |
| 当期变量 | T恤、白糖价格、冰箱哪个牌子好、茶叶价格、鸡肉价格、食盐价格、猪肉价格、快递查询、汽油价格、手机、数码相机哪个牌子好、医疗器械、装修材料 | T恤、白糖价格、冰箱哪个牌子好、茶叶价格、鸡肉价格、食盐价格、猪肉价格、快递查询、汽油价格、手机、数码相机哪个牌子好、医疗器械、装修材料 |
| 滞后一期变量 | 冰箱哪个牌子好、茶叶价格、电风扇什么牌子好、红薯价格、家具品牌、坚果、女鞋、酸奶、休闲裤、猪肉价格、床上用品、景点门票、手机、首饰、水电费 | 冰箱哪个牌子好、茶叶价格、电风扇什么牌子好、红薯价格、家具品牌、坚果、女鞋、酸奶、休闲裤、猪肉价格、床上用品、景点门票、手机、首饰、水电费 |
| 滞后二期变量 | 大米价格、红薯价格、黄豆价格、鸡蛋价格、空调排行榜、牛奶品牌排行榜、啤酒品牌、笔记本电脑排名、床上用品、中药材 | 大米价格、红薯价格、黄豆价格、鸡蛋价格、空调排行榜、牛奶品牌排行榜、啤酒品牌、笔记本电脑排名、床上用品、中药材 |

从表3可以看出，这两个模型选择出来的变量全部相同。这表明城镇CPI和农村CPI能够被同样的一些消费品的网络搜索行为所影响，包括当期的影响和滞后期的影响，即网民对某一种物品的搜索量的变化会同时对城镇CPI和农村CPI产生显著的影响，说明农村居民和城镇居民所关注的主要消费品近似。然而，同一个变量的估计系数在两个模型中存在细微的差异，农村居民回归系数的绝对值要高于城镇居民回归系数的绝对值，这说明在城镇地区和乡村地区，网络搜索对CPI的影响程度表现不同[[3]](#footnote-5)。

根据上述我们设定的实时预测模型，结合Elastic Net估计方法，分别得出全国CPI、城镇CPI和农村CPI的拟合值，分别用CPIF、SCPIF、RCPIF表示。三个模型的预测情况如下图1、图2和图3所示。从中可以看出，真实的CPI与拟合的CPI走势基本相吻合，模型估计效果较好，进一步表明本文所构建的实时预测模型的预测效果较好。

图1全国CPI的真实值与拟合值

图2城镇CPI的真实值与拟合值

图3农村CPI的真实值与拟合值

**四、小结**

由于CPI的传统计算方法为统计部门指派专人到指定的地点记录当月的各类商品价格，然后，通过加权平均算出最终的CPI，最后在每月上中旬对外公布，这种传统统计CPI计算方法具有时间上的滞后性。本文基于网络搜索数据，结合Elastic Net估计方法，构建的CPI实时预测模型，并进行了实证检验，研究结果显示模型估计效果较好，预测出的CPI与真实的CPI相吻合，从而避免了传统计算CPI方法上的时间滞后性。

**参考文献：**

### [1]  [Hui Zou and Hastie Trevor（2005）. Regularization and Variable Selection via the Elastic Net. Journal of Royal StatisticSociety B.，2005;67(2):301-20](http://xueshu.baidu.com/s?wd=paperuri%3A%2860e00a73a89b2ed8e2ba74f910ace5ef%29&filter=sc_long_sign&sc_ks_para=q%3DZou%20H%2C%20Hastie%20T.%20Regularization%20and%20variable%20selection%20via%20the%20elastic%20net.%20J%20R%20Statist%20Soc%20B.%202005%3B67%282%29%3A301-20&sc_us=15482582928652185397&tn=SE_baiduxueshu_c1gjeupa&ie=utf-8).

**案例使用说明**

**1.教学目的与用途**

在互联网大数据时代蓬勃发展的背景下，指导学生学习如何运用网络搜索数据，通过学习Elastic-net估计方法进一步加深对大数据分析方法的了解与应用，并在此基础上学会建立有效的预测模型。本案例适用于计量经济学、大数据分析方法等统计课程，不同基础的学生均可选择部分或全部内容来学习。

**2.启发思考题**

（1）传统的CPI计算方法具有滞后性，在当前的大数据时代下，可否有效地利用海量实时数据来实现对CPI的预测？

（2）面对庞杂的高维度数据，如何提取并整合其中的有效成分？

（3）构建预测模型后，如何检测这种实时预测模型的预测效果？

**3.分析思路**

由于传统的CPI计算方法具有时间上的滞后性，而现今大数据时代下，海量实时数据可以即时获取，因此我们可尝试利用网络搜索数据来对CPI进行实时预测。大数据维度高，样本长度短，可对变量进行降维或者选择。本案例选取Elastic Net估计方法，在百度指数中选取最合适的关键词，在此基础上构建实时预测模型并进行实证检验。

**4.理论依据与分析**

本案例应用了Elastic-net估计方法原理。在大数据背景下，数据往往呈现高维度、样本长度短的特点，即大p小n。如果直接对其进行建模，会大大降低估计的精度。此外，解释变量当期和其滞后期往往存在高度相关关系，模型产生的多重共线性会进一步降低预测的准确性。因此，为了实现对该类型的数据进行有效地拟合，通常做法是对变量进行降维或者进行有效的变量选择。

最常用的降维方法就是因子分析。该种方法对高维信息进行降维，提取出若干个具有代表性的因子，但缺点是该因子的经济含义不明确。而Elastic-Net（Zou H and Hastie,2005）是一种有效的变量选择方法，该方法不仅避免了逐步回归进行变量选择时产生的模型稳定性低的特点，而且避免了Lasso方法只能得到过于稀疏模型的缺陷。下面对Elastic-Net估计方法的原理进行详细地介绍。

关于线性回归模型：

其中，，，

，，。

定义：

可得：

理论上，这种基于岭回归和Lasso回归的凸组合可以得到模型的稀疏解，但实践效果并非十分理想。Hui Zou and Hastie Trevor（2005）提出了对该方法进行尺度化 处理的做法，即使用下列式子解决了稀疏解的问题。

Elastic-Net参数估计过程如下：首先给出的参数空间，对每一个，结合LARS-EN算法[[4]](#footnote-6)产生elastic net回归的所有解决路径，然后根据K重交叉验证法，决定相应的另一个参数。最终，再根据均方误差最小的及相应的的值确定模型的。

**5.关键要点**

通过教学，希望学生能够：增强大数据时代下搜索和应用网络数据的能力；掌握运用Elastic-net估计方法原理来处理高维度数据的方法，并能够从百度指数中选取关键词；了解CPI传统预测方法的时滞性，并在此基础上学习构建构建实时预测模型并进行实证检验。

**6.建议课堂计划**

建议采用2-3课时展开教学。首先使学生了解CPI传统预测方法的时滞性，学习按类选取网络搜索关键词的操作方法；其次介绍Elastic-net估计方法的理论与基本思想，增强对高维度大数据的认识和处理能力；最后予以案例分析，引领学生构建有效的预测模型并进行实证检验。

**案例3**

**中国城市化与经济发展的相互关系研究**

**——协整与误差修正模型的应用分析**

**摘 要：**本案例探讨了我国城市化与经济发展之间的相互作用和相互影响。以我国1985-2014年的时间序列数据为样本，利用城市人口比率和人均GDP构建指标，依次进行单整、协整检验并建立误差修正模型，对城市化与经济发展的相互作用做动态分析。结果发现，城市化与经济增长之间存在长期均衡关系。在此基础上，做出小结并给出相关分析。

**关键词：**城市化；经济发展；协整

**一、案例背景**

伴随着经济的高速发展，我国近期开始把关注焦点转向城市特别是超级大城市的发展。众多已有研究显示，城市化与经济增长水平之间呈现显著的正相关关系。经济发展水平越高，城市化水平也越高。钱纳里等在20世纪10年代中期就实证分析了城市化与经济发展水平关系模式，国内的学者也较早地提出了城市化率与人均GDP的对数回归模型，城市化与经济发展水平具有高度的相关性，任何一个国家和地区都在这一程度上遵从这一规律。我国自改革开放几十年来的实践也证明了经济发展激发了城市化水平的提高，城市化又推进了国民经济的发展，二者相互作用，互动发展，构成了改革开放以来中国经济发展的重要动力过程及表现形态之一。

目前我国城市发展已经进入新的发展时期，但是上述相关资料的研究还仅仅局限于改革开放至21世纪初期，而对近十几年来高速发展的城市化与经济增长间的相互关系谈之甚少。大部分学者利用截面数据对二者之间的相关系数进行了测算，却很少利用时序数据就城市化与经济发展之间的交互关系进行分析，且很多研究都没有考虑到人口迁移限制对研究年份的剔除因素。

因此，本案例对1985-2014年城市化与经济增长做了动态计量分析，利用协整检验的方法剖析新时期二者发展的特点，探讨二者间的相互作用机制。

**二、理论模型构建**

引入Engle和Granger提出的协整理论。假定自变量序列为{X1},…, {Xk}，响应变量序列为{Yt}，构造回归模型如下：

假定回归残差序列{}平稳，则称{Yt}与{X1},…, {Xk}之间具有协整关系。

为有效衡量两序列之间是否具有长期均衡关系，使用Engle-Granger两步法进行协整检验。只有具有相同单整阶数的两个变量才可能存在协整关系，故先检验变量的单整性。之后建立两序列间的回归模型，并对残差序列{}进行平稳性检验。

当确定协整模型后，构建误差修正模型来解释序列间的短期波动关系。模型结构如下：

上式中为误差修正系数，表示误差修正项对当期波动的修正力度，则我们可通过ECM模型解释两序列间的短期波动关系。

**三、计量模型设定及检验**

**1.变量和数据**

考虑到数据获取性和可获得性两个因素，综合各类指标，利用国家统计局网站年度数据（ <http://data.stats.gov.cn/easyquery.htm?cn=C01>）所提供的城市人口比率（单位：%）和人均GDP指标分别代表城市化水平U和经济发展水平G，为消除可能存在的异方差，在stata中首先对指标进行对数处理，并分别用lnG和lnU代表处理后的人均GDP值和城市化水平。样本数据见附录1。对样本数据需要说明以下两点：

第一，数据选取起始时间为1985年。因为城市化发展的进程与人口转移不可分离，而人口转移又取决于一国的户籍管理制度。虽然我国的户籍管理制度确立于1951年，但总体而言，在20世纪80年代初之前，我国是限制人口转移的。在1985年之前，农村人口基本不允许转移，故无法根据之前的数据来判别城市化率与经济发展的关系，选取1985年为起始年。

第二，案例数据中采用的人均GDP为名义人均GDP。根据托达罗人口流动模型的原理，只要未来预期城市收入看上去要大于未来预期乡村收入，且农村劳动力获得城市工作机会的可能性较大时，人们从农村向城市的流动就会进行并继续下去，说明与城市化有关的人口流动是与“名义值”有关的，因此我们在数据中采用的人均GDP均为名义人均GDP。

**2.单整检验**

应用Stata统计软件，首先对城市化水平和人均GDP进行单位根检验，看它们的时间序列是否平稳。对序列LnG和LnU的平稳性进行ADF检验，检验结果如表2所示，表中ADF检验值都大于5%的临界值，说明LnG和LnU均为非平稳序列。

表2城市化水平与人均GDP系列ADF检验结果

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 变量 | 检验P值 | ADF检验值 | 临界值（5%） | 临界值（10%） | 结论 |
| **LnU****dLnU****d2LnU****LnG****dLnG****d2LnG** | 0.13830.69870.01170.14160.10200.0010 | -2.978-1.814-3.910-2.967-3.118-4.607 | -3.588-3.592-3.596-3.588-3.592-3.596 | -3.2-3.235-3.238-3.233-3.235-3.238 | 不平稳不平稳平稳不平稳不平稳平稳 |

对两个序列做一阶差分，再进行ADF检验，如上图所示，两序列仍然不平稳；进一步做两阶差分，进行ADF检验，检验值均小于5%和10%的临界值，则此时两个系列的二阶差分都为平稳序列。所以LnU和LnG都具有二阶单整性，即LnU～I(2)，LnG～I(2)。

**3.协整模型及检验**

协整理论主要用于寻找两个或多个非平稳变量之间的均衡关系。如果某两个或多个同阶时间序列变量的某种线性组合可以得到一个平稳的误差序列，则这些非平稳的时间序列之间存在长期均衡关系，即具有协整性。

因为LnU和LnG都具有二阶单整性，故用Engle-Granger检验两变量的协整关系。首先用OLS方法估计变量LnU对LnG的回归方程。则得到估计结果为：

LnU=1.6362+0.2184LnG

 t=(22.92) (27.38)

R2=0.9640 DW=0.1373 F=749.85

据上式设立残差序列**εt** ，并检验其平稳性。ADF检验统计量明显小于显著性水平为5%，10%时的临界值，且p=0.0475<а，则可估计残差序列为平稳序列。说明LnU和LnG之间存在协整关系，则城市化与经济增长之间存在长期动态均衡关系。这种动态均衡关系说明1985-2014年我国经济增长和城市化之间存在长期协调性，并不存在城市化滞后于经济增长的情况。

**4.误差修正模型**

因为协整关系只反映变量之间的长期均衡关系，则为弥补长期静态模型的额不足，可通过短期动态模型反映短期偏离长期均衡的修正机制。建立城市化与经济增长之间的误差修正模型如下：

ΔLnU=0.0375-0.0632ΔLnG-0.0829μt-1+εt

t=(9.18) (-2.32) (-2.65)

由统计结果可知，模型拟合效果很好。LnU误差修正项系数调整分别为-0.0829，这表明城市化水平对均衡关系呈现一种反向修正的机制。由此可推断，误差修正对于城市化水平的短期波动具有显著的调整力度。

**四、小结**

对本案例的学习，需要强调以下几点：

1.本案例根据1985—2014年的时间序列数据，对中国城市化水平与人均GDP进行动态计量分析。因为非平稳序列城市化水平lnU与人均GDP水平lnG经过二阶差分后变得平稳，均为二阶单整，且二者之间的线性组合为平稳序列，存在协整关系。这说明1985—2014年，我国城市化水平与经济增长之间存在长期均衡关系。根据误差修正模型，短期内城市化水平对均衡关系呈现一种反向修正效果。

2.在分析社会经济问题时，大多数经济时间序列都是非平稳的。对其展开分析时，要注意检验序列的平稳性，其中单位根检验法是一种常用的平稳性检验方法。协整是指多个非平稳经济变量的某种线性组合是平稳的。协整分析对于检验变量之间的长期均衡关系非常重要，同时也是区别真实回归与伪回归的有效方法。

3.除了长期均衡关系外，我们还要关注短期波动。任何一组相互协整的时间序列变量都存在误差修正机制。误差修正模型把长期关系和短期动态特征结合在一个模型中，既可以克服传统计量经济模型忽视伪回归的问题，又可以克服建立差分模型忽视水平变量信息的弱点。

**参考文献：**

[1]陈明星，陆大道，刘慧.中国城市化与经济发展水平关系的省际格局[J].地理科学,2010,65.

[2]侯风云. 中国农村劳动力剩余规模估计及外流规模影响因素的实证分析[J]. 中国农村经济，2004，（1）：108-120.

[3]国家统计局.年度数据. <http://data.stats.gov.cn/easyquery.htm?cn=C01>.

[4]赵显洲.我国城市化与经济发展相互关系的动态分析[J].中国软科学, 2006,(9): 116-118.

[5]李金昌，程开明.中国城市化与经济增长的动态计量分析[J].财经研究，2006,(9):19-29

[6]Henderson J V .Urbanization and economic development[J]. Annals of Economics and Finance,2003 ,4 (2) :275～341.

[7] Henderson J V. How urban concentration affects economic growth [R]. The World Bank Policy Research Working Paper, 2005, 2326 ,Washington D. C..

**附录：**

1985-2014年各年的样本数据

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| year | G | U | lnG | lnU |
| 198519861987198819891990199119921993199419951996199719981999200020012002200320042005200620072008200920102011201220132014 | 8609661116137115281654190323243015406650745878645768357199790286709450106001240014259166022033723912259633056736018395444332046629 | 23.7124.5225.3225.8126.2126.4126.9427.4627.9928.5129.0430.4831.9133.3534.7836.2237.6639.0940.5341.7642.9944.3445.8946.9948.3449.9551.2752.5753.754.77 | 6.7569326.8731647.0175067.2232967.3317157.4109527.5511877.7510458.0113558.3104158.5318858.6789728.772928.8298128.8816988.9748729.0676249.153779.2686099.4254519.5651449.7172789.92019710.0821410.1644310.3276810.4917710.5851710.6763710.74998 | 3.1658973.1994893.2315953.2507623.2661413.2737433.2936123.312733.3318473.3502553.3686743.4170713.4629193.5070583.5490423.5896123.6285983.6658673.7020423.7319393.7609673.7918873.8262473.8499353.8782593.9110223.9371063.9621463.9834134.003143 |

数据来源：国家统计局网站年度数据

**案例使用说明**

**1.教学目的与用途**

让学生掌握协整和误差修正模型的基本原理及应用，并熟悉时间序列计量经济学建模的方法和程序，在此基础上能够用统计软件进行有效练习。本案例适用于计量经济学、时间序列分析等统计课程，不同基础的学生均可选择部分或全部内容来学习。

**2.启发思考题**

（1）探讨时间序列数据可以直接进行回归吗？这样的回归究竟是真回归还是伪回归呢？

（2）有些序列自身的变化虽然是非平稳的，但是序列与序列之间却具有非常密切的长期均衡关系。如何体现变量间的长期均衡关系？这种均衡机制的建立需要哪些条件？

（3）现实经济中变量很少处在均衡点上，如何度量变量间的短期的或者是非均衡的关系？

**3.分析思路**

经典回归模型是建立在平稳数据变量基础上的。对于非平稳变量，不能使用经典回归模型，否则会出现虚假回归等诸多问题。而具有协整关系的经济变量间具有长期的稳定关系，是可以使用经典回归模型方法建立回归模型的。因此我们可使用Engle-Granger两步法来检验两变量间的协整关系，估计协整向量，并在此基础上构建误差修正模型，度量变量间的短期波动。

**4.理论依据与分析**

首先用ADF检验来判断时间序列的平稳性，只有具有相同单整阶数的两个变量才可能存在协整关系，故先检验变量的单整性。为有效衡量两序列之间是否具有长期均衡关系，随后引入Engle-Granger两步法进行协整检验。建立两序列间的回归模型，并对残差序列{}进行平稳性检验。

假定自变量序列为{X1},…, {Xk}，响应变量序列为{Yt}，构造回归模型如下：

假定回归残差序列{}平稳，则称{Yt}与{X1},…, {Xk}之间具有协整关系。

当确定协整模型后，构建误差修正模型来解释序列间的短期波动关系。误差修正模型是一种具有特定形式的计量经济学模型，它常常作为协整模型的补充模型出现。结构如下：

上式中为误差修正系数，表示误差修正项对当期波动的修正力度，则我们可通过ECM模型解释两序列间的短期波动关系。

**5.关键要点**

通过教学，要求学生能够：理解时间序列变量非平稳性造成的“伪回归”和单位根过程；掌握时间序列平稳性的基本检验方法；能够通过协整分析来检验变量间的长期均衡关系，并构建误差修正模型将长期关系和短期动态特征结合在一个模型中；掌握用统计软件进行时间序列计量经济学建模的方法和程序。

**6.建议课堂计划**

建议从理论讲解和实证操作两部分来完成教学。首先让学生从理论上了解时间序列平稳性、协整和误差修正模型的相关概念和重要性，讲解各种检验方法。在此基础之上，针对相关案例，有目标地带领学生应用统计软件完成时间序列数据的平稳性检验、单位根检验和协整检验。

**案例4**

**我国国内生产总值GDP对货币供应量M2的影响**

——时间序列模型的应用分析

**摘 要：**长期以来，货币供应量与总产出、价格之间的相互关系一直是货币金融学的热点研究领域；货币供应量作为货币政策的中介目标，在货币政策体系中占有重要地位，研究货币政策问题与经济活动的关系，更与一个国家的实际经济情况相联系，不同时期我国货币供应量对经济总量均有不同程度的影响，有大量文献研究表明，货币供应量M2对国内生产总值GDP存在正相关关系，为研究GDP如何影响M2的，本案例选取了2000-2015年我国国内生产总值GDP和货币供应量M2的数据，运用Stata软件，以货币供应量M2为因变量，GDP为自变量，建立一元线性回归模型，并对其自相关问题进行检验和修正，自相关检验方法主要有图示法、DW检验和LM检验，修正方法主要用广义最小二乘法。

**关键词：**自相关；D-W检验法；广义最小二乘法

**一、案例背景**

21世纪以来，中国国内生产总值在全球排名稳步上升，2016年GDP达到74.4万亿元，经济总量全球排名第二，2016年国内生产总值的增长率稳定在6.7%；M2是指广义货币供应量，并且M2=M1+城乡居民储蓄存款+企业存款中具有定期性质的存款+信托类存款+其他存款，M2不仅反映了现实购买力，也反映了潜在购买力，所以在研究货币供应量对经济总量的影响时，选取M2变量。因货币供给与经济总量影响的相互作用，所以在运用货币政策对宏观经济进行调控，其关键之一就是要控制货币供给量，而货币政策调控后的GDP对M2也会产生影响，本案例用GDP与M2的数据建立一元线性回归模型进行实证分析。

**二、理论背景**

曾有文献研究表明，M0、M1、M2利率和GDP之前存在长期稳定的协整关系，且M2与GDP之间存在显著的正相关关系，在产出（增长率）与货币供应量（增长率）的Granger因果关系分析中，从短期来看，估计出的误差修正模型M2对GDP的修正的影响程度要大于M1和M0，所以最后选择M2，分析其与GDP之间的影响。

在本案例分析中，采用一元线性回归模型：

 （3-1）

其中表示货币供应量M2，表示我国国内生产总值GDP，为随机误差项，为待估参数。

自相关的表示形式，一阶自回归模型：

，1（t=1,2,...,n） （3-2）

自相关又称为序列相关。

**三、经济应用案例**

**（一）数据和变量**

案例数据来源于2016年中国统计年鉴，选取了2000年到2015年的GDP和货币供应量M2两个变量，数据如表3-1所示：

表3-1 2000-2015年我国GDP和货币供应量M2 单位：亿元



资料来源：《中国统计年鉴（2016）》

**（二）模型估计**

**1.散点图**

如图3-1所示，解释变量GDP和被解释变量M2沿着一条直线趋势，初步判断变量之间存在正相关的线性关系，接着进行回归模型估计。



图3-1 散点图

以GDP为自变量，货币供应量M2为因变量，运用Eviews软件对一元线性回归模型进行估计，采用OLS法，求得回归方程，参数估计结果如表3-2所示，回归方程：

se： （21221.02） （0.055）

DW=0.665，=0.990

表3-2 回归估计结果



**2.自相关的检验**

在计量经济模型中，自相关现象是经常存在的，主要出现在时间序列模型中，在经济数据中，常见的是正自相关现象。本案例中运用图示法、DW检验和LM检验对已估计的回归模型进行自相关检验。

**2.1图示法**

**（1）残差时序图**

在估计结果的菜单栏点击View/Actual,Fitted,Residual/Residual Graph，得到残差时序图（图3-2），横轴为时间，纵轴为OLS残差。



图3-2 残差时序图

当＞0时，随后的残差也都大于0，或＜0，随后的残差也小于0，则准在正自相关，而负相关则是两个相继的残差具有正负相反的倾向；无自相关则是残差值随机分布在时间轴两侧。从图3-2可以初步判断随机误差项之间存在正自相关。

**（2）残差（，）散点图**

在命令栏内输入 genr e=resid，然后enter，得到序列e

再在命令栏内输入scat e e(-1)，enter得到残差散点图（图3-3），横轴为，纵轴为。



图3-3 残差散点图

若是残差散点图基本落在第Ⅰ、Ⅲ象限，表示存在正自相关；若大部分落在第Ⅱ、Ⅳ象限，表示存在负自相关性，若随机落在四个象限，则不存在自相关性。图中的散点图大致落在了第Ⅰ、Ⅲ象限，表示残差序列存在正自相关。

**2.2 DW检验**

在线性回归方程的OLS估计中得到DW值（见表3-2）。在本案例中，n=16，k=1，在0.05显著性水平下，由DW检验表查得,1.37,DW=0.665＜，则存在正自相关。

**2.3 LM检验**

在估计窗口选择View/Residual Diagnostics/Serial Correlation LM Test，在弹出的对话框里输入1，得到LM检验结果（表3-3），检验结果的P值均小于显著性水平0.05，拒绝原假设，序列存在一阶正自相关。

表3-3 LM检验结果



**（三）自相关的修正**

自相关虽然不影响OLS估计量的线性和无偏性，但会使OLS估计量失去有效性，会影响参数显著性检验，还有模型用于预测失效。

广义最小二乘法（generalized least squares，GLS）是消除模型一阶线性自相关，恢复估计量最小方差的性质的一种方法。广义最小二乘法将模型转化成满足经典回归假定的模型，再对转化后的模型应用普通最小二乘法，运用估计自相关系数=0.67，进行广义差分变换，得到广义差分模型：。

在命令栏输入

genr NGDP2=GDP-0.67\*GDP(-1)

genr NM2=M2-0.67\*M2(-1)

以NGDP2和NM2为样本用OLS进行估计，估计结果见表3-4，得到新回归方程：

se： （19181.68） （0.120）

DW=1.332，=0.960

表3-4 广义差分估计结果



差分后数据少一个，n=15，k=1，显著性水平0.10下，DW检验表查得，1.07，DW=1.3322.93，表示处理后的模型无明显的自相关。则原参数估计值应为：

==-127304

相应的标准差：==19181.68/(1-0.67)=57710.7

=0.12

**四、结论**

对本案例主要学习自相关的检验和修正方法：

（1）自相关检验方法有图示法、DW检验、LM检验等，图示法比较直观，但主观性很强，在图形比较复杂时，很难做出准确的判断，注意DW检验只能用于检验一阶自相关。

（2）自相关修正方法有一阶差分和广义差分法，一阶差分法只适用于模型存在完全一阶正自相关的情况，广义差分法适用于非完全一阶正自相关的情况。存在高阶自相关时，采用杜宾两步法和CO迭代法。

**参考文献**

[1]胡荣才，邓淇中，陈旭辉.计量经济学实验.北京：中国统计出版社，2011.5

[2]李宝仁.计量经济学（第2版）.北京：机械工业出版社，2014.11

[3]朱小檬，孙爱田.我国GDP与广义货币供应量M2关系的实证分析[J].商业时代,2011,(28):43-44.

[4]戴建军.我国货币供应与经济增长关系的理论和实证研究[D].湖南大学,2008.

**案例使用说明**

**一、教学目的与用途**：

适用于本科生计量经济学课程和时间序列分析课程，目的是让学生学习、使用EViews软件在回归模型中对自相关的检验及修正，并熟悉计量经济分析建模的方法和程序。

**二、启发思考题：**

（1）自相关概念及产生的主要原因。

（2）D-W检验与LM检验的差别。

（3）自相关的修正方法中另外两种怎么做。

**三、分析思路：**

计量经济案例一般会结合宏观经济背景进行研究，然后运用计量经济学知识进行建模分析，步骤一般是先收集和整理数据，分析时先是进行简单的描述统计分析，然后建立模型，参数估计，参数检验和残差检验，并对出现如自相关的问题进行处理。得到最后的模型，并对模型的参数进行解释。

**四、理论依据与分析：**

（一）计量经济学的相关理论：

1.自相关概念

自相关又称序列相关，一般指同一随机变量在不同时间和空间（或空间）上的取值之间存在相关关系。这里讨论的是回归模型中随机干扰项的逐次值之间是否相关。

若经典回归的基本假定4不成立，即

0（） (1-1)

则称存在自相关。

2.自相关产生原因

（1）被解释变量的自相关

（2）略去了自相关的解释变量

（3）错误地确定确定回归模型的数学形式

（4）随机干扰项自身的特性决定

（5）数据整理的影响

3.自相关检验

（1）图示法

残差时序图的判断方法：当＞0时，随后的残差也都大于0，或＜0，随后的残差也小于0，则准在正自相关，而负相关则是两个相继的残差具有正负相反的倾向；无自相关则是残差值随机分布在时间轴两侧。

残差（，）散点图判断方法：若是残差散点图基本落在第Ⅰ、Ⅲ象限，表示存在正自相关；若大部分落在第Ⅱ、Ⅳ象限，表示存在负自相关性，若随机落在四个象限，则不存在自相关性。

（2）DW检验：

杜宾-沃森（Durbin-Watson）检验法，简称DW检验法，其基本思想是，对于一阶自相关形式，，显然，当，不存在自相关，当，存在一阶自相关。DW检验法是通过构造统计量

 (1-2)

其中 来建立d与的关系，从而判断随机误差的自相关性。

 (1-3)

d

DW检验临界区域示意图

DW

应用DW检验时，应注意以下几个问题：

①DW检验只适用于检验一阶自相关

②样本容量n不能小于15，否则将难以得出明确的检验结论。

③这种检验法不适用于自回归模型。

（3）LM检验

如果把误差项带入多元回归方程，可得到：

 (1-4)

式中的滞后项无法估观测，可用最小二乘估计的残差来代替。然后对系数进行t检验，这一检验方法称为拉格朗日乘数（Lagrange multiplier，LM）检验。其特点如下：

①不需要假设服从正态分布

②对的t检验可使用标准正态分布的临界值，不存在无法判断的区间。

③在使用计量软件进行检验时，只需在回归方程中加入残差的滞后项，对其系数进行检验，操作十分简单。

4.自相关修正

广义最小二乘法（generalized least squares，GLS）是消除模型一阶线性自相关，恢复估计量最小方差的性质的一种方法。广义最小二乘法将模型转化成满足经典回归假定的模型，再对转化后的模型应用普通最小二乘法。

模型 （1-5）

存在一阶线性自相关：

，1（t=1,2,...,n）。 （1-6）

广义差分交换：

 （1-7）

进行广义差分变换后再用普通最小二乘法进行估计，得到广义差分模型：

 （1-8）

最后计算原模型的参数估计值：

 （1-9）

= （1-10）

**五、关键要点：**

案例教学中的关键知识点：计量经济学模型的建模步骤；自相关的概念及产生的原因以及后果；自相关的检验以及修正。

能力点：学生的软件操作能力，能运用Eviews软件进行相关的分析。

**六、建议课堂计划：**

先对学生讲授自相关的理论知识，在熟悉理论知识的前提下对案例进行学习，因为Eviews在计量经济学和时间序列分析中都有涉及到，建议在教学中安排一次课来，进行案例学习和软件的实际操作。并在学习案例时对案例中用到宏观经济的背景进行讲解。

**案例5**

**我国国内旅游收入的影响因素分析**

**——多元线性回归模型的应用分析**

**摘 要：**近年来，中国旅游业一直保持高速发展，旅游业作为国民经济新的增长点，在整个社会经济发展中的作用日益展现，所以定量地分析影响国内旅游市场发展的主要因素是具有现实意义的。当在构建关于旅游收入的回归模型中使用两个或两个以上的解释变量时，这些自变量往往会提供多余的信息，回归模型中就存在多重共线性。在实际经济问题中，所使用的自变量之间存在相关是很常见的，但是在回归分析中存在严重的共线性时就会产生问题，因此，必须对多重共线性进行检验以及处理。本案例以构建我国国内旅游收入影响因素的计量经济模型为背景，描述了如何在多元线性回归模型构建过程中利用统计软件判断、检验以及处理多重共线性。

**关键词**：旅游收入，多元回归，多重共线性，计量模型

**一、案例背景**

旅游业是战略性产业，资源消耗低，带动系数大，就业机会多，综合效益好。旅游业的发展对优化我国经济产业结构、增加我国在外汇方面的收入以及促进国际收支平衡具有深远的影响。虽然旅游业在我国的发展起步比较晚，但改革开放三十多年来，国内旅游业呈现不断发展的趋势，特别是进入90年代后，已初步形成了“大旅游、大产业、大发展”的基本格局。随着人民富裕程度的提高，闲暇时间和可支配收入的增多，不断推动旅游需求以较高的速度增长，尤其是国内旅游市场，其需求增长十分迅猛。据国家旅游局数据中心统计，2016年旅游总收入预计达到4.69万亿元，同比增长13.6%。国内游、入境游、出境游分别实现44.4亿、1.38亿、1.22亿人次。国内旅游收入达到3.9万亿元，是我国旅游业收入的主要来源。旅游业作为国民经济新的增长点，在整个社会经济发展中的作用日益显现。

有相关学者指出，未来旅游业在我国经济发展中的比重也将越来越多，旅游收入未来将直接反映某一旅游目的地国家和地区旅游经济的运行状况，是衡量旅游经济活动及其效果的一个不可缺少的综合性指标，也是某一国家或者地区经济发达与否的重要指标之一。当前我国正处于工业化、城镇化快速发展时期，日益增长的大众化、多样化消费需求为旅游业发展提供了新的机遇。因此要充分发挥旅游业在保增长、扩内需、调结构等方便的积极作用，所以定量地分析旅游业的影响因素，不仅有利于规划中国未来旅游产业的发展，并且对促进我国国内经济发展也有一定的意义。

本案例主要通过计量经济学模型来分析我国旅游业发展过程中的影响因素。

**二、理论模型构建**

在生产和经营活动中，经常要对变量之间的关系进行分析。回归分析是考察变量之间的数量关系，并通过一定的数学表达式将这种关系描述出来，进而确定一个或几个变量（自变量）的变化对另一个特定变量（因变量）的影响程度的一种分析方法。一元线性回归是一个主要影响因素作为自变量来解释因变量的变化，在现实问题研究中，因变量的变化往往受几个重要因素的影响，此时就需要用两个或两个以上的影响因素作为自变量来解释因变量的变化，这就是多元回归亦称多重回归。当多个自变量与因变量之间是线性关系时，所进行的回归分析就是多元线性回归分析，所建立的模型就是多元线性回归模型。

多元线性回归的理论模型为：

 （1.1）

其中,为模型的回归参数，k为自变量的个数

**三、计量模型构建**

**3.1 变量和数据选取**

旅游业是一个关联性、依赖性较强的产业，影响旅游业的指标因素也是多方面的，而各个因素的影响也不尽相同。国内旅游收入是多种因素共同影响的函数，包括国内旅游人数、旅游行业规模、居民可支配收入、游客人均花费及包括交通条件、服务设施、接待机构设施在内的旅游基础设施等。

经分析，影响我国国内旅游市场收入的主要因素，除了国内旅游人数和旅游支出以外，还可能与相关基础设施有关。为此，考虑的影响因素主要有国内旅游人数X2，城镇居民人均旅游支出X3，农村居民人均旅游支出X4，并以公路里程X5和铁路里程X6作为相关基础设施的代表。

为估计模型参数，从中国国家统计局网站收集国内旅游收入以及相关影响因素的1994—2015年的统计数据，如附表所示。

**3.2 回归模型设定**

根据选取的解释变量设定了如下形式的计量经济模型：

 （1.2）其中 ： ——第t年国内旅游收入 （亿元）

——国内旅游人数 （万人）

——城镇居民人均旅游支出 （元）

——农村居民人均旅游支出 （元）

——公路里程 （万公里）

——铁路里程 （万公里）

可以看到所选取的变量单位和数量级之间存在较大差异，并且大多数经济数据都是呈现右偏分布的形态并不满足正态分布的假定，鉴于这种状况，我们对所有变量取对数后再建立多元线性回归模型。取对数并不会改变变量之间原有的相关关系和数据的性质，只是压缩变量的尺度，并且可以一定程度的修正数据的右偏形态，使其更接近于正态分布。所以对所有变量取自然对数后设定了如下形式的多元线性回归模型：

 （1.3）

**3.3 参数估计以及检验结果**

利用Stata软件，输入Y、C、X2、X3、X4、X5、X6等数据，采用这些数据依照设定的模型形式进行OLS回归，结果如表：



表1 模型回归结果

由此可见，R2=0.9988，调整后的R2=0.9984，可决系数很高，F检验值2560.12，远大于临界值，说明模型的线性关系显著。但是当时，lnX5、lnX6系数的t检验不显著，判断可能存在严重的多重共线性。下面就对多重共线性进行检验。

1.综合判断法

由模型的估计结果可以看出，拟合优度很高，说明模型对样本的拟合很好；F=2560.12，相应的P=0.0000，说明回归方程的线性关系显著，即各解释变量联合起来确实对被解释变量“国内旅游收入”有显著影响；而在给定显著性水平下，不仅lnX5、lnX6系数的t检验不显著，而且X5（公路里程）系数的符号与现实的经济意义相反。所以，很可能存在严重的多重共线性。

2.相关系数检验

计算各自变量间的相关系数，选择X2、X3、X4、X5、X6数据，得到如下的相关系数矩阵以及相应的sig值：



表2 相关系数表

由相关系数矩阵可以看出：各自变量之间的相关系数较高，均超过了0.8，呈现高度正相关，证实确实存在较为严重多重共线性。

3.VIF方差膨胀因子法检验

定义解释变量Xk的方差膨胀因子也称方差扩大因子（Variance Inflation Factor, VIF）为

  （1.4）

方差膨胀因子VIF的值越大，Xk的多重共线性问题就越严重。一般来说，VIF的值大于10，就认为存在较为严重的多重共线性。

运用Stata软件，可以得到各解释变量的方差膨胀因子的值。结果如下：



表3 解释变量方差膨胀因子表

可以看出，方差膨胀因子VIF的值均超过了10，存在严重的多重共线性。

结合综合判断法、简单相关系数检验法和VIF方差膨胀因子法分析的结果可以知道，本案例的回归变量间确实存在较为严重的多重共线性。

**3.4 多重共线性的解决**

下面我们将采用逐步回归法来解决多重共线性问题。

首先，分别作lnY对lnX2、lnX3、lnX4、lnX5、lnX6的一元回归，结果整理如下所示：

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 变量 | lnX2 | lnX3 | lnX4 | lnX5 | lnX6 |
| 参数估计值 | 1.5869 | 4.1240 | 1.4738 | 1.8074 | 5.2406 |
| t 统计量 | 30.05 | 12.21 | 10.66 | 12.14 | 21.49 |
|  | 0.9783 | 0.8816 | 0.8504 | 0.8805 | 0.9585 |

 表4 一元回归结果表

按的大小排序为：lnX2、lnX6、lnX3、lnX5、lnX4。以lnX2为基础，顺次加入其他变量做逐步回归。结果如下所示：

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | lnX2 | lnX3 | lnX4 | lnX5 | lnX6 |  |
| lnX2，lnX3 | 1.2113(15.58) | 1.1444(5.38) |  |  |  | 0.9914 |
| lnX2，lnX4 | 1.2095(34.90) |  | 0.4369(12.66) |  |  | 0.9977 |
| lnX2，lnX5 | 1.5654(9.27) |  |  | 0.0273(0.13) |  | 0.9784 |
| lnX2，lnX6 | 1.1667(4.85) |  |  |  | 1.4336(1.79) | 0.9814 |

表5加入两个解释变量后的回归结果

比较发现，新加入lnX4的回归模型的R2=0.9977，改进最大，而且各参数的t检验显著，选择保留lnX4，再加入其他新变量逐步回归，结果如下所示：

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | lnX2 | lnX3 | lnX4 | lnX5 | lnX6 |  |
| lnX2,lnX4,lnX3 | 1.1641(31.91) | 0.3288(2.37) | 0.3645(8.39) |  |  | 0.9983 |
| lnX2,lnX4,lnX5 | 1.2214(19.40) |  | 0.4373(12.33) | -0.0156(-0.23) |  | 0.9977 |
| lnX2,lnX4，lnX6 | 1.1124(13.33) |  | 0.4229(11.85) |  | 0.3723(1.28) | 0.9979 |

表6 加入三个解释变量后的回归结果

在lnX2，lnX4的基础上加入lnX3后的有所改善，且各参数的t检验显著。而加入lnX5时，没有明显变化，且lnX5的系数的t检验变得不显著并且系数符号不符合经济意义。加入X6后，有所提升，但是X6的参数的t检验不显著，且的提升没有加入lnX3的效果好。所以保留lnX3，再加入其他新变量继续逐步回归，结果如下所示：

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | lnX2 | lnX3 | lnX4 | lnX5 | lnX6 |  |
| lnX2,lnX4,lnX3,lnX5 | 1.2439(24.38) | 0.5060(3.30) | 0.3292(7.58) | -0.1362(-2.08) |  | 0.9986 |
| lnX2,lnX4,lnX3,lnX6 | 0.9977(13.08) | 0.4142(3.23) | 0.3235(7.67) |  | 0.5928(2.41) | 0.9987 |

表7 加入四个解释变量后的回归结果

当加入lnX5后，有所增加，但其参数的t检验在0.05的显著性水平下不显著，并且系数符号为负号，与预期的相反。加入lnX6后，也有所改善，并且所有解释变量的系数均通过t检验，符号也合理，所以lnX6可以加入方程中，X5需要剔除。

最后剔除变量lnX5，修正严重多重共线性影响后的回归结果为：



表8 逐步回归法确定的最终回归结果

因此，最终国内旅游收入的回归方程可以写为：

 （1.5）

 **t** = (-16.37) (13.08) (3.23) (7.67) (2.41)

 = 0.9987  = 0.9984  **F**=3254.72 **P**=0.0000

这表明，在其他因素不变的情况下，国内旅游人数、城镇居民人均旅游支出、农村居民人均旅游支出、铁路里程分别变动1%，国内旅游收入就会同方向分别变动0.9977%、0.4124%、0.3235%和0.5928%，其中国内旅游人数对国内旅游收入的影响最大。

**四、案例小结**

用逐步回归筛选变量的方法从有共线性的变量组中筛选出对因变量影响显著的若干个变量来建立最优回归式。不仅克服了共线性问题，且使得回归式简化。最终构建的模型的经济意义也是合理的。

对本案例的学习，还需要强调以下几点：

1、如果在解释变量中，某一解释变量可由其他解释变量线性表出，则存在“严重多重共线性”。严重多重共线性现实中较少出现，在实践中更为常见的是近似（非严格）的多重共线性，即某个解释变量对其余解释变量的辅助回归具有很高的可决系数但并不是完全共线，在这种情况下，OLS估计量仍然是无偏的，但是方差会变大，参数的估计将变得不准确，对模型的统计解释也很可能存在错误，所以我们要对多重共线性进行处理。

2、在实际问题中，我们更关心的是多重共线性的严重程度而不是是否存在多重共线性，解释变量之间的相关性普遍存在，大多数经济模型的构建中都会存在多重共线性，但只要对模型设定和经济意义解释没有显著影响，一般不对其做处理，只有当多重共线性严重到违背我们的假设与理论，影响模型现实意义的解释时，我们才会采取方法减轻多重共线性的影响。

3、常用的多重共线性的判别方法有三种：综合判断法、相关系数法和方差膨胀因子法。

4、如果存在严重的多重共线性，可以采取增大样本容量、逐步回归对变量进行筛选、人为剔除导致严重共线性且对模型不重要的变量，岭回归或对模型设定进行修改等等方法解决。

5、逐步回归法虽然实现了减轻多重共线性的目的，但在某些模型的设定中，一些重要的影响因素可能会一并从模型中被剔除出去，虽然本案例中没有出现这种情况，但是在其他经济计量模型中使用逐步回归法时需要注意。

6、多重共线性带来的主要麻烦是对单个回归系数的解释和检验，但是在求被解释变量的置信区间和预测区间时一般不会受其影响。如果仅仅是为了估计或者预测，可以不处理多重共线性。

**附录：**

1994——2015年我国的国内旅游收入以及相关影响因素数据

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 年份 | 国内旅游收入Y/亿元 | 国内旅游人数X2/万人次 | 城镇居民人均旅游支出X3/元 | 农村居民人均旅游支出X4/元 | 公路里程X5/万km | 铁路里程X6/万km |
| 1994 | 1023.5 | 52400 | 414.7 | 54.9 | 111.78 | 5.9 |
| 1995 | 1375.7 | 62900 | 464 | 61.5 | 115.7 | 6.24 |
| 1996 | 1638.4 | 64000 | 534.1 | 70.5 | 118.58 | 6.49 |
| 1997 | 2112.7 | 64400 | 599.8 | 145.7 | 122.64 | 6.6 |
| 1998 | 2391.2 | 69500 | 607 | 197 | 127.85 | 6.64 |
| 1999 | 2831.9 | 71900 | 614.8 | 249.5 | 135.17 | 6.74 |
| 2000 | 3175.5 | 74400 | 678.6 | 226.6 | 167.98 | 6.87 |
| 2001 | 3522.4 | 78400 | 708.3 | 212.7 | 169.8 | 7.01 |
| 2002 | 3878.4 | 87000 | 739.7 | 209.1 | 176.52 | 7.19 |
| 2003 | 3442.3 | 87800 | 684.9 | 200 | 180.98 | 7.3 |
| 2004 | 4710.7 | 110200 | 731.8 | 210.2 | 187.07 | 7.44 |
| 2005 | 5285.9 | 121200 | 737.1 | 227.6 | 334.52 | 7.54 |
| 2006 | 6229.7 | 139400 | 766.4 | 221.9 | 345.7 | 7.71 |
| 2007 | 7770.6 | 161000 | 906.9 | 222.5 | 358.37 | 7.8 |
| 2008 | 8749.3 | 171200 | 849.4 | 275.3 | 373.02 | 7.97 |
| 2009 | 10183.7 | 190200 | 801.1 | 295.3 | 386.08 | 8.55 |
| 2010 | 12579.8 | 210300 | 883 | 306 | 400.82 | 9.12 |
| 2011 | 19305.4 | 264100 | 877.8 | 471.4 | 410.64 | 9.32 |
| 2012 | 22706.2 | 295700 | 914.5 | 491 | 423.75 | 9.76 |
| 2013 | 26276.1 | 326200 | 946.6 | 518.9 | 435.62 | 10.31 |
| 2014 | 30311.9 | 361100 | 975.4 | 540.2 | 446.39 | 11.18 |
| 2015 | 34195.1 | 400000 | 985.5 | 554.2 | 457.73 | 12.1 |

数据来自：中国国家统计局:[www.stats.gov.cn](http://www.stats.gov.cn)

**参考文献：**

[1].庞皓.计量经济学[M].北京：科学出版社，2007.

[2].吴亮亮.我国国内旅游收入的主要影响因素的计量模型分析[J].经济视野，2016.16.

[3].宫庆硕，冯爱芬，蔡雪瑞.基于多元统计回归的国内旅游收入预测研究[J].农村经济与科技，2017.28.

[4].何红霞，李锴，梁磊.我国国内旅游收入影响因素的实证分析[J].现代物业，2010，02：21-23.

[5].郑媛媛.国内旅游影响因素分析与模型评估[J].成都电子机械高等专科学校学报，2005，02：63-68.

[6].徐勇.中国国内旅游发展及其经济增长关系的计量分析[D].上海：华东师范大学，2012.

[7].袁翊茗，邵欣欣.我国国内旅游收入影响因素的计量分析——基于1994-2010年的数据[J].科技广场，2010，10：97-100.

**案例使用说明**

**一、教学目的与用途**

1.本案例主要适用于计量经济学课程。

2.本案例的教学目的在于通过构建实际经济问题的计量经济模型，加强学生对于回归模型中多重共线性的识别、检验及处理的能力，从旅游业发展的影响因素分析与多元线性回归模型的构建入手，深入理解多重共线性的概念和表现，掌握如何利用Stata软件检验和处理多重共线性问题。

**二、启发思考题**

1.在整个方程线性关系显著的情况下，如果只关心整个方程的预测能力，多重共线性是否一定要处理？

2.如果更关心某几个解释变量对被解释变量的影响，是否一定要处理回归模型中的多重共线性问题？

3.如果在采用逐步回归方法筛选变量来消除多重共线性的过程中，我们最关心的或者经验已知的对被解释变量影响显著的某个解释变量被剔除，我们是否应该继续采纳逐步回归确定的模型进行分析？如果不采纳，怎样消除多重共线性？

4.模型中解释变量相关系数过高一定是问题吗？相关系数的高低是否能决定和控制变量的取舍。

**三、分析思路**

教师可以根据自己的教学目标（目的）来灵活使用本案例。这里提出本案例的分析思路，仅供参考。

1.分析影响旅游业发展的相关因素，并选取解释变量，运用多元回归模型理论设定国内旅游收入影响因素的计量经济模型。

2.运用Stata软件和相关数据对国内旅游收入建立多元线性回归模型，并对回归结果以及系数的显著性进行分析判断是否存在多重共线性。

3.运用Stata软件对多重共线性进行检验，并采取相应的方法解决模型中存在的多重共线性。

4.采用逐步回归法消除多重共线性，得到最终的回归模型，对多重共线性的解决结果进行经济意义解释和进一步的思考。

**四、理论依据与分析**

 1.多元回归模型理论

多元回归分析就是用两个或两个以上的影响因素作为自变量来解释因变量的变化的一种分析方法。而旅游业是商业、交通业、运输业等的结合，是一种关联性较强的综合产业，影响旅游业发展的因素有很多，所以对其建立多元回归模型是合理并且可行的。

 2.多重共线性的成因

当回归模型中两个或两个以上的解释变量彼此相关时，则称回归模型中存在多重共线性。共线性产生的原因是多方面的。一是没有足够多的样本数据；二是选取的解释变量之间客观上就有共线性的关系；三是模型中包含滞后变量；还有可能由其他因素导致，如数据采集所用的方法、选取的是截面数据、模型设定等等。

 3.多重共线性的危害

在实际问题中，所选取的解释变量之间存在相关关系是一件很平常的事，但是共线性一直受到人们的关注，这是因为严重的共线性会对回归分析造成危害和问题。如果出现完全共线性，则回归模型中的系数将无法确定。而对于高度共线性的模型，回归系数尽管可以估计，但可能会招致以下后果：一是方差和协方差都将扩大，估计参数精度将大大降低，从而使某些系数的t统计值不显著；二是回归参数估计量不稳定，加上或减少一个变量会使参数估计发生较大变化；另外，一个观测值稍有变化也会对估计量和标准误的值产生很大的影响。最为常见的后果是估计量的符号会发生同客观实际不一致的情况。

4.多重共线性的处理

目前有十余种处理共线性问题的方法，这里只对几种常用的方法进行简单的讨论。

（1）剔除一些不重要的变量。当面临严重的共线性是，一种最简单的方法就是去掉那些不太重要的变量。如何判断某个变量是否重要是此方法的关键，因为从模型中删除一个变量可能导致模型设定的偏误。因此，变量的选择可以采用向前选择、向后剔除、逐步回归和最优子集等方法，还可以结合实际经验或者相关理论来考虑。

（2）增大样本量。在允许的情况下可以增大样本数据的个数来消除多重共线性，这样可以避免随意剔除变量对模型造成的影响，但实际问题中，样本量的扩大往往不是非常容易的。

（3）主成分估计。这是一种处理严重共线性的有偏估计方法。它利用主成分分析对自变量进行变量综合，然后将综合后的新变量作为解释变量，再进行OLS估计，得到主成分估计。然而，主成分估计的结果必然受到重叠信息的影响，且主成分的实际含义也不明确。

（4）岭回归。岭回归也是有偏估计方法。当出现严重共线性是，岭估计往往比OLS估计量更稳定，以及有更小的协方差矩阵。岭估计的最大困难是最有K值的选择。

目前处理多重共线性的方法没有一种具有突出的优势，所以还是要根据研究目的来选择适合的方法处理多重共线性。

**五、关键要点**

1.在实际问题中，共线性是普遍存在的。目前处理共线性的方法都是把共线性的部分分离出来，使它们对因变量的解释只出现一次，但在代数处理上无法对得到的数据进行这样的分离，所以共线性是无法完全解决的。

2.存在共线性不代表就一定会对回归分析产生严重的危害，并非严重的共线性就一定需要处理，也并非不严重的共线性就不会对分析产生严重的后果。所以对于共线性是否需要处理是由研究目的和研究要求来确定的，如果我们关心的是整体模型的预测估计能力，那么只要模型的线性关系是显著的，就无需考虑多重共线性。

3.逐步回归选择变量可能会将我们所关心的变量或者对解释因变量变化情况具有关键作用的自变量从模型中剔除，所以在实际问题中使用逐步回归法需要格外注意这个问题。

**六、建议课堂计划**

本案例可以作为专门的案例讨论课来进行。如下是按照时间进度提供的课堂计划建议，仅供参考。整个案例课的课堂时间控制在90分钟。

1.课前计划：

请学生在课前阅读案例并查阅与旅游业发展以及影响因素相关的资料，提出启发思考题，进行小组讨论，并对问题进行初步思考。

2.课中计划：

（1）简要的课堂前言和导引。老师说明本次案例分析课的主题，研讨形式与目标确定（5分钟）。

（2）分组讨论。以5人为一小组，组长引导组员对思考题进行研讨，并思考如何构建旅游收入的计量经济模型（35分钟）。

（3）小组发言。请每小组代表发言（每组5分钟左右，控制在35分钟）。

（4）总结提升。老师根据同学发言，再详细讲解多重共线性的相关知识，加深学生对这一概念的掌握，以及在实际建模过程中的检验和处理（15分钟）。

3.课后计划：

请学生运用所学的理论、方法与工具，采用PPT报告的形式选取自己感兴趣的问题和方向，寻找相关数据建立多元回归模型，并对多重共线性进行判断和检验，若是模型存在共线性，则采用合适的方法进行处理。

**案例6**

**我国各地区国内生产总值对最终消费支出的影响分析**

**摘 要：**随着我国经济迅速发展，我国的国内生产总值不断提高，在2016年达到74.4万亿元，经济总量跃居世界第二。根据经济学的相关理论，收入是影响消费水平的最主要因素。国内生产总值作为我国总收入水平的代表，影响着我国的最终消费水平。本案例用国内生产总值代表总收入，用居民最终消费表示总消费，建立计量经济模型。选取全国31个省、市、自治区2015年国内生产总值（GDP）和最终消费支出的截面数据，建立了一个简单的一元线性回归模型，利用Stata软件进行分析，检验了模型存在的异方差问题并对模型进行了修正。

**关键词**：计量经济模型；一元线性回归；Stata软件

**一、案例背景**

国内生产总值（Gross Domestic Product，简称GDP）是指在一定时期内（通常为一年），一个国家或地区经济中所生产的全部最终产品和劳务的价值，常被公认为衡量国家经济发展状况的最佳指标。一般来说，国内生产总值由四部分组成，其中包括居民消费(C)、投资(I)、政府购买支出(G)和净出口(NX)，即GDP=C+I+G+NX。将居民消费和政府购买支出之和作为最终消费支出，则最终消费支出影响国内生产总值，根据消费与收入理论，收入是影响消费行为的重要因素，因此国内生产总值作为总收入也影响着一国最终消费水平，本案例即用简单线性模型来衡量国内生产总值对最终消费支出的影响。

**二、理论模型构建**

消费量由什么决定？在现实生活中，影响各个家户消费的因素很多，如收入水平、商品的价格水平、利率水平、收入分配状况、消费者偏好、家庭财产状况、消费者信贷状况、社会保障制度、风俗习惯等等。凯恩斯认为，这些因素中具有决定意义的是收入，关于收入和消费的关系存在一条基本心理规律：随着收入增加，消费也会增加，但消费的增加不如收入增加多（因此边际消费倾向大于0小于1），消费和收入的这种关系称作消费函数。若消费与收入之间存在线性关系，则消费函数可表示为，其中X表示收入，Y表示消费。为边际消费倾向。

以上为家户的消费函数，宏观经济分析的是整个社会的消费函数，即本案例中总消费与总收入的关系。社会消费函数受到国民收入的分配、政府税收政策、未分配利润在利润中所占比例等因素的影响，因此社会消费曲线并非家户消费曲线的简单加总，但在考虑了种种限制之后，社会消费曲线的基本形状仍和家户消费曲线有很大的相似之处。因此，本文中关于总收入和总消费的社会消费函数仍采用了家户消费函数的形式。即对于总收入和总消费的线性关系建立一元线性回归模型：，其中Y表示总消费，X表示总收入，表示边际消费倾向。

**三、计量模型建立及分析**

1. **变量和数据**

案例选取了2015年我国各省市自治区的国内生产总值和最终消费支出两个变量，下表列出了建模所用的相关数据：

 表3-1最终消费和国内生产总值数据 单位：亿元

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 地区 | 最终消费 | 国内生产总值 | 地区 | 最终消费 | 国内生产总值 |
| 北京 | 14503.6 | 23014.59 | 湖北 | 13799.7 | 31226.4 4 |
| 天津 | 7155.66 | 16538.19 | 湖南 | 14755.76 | 28902.21 |
| 河北 | 13197.78 | 29806.36 | 广东 | 37211.27 | 72812.55 |
| 山西 | 7134.7 | 12766.49 | 广西 | 8878.53 | 16803.12 |
| 内蒙古 | 7452.82 | 17831.51 | 海南 | 2242.65 | 3702.76 |
| 辽宁 | 13019.49 | 28669.02 | 重庆 | 7503.21 | 15717.27 |
| 吉林 | 5593.22 | 15507.88 | 四川 | 15774.96 | 30053.1 |
| 黑龙江 | 8986.69 | 15083.67 | 贵州 | 5957.73 | 10502.56 |
| 上海 | 14854.5 | 25123.45 | 云南 | 8855.33 | 13619.17 |
| 江苏 | 35041.42 | 70116.38 | 西藏 | 820.01 | 1026.39 |
| 浙江 | 20936.3 | 42886.49 | 陕西 | 8199.95 | 18021.87 |
| 安徽 | 10970.5 | 22005.63 | 甘肃 | 4374.1928 | 6790.3228 |
| 福建 | 10328.9 | 25979.82 | 青海 | 1485.98 | 2417.05 |
| 江西 | 8418.28 | 16723.78 | 宁夏 | 1719.65 | 2911.77 |
| 山东 | 26144.42 | 63002.33 | 新疆 | 5639.84 | 9324.8 |
| 河南 | 18722.62 | 37002.16 |  |  |  |

资料来源：《中国统计年鉴（2016）》

1. **模型建立**
2. **散点图**



图3-1散点图

如上图所示，随着解释变量X增加，被解释变量Y的值不断变大，基本上分布在一条直线附近，X，Y之间可能存在线性关系，尝试拟合一元线性回归模型。

1. **普通最小二乘（OLS）估计**

建立回归模型： （3.1）

其中，表示最终消费支出，表示各地区国内生产总值，为随机误差项，待估参数为和。

用Stata软件进行OLS估计，命令为：reg y x。结果如下：



图3-2回归结果

即

 (1.19) (30.29)

 

1. **模型检验**

一般经验告诉我们，对于采用截面数据作为样本的计量经济学问题，由于在不同样本点上解释变量以外的其他因素的差异较大，往往存在异方差性。对我们建立的模型进行异方差检验。

1. **图示检验法**



图3-3残差平方图

如上图，横坐标为解释变量X，纵坐标为进行最小二乘估计后得到的残差的平方，是随机干扰项方差的估计值。由上图可以看出，残差的平方随国内生产总值的增加而增加，由此指出了异方差的存在。但图示检验法只能进行大概的判断，我们需要其他更为严格的方法。以布殊-帕甘检验为例。

1. **布罗施-帕甘检验（Breusch-Pagan test，BP检验）**

第一步：对（3.1）式进行OLS回归，结果如图3-2所示。

第二步：使用predict命令生成残差e，并生成残差e的平方e2。

第三步：做e2对解释变量X的回归，即（3.2）式，根据该方程总体显著性检验的F统计量来检验同方差假定。

 （3.2）



图3-4 BP检验结果

结果如上图所示，模型整体显著性检验的F统计量对应的P值为0.0057，这意味着无论显著性水平为0,01还是0.05，都认为同方差假定不成立，即存在异方差问题。

1. **模型修正**

当模型出现异方差问题时，虽然OLS估计量仍具有线性性、无偏性，但变量的显著性检验和模型的预测都失效，无法进一步的进行推断，因此我们要对模型进行修正，本案例中采用如下两种方法。

1. **可行的广义最小二乘法**

第一步：对（3.1）式进行OLS回归，结果如图3-2所示。

第二步：使用predict命令生成残差e，并生成残差e的平方e2，以及残差平方的自然对数lne2。

第三步：做lne2对解释变量X的回归，即（3.3）式，得到拟合值g的指数h=exp（g）（注意：这里的exp是Stata指数函数的命令）。

 （3.3）

第四步：以1/h为权数对模型（3.1）进行加权最小二乘估计。即生成新的被解释变量和解释变量：，做关于的回归，即: 

回归结果如下：



图3-5可行的广义最小二乘估计结果

下面我们用BP检验来检验是否加权的模型不再存在异方差性。做残差u2关于加权后的新变量的回归，结果如下：



图3-6 BP检验结果

方程整体显著性检验的F统计量对应的P值为0.1620，即使在0.1的显著性水平下，也不能拒绝同方差假定，即加权后的模型不存在异方差。

1. **异方差稳健标准误法**

我们也可以使用异方差稳健标准误法来修正模型的异方差问题，结果如下图所示：



图3-7异方差稳健标准误法结果

可以看出，估计的参数与普通最小二乘法的结果相同，只是由于参数的标准差得到了修正，从而使得t检验值与普通最小二乘法的结果不同。

**四、结论**

1.本案例建立了国内生产总值与最终消费之间的一元线性回归方程，回归系数大于0小于1，符合凯恩斯的绝对收入消费理论。说明社会总消费受总收入的影响。但是回归系数小于0.5，即收入增加1单位，总消费平均增加不足0.5单位，社会整体边际消费倾向较低，一定程度上反映了国民收入分配的不均等。

2.消费作为拉动经济增长的三驾马车之一，也受到收入水平的影响，对于消费水平不高的农村地区，想要提高消费水平还是应该从根本上提高收入水平。

**参考文献**

[1]赵国庆.应用计量经济学.北京：中国人民大学出版社，2011.8

[2]刘泽云，孙志军.计量经济学实验教程.北京：北京师范大学出版社，2011.9

[3]李子奈，潘文卿.计量经济学（第三版）.北京：高等教育出版社，2010.3

[4]胡志宁.Stata/EViews计量经济分析（第二版）.北京：中国人民 大学出版社，2016.1

**案例使用说明**

**一、教学目的与用途**

适用于本科生计量经济学课程，目的是让学生掌握异方差模型的基本原理及其应用，并熟悉计量经济分析建模的方法和程序。

**二、启发思考题**

1.本案例中利用的异方差的检验及修正方法在多元回归模型中怎样应用？

2.结合宏观经济学的相关理论，思考社会消费函数中边际消费倾向代表什么，它的数值大小说明什么？

3.计量经济学的建模步骤在实际应用中是否可以根据需要进行调整？

**三、分析思路**

拿到计量经济案例是应先熟悉案例背后反映的相关经济理论，根据经济理论等来设计理论模型，必须遵循从“一般到简单的原则”，之后进行样本数据的收集与描述；模型参数的估计和模型的检验；最后利用模型进行相应的分析。

**四、理论依据与分析**

（一）计量经济学的相关理论：

1.异方差的定义：

对于模型：  （1.1）

同方差假定为：

如果出现：

即对于不同的样本点，随机干扰项的方差不再是常数，而是互不相同，则认为出现了异方差性。

2.异方差的后果：参数估计量非有效；变量的显著性检验失去意义；模型的预测失效。

3.异方差的检验：

（1）图示检验法：既可用Y-X的散点图进行判断，也可用的散点图来判断。对前者看是否存在明显的散点扩大 、缩小或复杂型趋势（即不在一个固定的带型域中）；对后者看是否形成一条斜率为零的直线。

（2）布罗施-帕甘检验（BP检验）：

该检验的基础思想是对于 （1.2）

进行假设检验 

这实际上是对模型（2）进行回归总体的显著性检验。如果拒绝了原假设，说明与一个或多个解释变量相关，从而模型（1.1）存在异方差。由此可以总结出BP检验的三个步骤：

第一步：对模型(1.1)进行OLS估计，得到残差平方

第二步：对模型（1.2）进行OLS估计，并进行回归总体显著性检验。

第三步：如果回归总体显著性检验的p值足够小，就可以拒绝同方差假定，认为存在异方差。反之则无法拒绝同方差假定。

4.异方差的修正

（1）异方差稳健标准误法：

由于回归模型随机干扰项存在异方差时，OLS法只影响到了参数估计量的方差或标准差的正确估计，从而无法保证OLS估计量的有效性，但并不影响估计量的无偏性与一致性。因此我们仍可采用OLS估计量但修改相应的方差。

如在一元回归中，估计的斜率正确的方差应为：，于是用普通最小二乘法估计的残差的平方作为相应的的代表，则用下式作为的估计：，该式的平方根成为的异方差稳健标准误。这种方法成为异方差稳健标准误法。

（2）加权最小二乘法：

i）异方差具有倍乘常数的形式。

假设异方差可以写成，其中的诀窍在于找出的构成。如果我们把整个方程式除以，因为是解释变量x的函数，那么，而且因为我们知道，。通过这种方法便可得到一个同方差的模型。例如总体回归模型为，

模型整体除以，得到如下同方差的模型：

ii）可行的广义最小二乘估计。

当我们不知道异方差的确切形式时，需要对进行估计。通常，我们首先建立一个相对灵活的模型，比如：

因为我们不知道那些系数的值，必须要估计他们。我们的假设意味着，，或者方程两边都取对数转换为，其中和解释变量不相关，互相独立。我们知道是的估计，并且通过OLS可以得到。回归后，的估计值实际是，它的倒数就是我们需要的权数。

总之，我们先回归最初的OLS模型，保存残差，计算残差的平方，接着取对数。然后以为被解释变量对所有解释变量进行回归并得到拟合值。最后我们以作为加权值，以便运行加权最小二乘回归。

（二）宏观经济学相关理论：凯恩斯的绝对收入消费理论

家户消费函数：现实生活中，影响各个家户消费的因素很多，如收入水平、商品的价格水平、利率水平、收入分配状况、消费者偏好、家庭财产状况、消费者信贷状况、社会保障制度、风俗习惯等等。凯恩斯认为，这些因素中具有决定意义的是收入，关于收入和消费的关系存在一条基本心理规律：随着收入增加，消费也会增加，但消费的增加不如收入增加多（因此边际消费倾向大于0小于1），消费和收入的这种关系称作消费函数。若消费与收入之间存在线性关系，则消费函数可表示为，其中X表示收入，Y表示消费。为边际消费倾向。

社会消费函数：宏观经济分析的是整个社会的消费函数，即总消费和总收入之间的关系。社会消费函数是家户消费函数的总和，然而，社会消费函数受到国民收入的分配、政府税收政策、未分配利润在利润中所占比例等因素的影响，因此社会消费曲线并非家户消费曲线的简单加总，但在考虑了种种限制之后，社会消费曲线的基本形状仍和家户消费曲线有很大的相似之处。

**五、关键要点**

案例分析的关键是设定正确的模型形式，如果模型设定不正确，则接下来的一切分析都没有意义。

案例教学中的关键知识点：计量经济学模型的建模步骤；异方差的定义，存在异方差问题的后果；异方差的检验以及修正。

能力点：充分发挥学生的动手实操能力，能运用Stata软件进行相关的分析。

**六、建议课堂计划**

建议在教学中安排一节课时间来进行案例学习。同时在案例学习之前，应该讲授过异方差问题的相关理论，并在学习案例时对案例中用到宏观经济的相关理论进行讲解。

**案例7**

**中国居民人均消费函数**

**——豪斯曼内生性检验的应用分析**

**摘 要：**本案例描述了通过建立中国居民人均消费函数的回归模型，寻找模型中疑似具有内生性的随机解释变量，并对其进行检验的过程。

**关键词：**随机解释变量、内生性、豪斯曼检验

**一、案例背景**

英国经济学家凯恩斯在《就业利息和货币通论》（1936）一书中提出的消费函数是宏观经济学最重要的范畴之一，它反映的是消费支出水平与个人可支配收入之间的关系。消费函数的一个假设前提是，消费和收入之间存在着一种以经验为依据的稳定关系。而这种稳定的关系可被简化为线性函数形式表示，即：

其中，Y表示当年个人消费总量，X表示当年人均可支配收入，

然而，居民消费水平具有一定的惯性，即当年的消费支出在一定程度上受到上一年消费支出的影响。因此，考虑将前一年的个人可支配收入作为另一个解释变量引入到模型中。

**二、理论模型的构建**

引入上一年的消费支出对消费函数模型进行扩展，扩展后的模型与原模型具有相同的假设条件。扩展后的模型表示为：

其中，Y表示当年的个人消费总量，表示当年的可支配收入，表示上一年的消费支出，分别为参数，为随机干扰项。

根据线性回归的基本假设，解释变量应与随机干扰项相互独立。但是，考虑到在居民人均消费支出由人均可支配收入决定的同时，人均消费支出又反过来影响着同期居民人均可支配收入，那么那些未在模型中独立列出而纳入到随机干扰项的影响居民人均消费支出的因素，也有可能影响着居民人均可支配收入。因此有理由怀疑居民人均可支配收入与随机干扰项同期相关。而对于前一年的居民人均消费支出而言，它影响当年的消费支出，而当年的消费支出不会影响前一年的消费支出。因此可认为前一年的人均消费支出是同期外生变量。

因此，为了保证模型构建的合理性，我们不光要建立模型，还需要对居民人均可支配收入这一变量进行内生性检验，用到的方法为豪斯曼内生性检验法。

**三、计量检验**

**1、变量和数据**

为构建需要的计量经济学模型并对其进行检验，表1给出了2015年中国内地分地区居民人均消费支出和人均可支配收入。

表1 2015年中国内地31个省区居民人均消费支出与人均可支配收入

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 地区 | 消费支出 | 可支配收入 | 地区 | 消费支出 | 可支配收入 |
| 北京天津河北山西内蒙古辽宁吉林黑龙江上海江苏浙江安徽福建江西山东河南 | 33802.824162.513030.711729.117178.517199.813763.913402.534783.620555.624116.912840.118850.212403.414578.411835.1 | 48458.031291.418118.117853.722310.124575.618683.718592.749867.229538.935537.118362.625404.418437.122703.217124.8 | 湖北湖南广东广西海南重庆四川贵州云南西藏陕西甘肃青海宁夏新疆 | 14316.514267.320975.711401.013575.015139.013632.110413.811005.48245.813087.210950.813611.313815.612867.4 | 20025.619317.527858.916873.418979.020110.117221.013696.615222.612254.317395.013466.615812.717329.116859.1 |

表2给出了2014年中国内地分地区居民人均消费支出和人均可支配收入。

表2 2014年中国内地31个省区居民人均消费支出与人均可支配收入

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 地区 | 消费支出 | 可支配收入 | 地区 | 消费支出 | 可支配收入 |
| 北京天津河北山西内蒙古辽宁吉林黑龙江上海江苏浙江安徽福建江西山东河南 | 31102.922343.011931.510863.816258.116068.013026.012768.833064.819163.622552.011727.017644.511088.913328.911000.4 | 44488.6 28832.3 16647.4 16538.3 20559.3 22820.2 17520.4 17404.4 45965.8 27172.832657.6 16795.5 23330.9 16734.2 20864.2 15695.2 | 湖北湖南广东广西海南重庆四川贵州云南西藏陕西甘肃青海宁夏新疆 | 12928.313288.719205.510274.312470.613810.612368.49303.49869.57317.012203.69847.612604.812484.511903.7 | 18283.2 17621.7 25685.0 15557.1 17476.5 18351.9 15749.0 12371.1 13772.2 10730.2 15836.7 12184.7 14374.0 15906.8 15096.6 |

在模型中，为了方便对变量含义的理解，我们用CPC15和CPC14分别表示2015和2014两年的居民人均消费支出，而用DI15和DI14分别表示2015和2014两年的居民人均可支配收入。

**2、计量模型设定**

建立截面数据模型进行回归分析。我们建立关于2015年居民人均消费支出与2015年居民人均可支配收入以及2014年居民人均消费支出的线性回归模型。模型的形式如下：

**3、检验结果及应用分析**

应用Stata软件,可得到对全国相关数据的估计和检验结果。估计和检验结果见图1。



图1

在上述得到的回归模型中，根据之前的分析，我们怀疑DI15是与随机误差项同期相关的内生变量。

为检验DI15是否为内生变量，我们使用豪斯曼内生性检验法。根据工具变量的选择原则，我们将CPC14作为DI15的工具变量，因为前一年的居民人均可支配收入会与当年的人均可支配收入有较强的相关性，但由于当年的消费支出不会影响前一年的可支配收入，因此CPC14与原模型的随机干扰项也不会存在同期相关。

将DI15关于CPC14和DI14利用OLS进行辅助回归。估计和检验结果见图2。



图2

接下来记录残差序列，并将其加入原模型后进行OLS估计得的估计和检验结果见图3。



图3

t检验表明，前的参数显著不为0。因此判断居民人均可支配收入确实是内生变量。

**四、小结**

对本案例的学习，需要强调以下几点：

（1）在计量经济学建模过程中，要检验选取的解释变量是否满足基本假设，其中，随机解释变量的内生性值得进一步关注。

（2）在选取工具变量进行豪斯曼检验时，选取的工具变量要满足基本要求。

（3）工具变量法并没有改变原模型，只是在原模型的参数估计过程中用工具变量“替代”随机解释变量。

（4）判断X与是否同期相关，实际上等价于判断v与是否是同期相关的。

此外，通过对案例的进一步分析，不妨思考以下几个问题：

（1）模型中出现了内生变量，若仍采用OLS估计，会对参数的估计量会产生什么影响？

（2）若模型中含有内生变量，应采用哪些方法对模型的参数进行估计？

（3）若一个随机解释变量可以找到多个相互独立的工具变量，该怎样充分利用这些工具变量的信息？

（4）若原回归模型有多个随机解释变量被怀疑与随机干扰项同期相关，该如何判断哪些解释变量确实是内生变量？

**参考文献：**

[1] 李子奈，潘文卿. 计量经济学. 高等教育出版社. 2014.

[2] 胡志宁. Stata/Eviews计量经济分析. 中国人民大学出版社. 2010.

[3] [美]保罗▪萨缪尔森，威廉▪诺德豪斯. 经济学. 人民邮电出版社. 2016.

**案例使用说明**

**1.教学目的与用途**

本案例适用于本科阶段计量经济学课程，目标是为了使学生理解并会应用豪斯曼内生性检验法。

**2.启发思考题**

（1）模型中出现了内生变量，若仍采用OLS估计，会对参数的估计量产生什么影响？

（2）若模型中含有内生变量，应采用哪些方法对模型的参数进行估计？

（3）若一个随机解释变量可以找到多个相互独立的工具变量，该怎样充分利用这些工具变量的信息?

（4）若原回归模型有多个随机解释变量被怀疑与随机干扰项同期相关，该如何判断哪些解释变量确实是内生变量？

**3.分析思路**

本案例通过对经济学变量内在关系的理解，试图将上一年的人均消费支出引入消费函数。但是根据经验，新引入的解释变量可能与随机干扰项是同期相关的。因此，为了确保回归模型的合理性，需要对该新增解释变量进行内生性检验。

**4.理论依据与分析**

本案例需要一定的宏观经济学中有关消费函数的知识，这可以帮助我们更好的理解模型中变量的含义和彼此之间的经济关系。

此外，还需要计量经济学中经典单方程一元线性回归模型和放宽基本假定的模型中的相关知识，利于对模型进行计量经济学检验。

**5.关键要点**

本案例的关键要点在于首先要能找到具有内生性的变量，这需要对模型中变量之间的关系具有非常明确的认识。

其次，还需要对豪斯曼内生性检验的基本思想有明确的认识，认识到该方法的内在原理。

需要具备一定的用Stata软件进行计量经济学进分析的能力。

**6.建议课堂计划**

本案例建议使用1课时。前半课时用于基本原理的讲解，后半课时可让学生进行上机操作。

1. 详见参考文献Hui Zou and Hastie Trevor（2005）。 [↑](#footnote-ref-2)
2. 由于筛选出的变量过多，限于篇幅，这里仅列出被剔除掉的变量。 [↑](#footnote-ref-4)
3. 由于预测模型变量较多，限于篇幅，在此没有写出模型的具体形式。 [↑](#footnote-ref-5)
4. 详见参考文献Hui Zou and Hastie Trevor（2005）。 [↑](#footnote-ref-6)