Finance 金融, 2019, 9(4), 390-408
Published Online July 2019 in Hans. http://www.hanspub.org/journal/fin
https://doi.org/10.12677/fin.2019.94047

Hans汉斯

The Influence of Chinese Population Features on Economic Development

Pailiang Ye, Hanyang Wei, Zhengyi Qi, Ruisong Ye*

Department of Mathematics, Shantou University, Shantou Guangdong
Email: *rsye@stu.edu.cn

Received: Jul. 7th, 2019; accepted: Jul. 21st, 2019; published: Jul. 29th, 2019

Abstract

This paper mainly focuses on the relationship between the potential factors with economic indices in order to provide developing suggestions. First, we preprocess the original data, including data reading, deleting, rename, standardization and so on, to make the data easily read and recognized. Then, we use correlation analysis to tell which factor is positive and negative and the death is not a strong linear factor to economic indexes, while others are. After that, we use three methods to explore the relationship between one factor and one of the economic indexes, including linear regression, polynomial regression as well as nonlinear regression, and we find out utilizing the residual is the least to determine the better way to match the relationship. Finally, we use five methods to do the multiple factors analysis, including directly using multiple regression, PCA + multiple regression, stepwise regression, time series analysis, to determine the best model for predicting the economic development among them.

Keywords

Linear Regression, Polynomial Regression, Nonlinear Regression, Multiple Regression, PCA + Multiple Regression, Stepwise Regression, Time Series Analysis

中国人口特征对经济发展的影响

叶派良，韦汉阳，戚正奕，叶瑞松*

汕头大学数学系，广东 汕头
Email: *rsye@stu.edu.cn

收稿日期：2019年7月7日；录用日期：2019年7月21日；发布日期：2019年7月29日

摘　要

本文主要关注潜在的人口因素与经济指标之间的关系，以期提供经济发展的建议。首先，对原始数据进

*通讯作者。

文章引用: 叶派良, 韦汉阳, 戚正奕, 叶瑞松. 中国人口特征对经济发展的影响[J]. 金融, 2019, 9(4): 390-408.
DOI: 10.12677/fin.2019.94047

行预处理，包括数据获取、截取、重命名、标准化等，使数据易于读取和识别。然后用相关分析来判断因素存在的正或负相关影响，以及哪个因素是显著的，哪个因素是不显著的。之后，使用线性回归、多项式回归、非线性回归三种方法来探索一个因素与经济指标之间的关系，找到更好的匹配方法。最后通过使用四种方法进行多因素分析，包括直接使用多元线性回归，主成分分析(PCA) + 多元线性回归，逐步回归，时间序列，找到一个最为合适的回归方法。

关键词

线性回归，多项式回归，非线性回归，多元回归，PCA + 多元线性回归，逐步回归，时间序列分析

1. 引言

人口与经济的关系是人类社会经济发展研究中最基本、最重要的问题之一。人口不仅作为生产力的决定性因素参与直接生产过程，而且作为消费者主体成为生产过程的目的和归宿。如果人口太少，就会缺乏劳动力投入生产，而人口过剩需要太多的资源来支持。可见，人口的发展，一方面可以促进经济的发展，另一方面，限制了经济的发展。作为世界上人口最多的国家，人口分布一直是我国经济发展的关键。自 1954 年以来，我国一直把人口政策作为调节人口和促进经济发展的手段。其中最著名的计划生育政策有效地控制了中国人口增长的速度，创造了 90 年代的人口红利，促进了经济的发展，是创造“中国经济奇迹”的重要因素之一。但是近年来，从高到低生育率迅速的转变之后，我们的人口的主要矛盾不再是增长过快，而是增长放缓，失去人口红利，靠近超低生育水平，老龄化的人口和出生性别比的失衡。数据显示，在 2017 年，我国总出生人口为 1723 万人，出生率为 12.43‰相比 2016 年的 1786 万人，出生率为 12.95‰，2017 年出生人口和人口出生率都有小幅下降。这主要是因为 2017 年在二孩数量增加的同时，一孩数量比 2016 年减少了 249 万人。也就是说造成我国人口出生率水平降低的原因主要是有相当一部分夫妇连一胎都拒绝生育。根据这一趋势，在目前因政策激励产生的二胎需求在几年内释放完毕之后，我国人口增速可能会迎来断崖式的下跌。在这个时代大背景下，进行人口数量、人口质量、年龄结构等各种人口指标对人口预测，GDP 调节的研究，对国家未来人口政策的制定，优化我国人口结构都有着重要意义。

胡鞍钢教授在《社会与发展——中国社会发展地区差异研究》[1]一书中对各人口指标对经济的影响进行了全面的定性分析，而在此基础上本文则应用概率统计的一些方法分析中国人口的一些特征因素对经济发展的影响。首先用相关分析来判断因素存在的正或负相关影响，以及哪个因素是显著的，哪个因素是不显著的。然后使用一元线性回归[2]、多项式回归[3]、非线性回归[4]三种方法来探索一个因素与经济指标之间的关系，找到更好的匹配方法。最后通过使用多元线性回归[5]，主成分分析(PCA) [6]结合多元线性回归，逐步回归[7]，时间序列分析[8]四种方法进行多因素分析，找到一个最为合适的回归方法。我们的数据来自国家统计局网站 http://www.stats.gov.cn/tjsj/。数据范围选取 1958 年至 2017 年的数据，包括 1) 经济指标：国内生产总值(亿元)、第一产业增加值(亿元)、第二产业增加值(亿元)、第三产业增加值(亿元)、人均国内生产总值(元)；2) 潜在影响经济指标的相关人口因素：居民消费水平(元)、农村居民消费水平(元)、城镇居民消费水平(元)、人口自然增长率、人口死亡率、普通本专科毕业生人数(万人)。

2. 原始数据预处理

由于原始的数据中有些年份的数据缺失或是极端数据，不合适做分析，所以有必要对原始的数据进行一些处理。对得到的数据进行初步的缺失检查。由于国家统计局的数据是工整的收集在 excel 文档中，所以针对数据的重命名，可以采取直接在 excel 文件中重新整理。然后，直接使用 R 语言里的“read.delim(“clipboard”)”来获得数据集。在读取数据之后，使用“complete.cases()”函数来检查是否有缺失值，得到缺失值数量是“0”，这意味着没有缺失值。

对检验过后的数据进行做标准化处理。注意到某些数值的格式不相同，而可能会由于不同数据的精度并不相同，从而影响结果，所以我们将所有数据标准化为数据型。然后，使用 EDA 分析[9]，其中包括“直方图”，“散点图”，“箱线图”，“正太 Q-Q 图”的四种图从四个直观的维度来获得数据的基本信息并且在此基础上面进行合理的分析。通过分析，我们可以得到如下的分析结果：1) 国内生产总值和死亡率的箱线图里面表示，都存在一些异常值，意味着必须要进行异常值的分析和消除。2) 从直方图和散点图来看，可以知道，不论是国内生产总值还是死亡率，数值较低的点占主要的数量。3) 国内生产总值和死亡率不符合正态分布特征，意味着不可以使用常规的概率分布的性质来进行分析和解释。同理，还可以对其他变量进行近似的分析：都存在一定数量的异常值，数值较低的点占据主要部分。

为了删除异常值，通过查看出现异常值的地方，发现数值都是相对大了很多，而且除了自然增长率和死亡率，其他因素和指标都是由于近年的数据骤增而爆发性的出现异常点。所以，可以得出一个结论，近年来的高速发展，使得数据超过了四分之三分位点，因而才识别成异常值。然而，这种经济高速增长所带来的数据并不是异常的，是符合指数增长的，所以使用取对数，即取对数 log 函数就可以将这种大部分异常值情况消除，剩下的异常值才是真实情况下的异常值。按照上述逻辑，通过检查取对数之后的异常数据的来源，发现主要是 1958 年到 1968 年的数据。通过查阅历史，发现主要是“大饥荒事件”产生了异常值。由现有的研究[10]，我们知道这个事件可以影响未来 10 年的情况，可见，死亡率的剧增是有原因的。在这个情况下，我们决定把这段时间的数据删除，这使得所有的数据都是 1968 年到 2017 年的。然而，近年来仍有一个不正常的点。如果我们决定删减数据，会降低这个项目的可信度。如果我们决定改变这个数据，也不能反映真实情况，是一个进退两难的问题。巧合的是，在接下来的研究中，死亡率并不是决定经济指标的一个相对强线性相关的因素，这意味着我们可以忽略这个问题。

3. 数据分析和结果解释

此节主要是对数据进行可视化处理，获得直观的信息。

3.1. 数据分析

从图 1 中可以看出，GDP(国内生产总值)的增长趋势最大，总量最大。第二产业增加值的增长趋势和增加的量和第三产业相似，但第一产业增长随着时间的推移而放缓，但四个经济指标都表明他们有相似的趋势，有一个近似的线性增长，说明取了对数之后，对于数据的处理可以得到简化。

为了说明截取的原因，我们在这里给出取对数前的数据图示。从图 2 中可以看出，1958 年至 1963 年间，自然增长率在 1959 后急剧下降达到−4.57%并且快速上升到峰值 33.5%，之后有逐渐下降的趋势。此外，1958~1961 年的死亡率呈现出激增的现象，我国在这一阶段遭遇了历史上的大饥荒，导致死亡人数出现了这样的突然增加，自然增长率下降到负数的情况。1962 年后，随着饥荒的结束和当时医疗技术的进步，死亡人数总体呈下降趋势。相反，由于政策的宣传，自然增长率达到了 33.5%，在那之后，由于政策的改变与计划生育政策的实施，呈现了一个下降的趋势。由于这段时间数据波动太大，因此再一次得证：需要删除这 10 年的数据。

从图 3 中可以看出，虽然 1968 年以后政策因素也发生了很大的变化，如 1971 年附近，由于某些因素，导致本科生毕业生波动较为剧烈，但是从图 1 可以看出，1971 年附近的经济发展也是减缓的。因此删除 1971 年附近数据也会很影响模型准确性，所以文章没有因为这个波动的影响而删除数据。实际上，通过这个总体的趋势，可以从宏观角度看出，它与经济指标有一定的关系，为后文的分析提供了初步的判断依据。

图 4 以经济学为基础，从居民总消费水平、农村居民消费水平、城镇居民消费水平三个维度来探讨经济发展程度。从图 8 看出，三个消费水平整体呈现一个线性增长趋势，和经济指标相似，因此可以得出结论：它们和经济指标都有一定的联系。

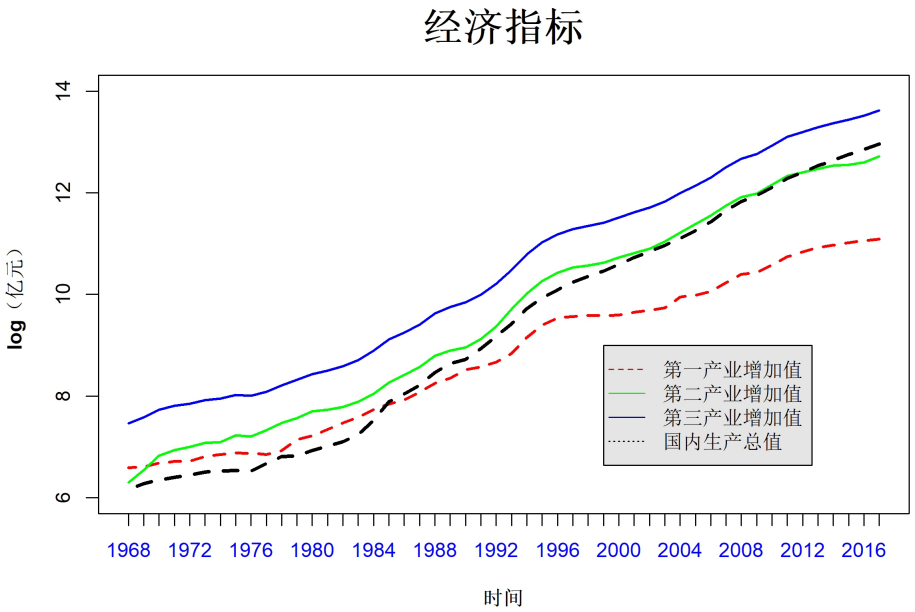

Figure 1. The visualization of truncated economic indexes
图 1. 截取后经济指标的可视化

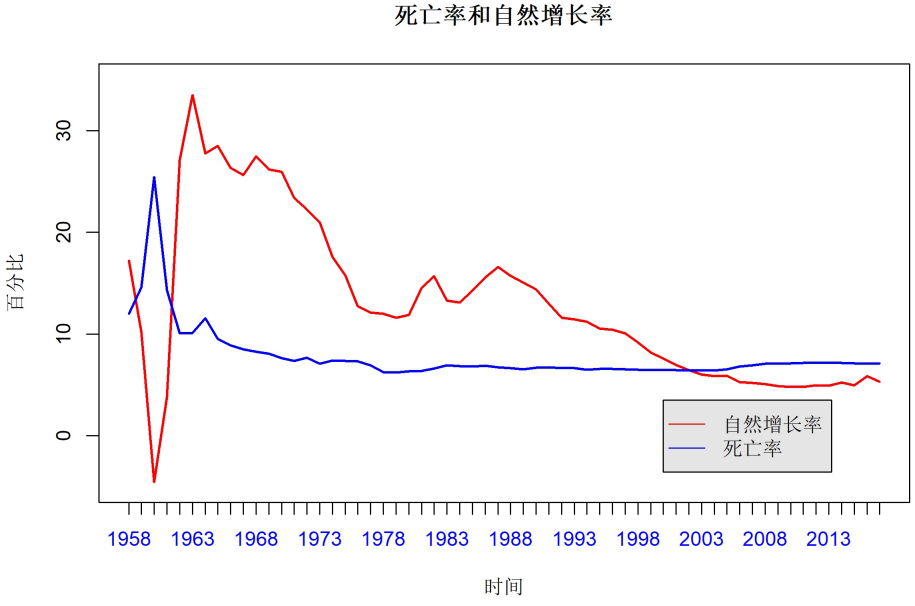

Figure 2. The visualization of untruncated death and natural growth
图 2. 截取前的死亡率与自然增长率

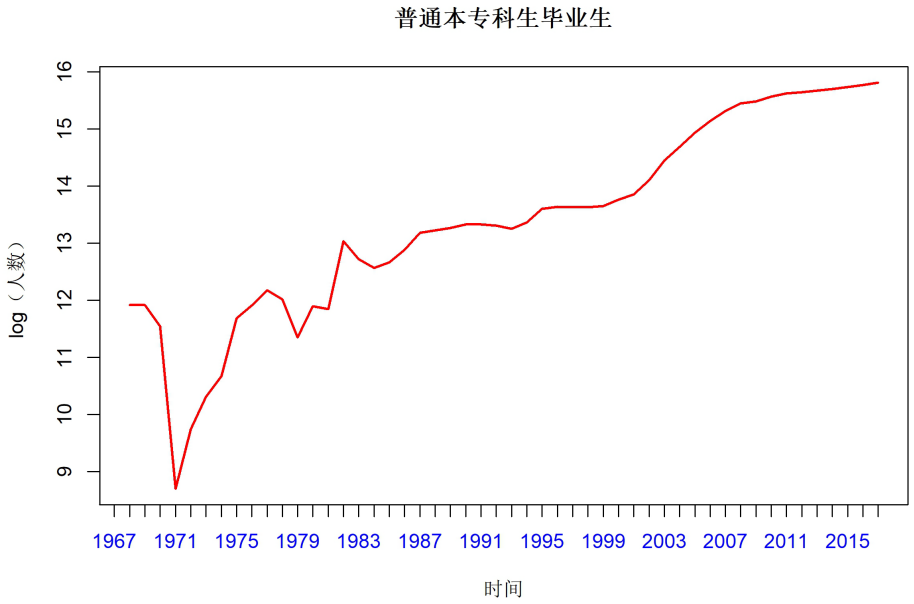

Figure 3. The visualization of truncated graduates
图 3. 截取后普通本专科毕业生

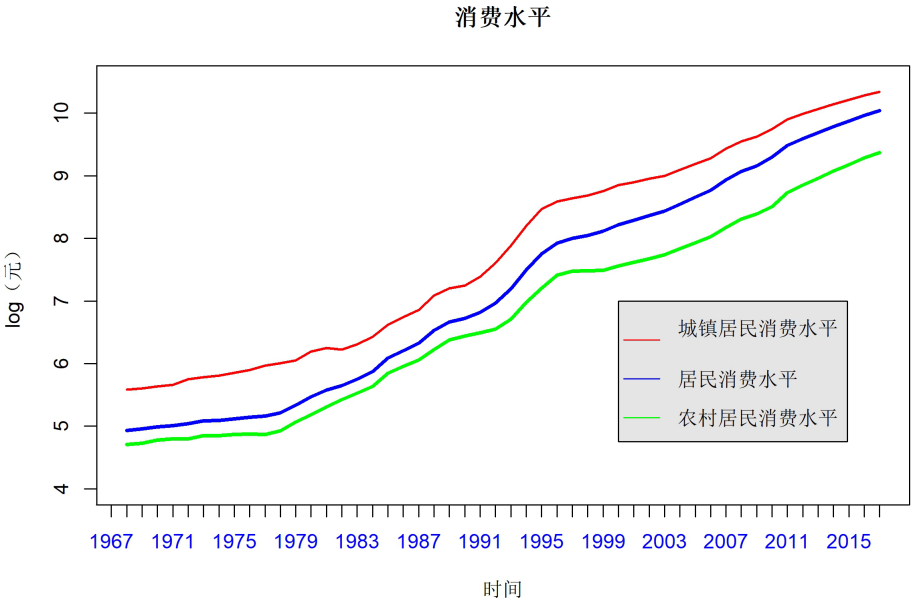

Figure 4. The visualization of truncated household consumption level
图 4. 截取后的居民消费水平

3.2. 相关分析

这一小节主要对所有变量进行强线性相关性的探索，从而筛选出相关性强的变量，为后面具体关系的探索提供理论基础。相关性系数的计算公式为

$$Cov(X,Y)=E(XY)-E(X)E(Y)$$

$$\rho_{XY}=\frac{Cov(X,Y)}{\sqrt{D(X)}\sqrt{D(Y)}}$$

其中 $Cov(X,Y)$代表随机变量 X,Y 之间的协方差，$E(X)$代表 X 的期望，$D(X)$代表 X 的方差。相关系数用来刻画随机变量 X 和 Y 之间的线性相关性，其值介于−1 与 1 之间。正值为正相关，负值为负相关，且绝对值越接近 1 线形相关性越强[4]。

为检查各个因素之间的相关性，选择出具有强线性关系的指标来进行关系的探索，所以将使用以“person”为基础的相关性分析并且从 R 语言的“corrgram”包中调用函数“corrgram”来进行操作。通过使用该方法，可以初步筛选出相关性较强的变量，为后续研究这些变量与应变量之间的关系奠定基础。结果如图 5 所示。从图中我们可以得到几个结果：

1) 死亡率与其他各项经济指标的线性相关性都相对较低，因此在这个探索强相关因素的框架下，死亡率对经济的影响将不在下文讨论。

2) 经济指标与自然增长率呈负相关关系。对这一结果的可能的解释是：经济越好，人们要孩子的愿望就越低。因此，较低的自然增长率可以在一定程度上反映经济的发展，具体的关系应当在后文继续探索。

3) 本科毕业生人数和三个消费水平这四个变量都与经济指标都有很强的正相关关系，所以可以看出：这些因素增加，对经济的影响相对较强，并且是正向影响的，因此具体的关系在后文也会继续探索。

4) 并且，注意到所有三个消费水平之间有很高的相关性，这就意味着如果我们只需寻找定性关系，而不是定量的关系，只需要探索其中一个消费水平与经济指标的关系，就可以得出另外两个消费水平和经济指标的关系，从而在后文的建议里面，能够给出较为宏观的建议。

5) 从相关图中可以看出，用来衡量经济好坏的五个经济指标两两之间的相关系数相当高，这意味着中国在经济增长方面不是一个跛足巨人，而是同时从多个方面发展起来的。当然，在后续工作中，这一信息也可以用于这种情况下：对于 GDP 的准确趋势判断，我们可以从其他经济指标中推断出同样的趋势。同理，由于此文更关注宏观变化，在以后的研究中，将使用国内生产总值，即 GDP，作为检验经济好坏的指标，在得到 GDP 和人口因素的关系之后，由于经济指标两两之间的强相关性的性质，也得到其他经济指标和人口因素也会有相似的关系。

相关图

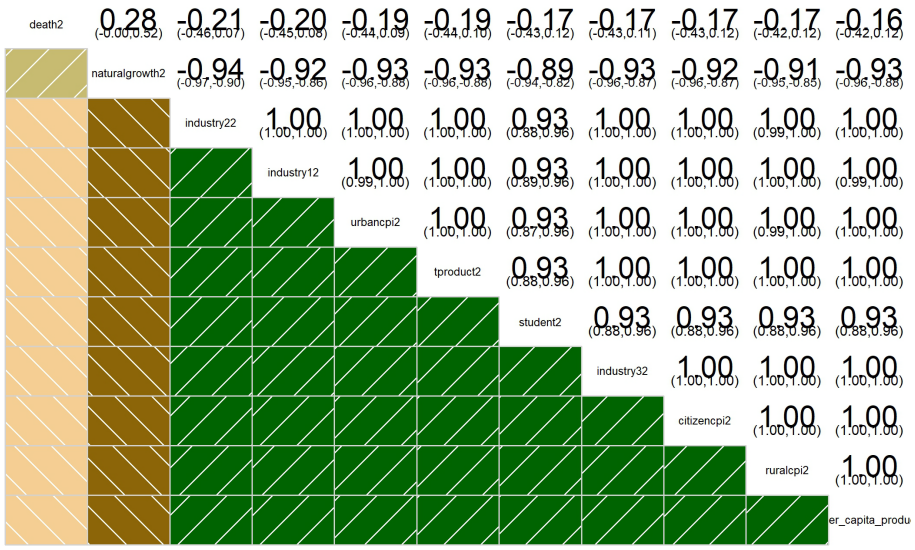

Figure 5. Correlation among factors
图 5. 解释变量的相关性

3.3. 单因素与经济指标的关系

这一小节主要对单个人口因素和经济指标的关系进行探讨。为定量探究变量间的关系，本节采用回归的方法进行分析。这是利用称为线性回归方程的最小二乘法对一个或多个自变量和因变量之间关系进行建模的一种回归分析。这种函数是一个或多个称为回归数的模型参数的线性组合。只有一个自变量的情况称为一元回归，大于一个自变量情况的叫做多元回归[11]。

$$Y_i = \sum_i^n \alpha_i X_i + \varepsilon$$

α_i 为线性回归的系数，ε 为误差。当 $n = 1$ 时候为一元线性回归，反之为多元线性回归。基于相关性分析的结果，本文可以把两两之间关系的探索进行简化，即只需要探索“人口自然增长率”、“学生人数”、“城市居民消费水平”各自成为单变量与因变量 GDP 之间的关系。因此，我们可以知道各因素与经济指标之间的所有关系。在本文中，分别使用三种回归方法来进行探索，并在其中选择一种最好的回归方式，从而得到他们的具体的关系。此外，如果我们想知道哪种方法是最好的，我们需要建立一个标准。在本文中，决定使用残差平方之和来表示从各散点到回归线的距离，最终，选择残差平方和最小的方法作为其变量和 GDP 的关系。

人口自然增长率和 GDP：图 6~图 8 和表 1 给出人口增长率与 GDP 一个因子之间的线性回归、多项式回归和非线性回归的结果。从中可以得到以下的分析结果。

1) 由结果可知，“人口自然增长率”与 GDP 之间存在较好的线性关系，利用一元多项式回归(次数 = 3)可以较好地拟合该模型。因此，在“人口自然增长率”和 GDP 之间，可以认为存在一个多项式回归关系，即，这意味着对原数据取对数之后，人口自然增长率和 GDP 之间会有一个多项式关系。

2) 为了确保没有过度回归，只选择了二次和三次多项式回归。然而，在自然增长率的数据中，因为复杂的政策因素，波动还是比较大的，导致结果可能不够精准。我们可以看到在回归方程中自然增长率与 GDP 呈现了负相关的关系，这和我们日常认知不符，初步分析我们认为是自然增长率的下滑对 GDP 产生的负面影响被技术发展和人口素质的提高等其他因素产生的正面影响所抵消掉，与此同时经济社会的发展会对自然增长率的增长形成压力。从社会的角度来看，它确实符合人们的行为模式。随着经济的发展，社会观念的开放，越来越多高学历的女性进入职场，对于事业女性来说要孩子对自己的事业无疑是一项巨大的冲击。越来越多的晚婚、晚孕族，也极大的影响了出生率。另外，随着社会福利的完善与收入的增长人们并不需要生育更多的孩子来为老年的生活作保障，他们把多余的时间精力金钱，用于自己享受，或者人们想要存更多的钱提供更好的物质条件给孩子，然而当条件达不到的时候，往往选择放弃生孩子。因此，他们生育的计划会被推迟，使总的生育率降低。

Table 1. Data from regression analysis
表 1. 回归分析的数据

	残差平方和	Rank
单因素线性回归	0.5174033	4
单因素多项式回归(次数 = 2)	0.4832927	2
单因素多项式回归(次数 = 3)	0.4723901	1
非线性回归的一个因素(指数)	0.5125726	3
非线性回归的一个因素(倒指数)	0.5894593	5

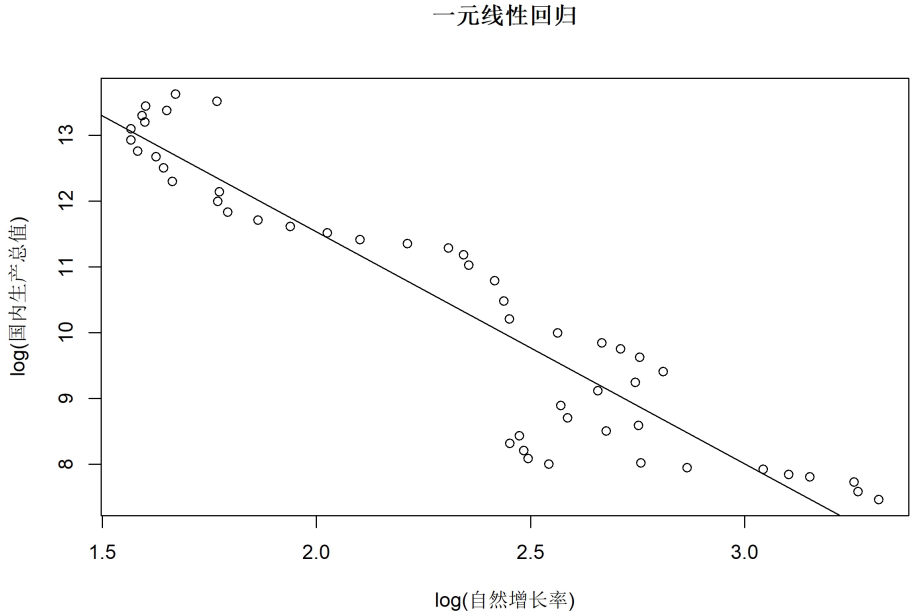

Figure 6. Linear regression of one factor between natural growth and GDP
图 6. 人口自然增长率与 GDP 之间一个因子的线性回归

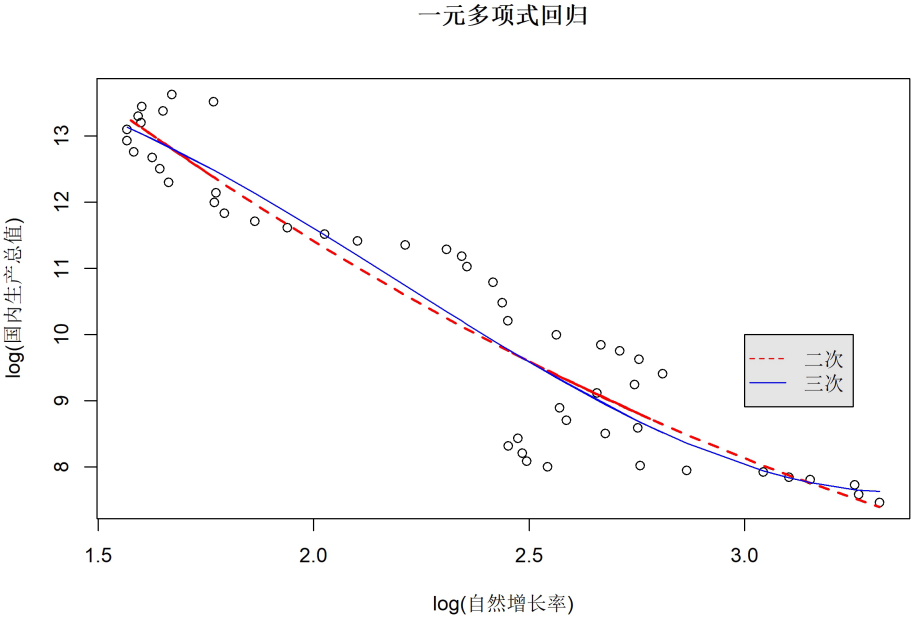

Figure 7. Polynomial regression of one factor between natural growth and GDP
图 7. 人口自然增长率与 GDP 之间一个因子的多项式回归

学生人数和 GDP：图 9~图 11 和表 2 给出学生人数与 GDP 之间单因素线性回归、多项式回归和非线性回归的结果。从中可以得到以下的分析结果。

1) 在前面段的年限里面，普通本专科学生毕业人数处于较低的数值，从而导致前面段的回归效果都会较差。

2) 通过具体的毕业人数，通过后半段的分析，在政策干扰的减少下，规律变得更加明显：在普通本专科毕业生人数的趋势可以看出，自改革开放以来，我国确定的优先发展教育的战略产生了巨大的作用。

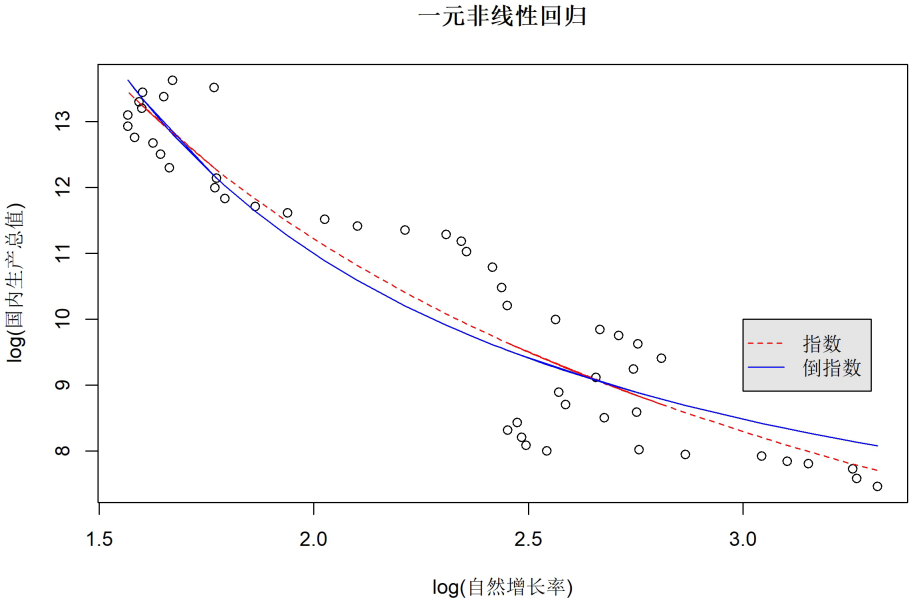

Figure 8. Nonlinear regression of one factor between natural growth and GDP
图 8. 人口自然增长率与 GDP 之间一个因子的非线性回归

普通本专科毕业生人数呈爆发性增长。从 1978 年的 16.5 万人达到 2016 年的 704.18 万人，教育事业的发展也取得了显著的成绩，对推动了我国经济的发展起到了举足轻重的作用。我们的回归分析也证明了这一点。本专科毕业生人数与 GDP 在次数为三的多项式回归中平均残差绝对值仅有 0.2104003。在各种方程中拟合效果最好，而次数为三也恰恰表明了这项指标对 GPD 的推动作用。可以说社会经济的发展离不开高素质人才，随着高素质人才的增多，他们对社会的贡献越大，从而经济会变得更好。所以中国在狠抓普及义务教育，扫除文盲这两大重点的同时，也要积极发展高等教育，完善国内本专科的教育制度，提升专科，本科学校的办学质量，为社会源源不断地输出高素质人才，在长期逐渐提高人口的文化素质水平。

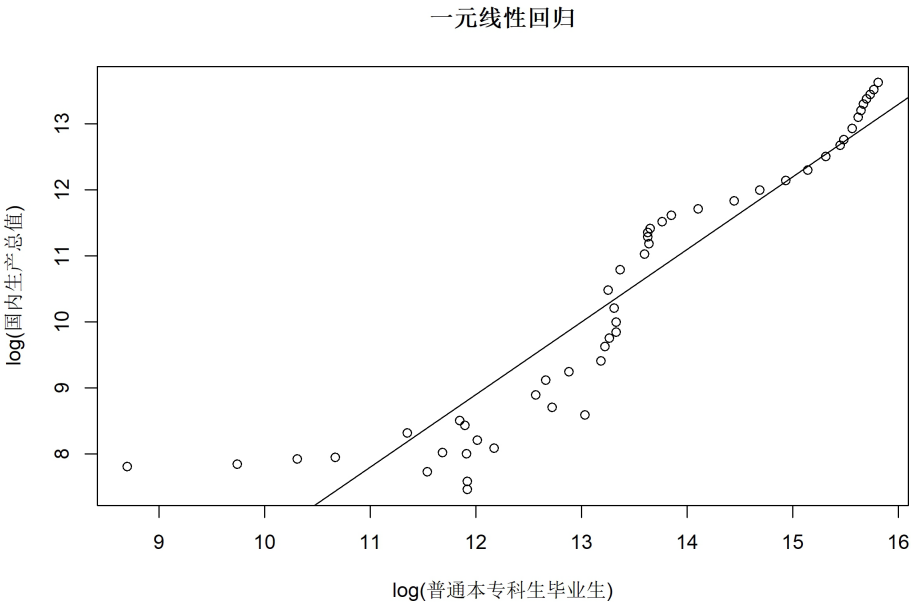

Figure 9. Linear regression of one factor between graduates and GDP
图 9. 学生人数与 GDP 之间一个因子的线性回归

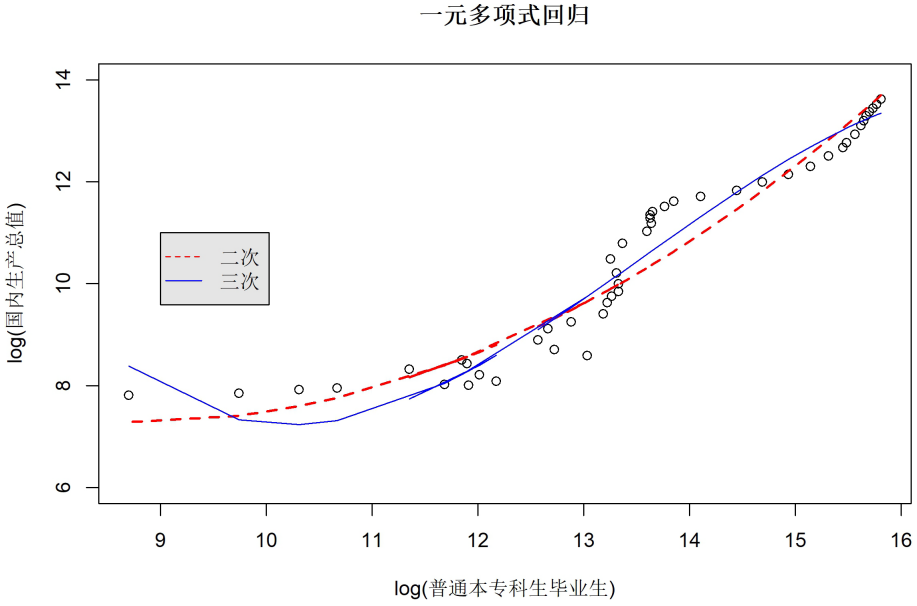

Figure 10. Polynomial regression of one factor between graduates and GDP
图 10. 学生人数与 GDP 之间一个因子的多项式回归

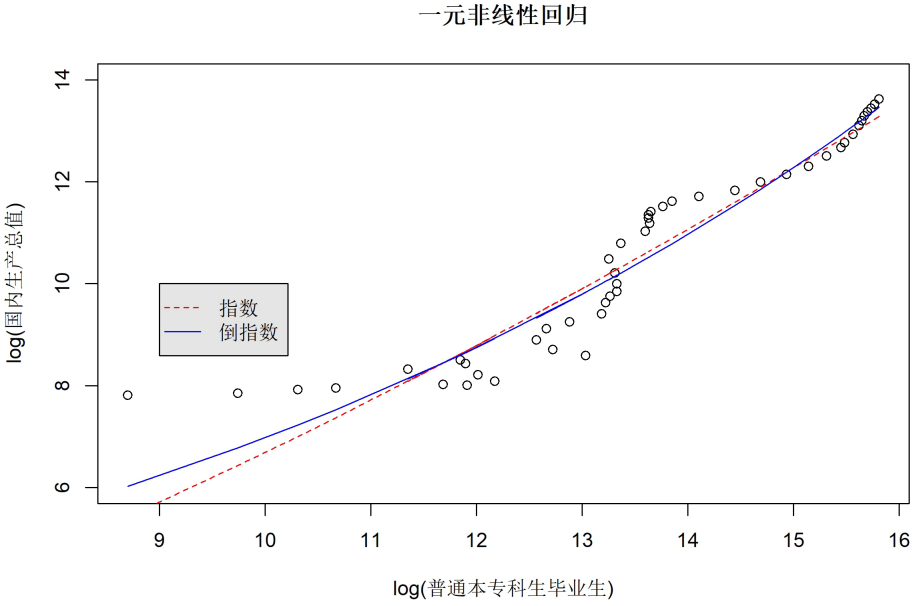

Figure 11. Nonlinear regression of one factor between graduates and GDP
图 11. 学生人数与 GDP 之间一个因子的非线性回归

Table 2. Data from regression analysis
表 2. 回归分析的数据

	平均残差绝对值	Rank
单因素线性回归	0.5195956	5
单因素多项式回归(次数 = 2)	0.2958184	2
单因素多项式回归(次数 = 3)	0.2104003	1
单因素非线性回归(指数)	0.4521024	4
单因素非线性回归(非指数)	0.3696944	3

城市居民消费水平与 GDP：图 12~图 13 和表 3 给出城市居民消费水平与 GDP 之间单因素线性回归、多项式回归的结果。1、由于前面已经讨论过消费水平之间的相关性非常强，所以只需要找出其中一个消费水平与 GDP 的关系就可以类似找到了所有消费水平和经济指标的关系。具体来说，本文利用城市居民的消费水平来探讨消费水平与经济指标之间的关系。2、从上述三种方法来看，效果都很不错，而之所以这种线性关系如此明显，主要是因为消费水平与经济指标之间的关系存在。从回归结果可以看出，城市居民消费水平与 GDP 呈正相关的关系，而且 GDP 和城市居民消费水平的增长是相辅相成，并且幅度相似，否则不能出现如此显著的正相关关系。使用经济学理论也可以得出同样的结果，城市居民消费水平的提高，反映出市场需求的增加，从而可以带动供给的提高，有利于促进企业生产。而企业的营运情况良好，则又会使居民收入提高，从而有更高的居民消费水平。可见我们要发展国民经济，就需要提高消费水平，拉动内需，形成经济发展良性的循环。

3) 由于线性关系的显著，此处不再讨论非线性回归的结果。

3.4. 多因素与经济指标的关系

本节主要是对多因素综合考虑的情况下，与经济指标的关系进行探索，并选取最佳方法进行经济指标的预测。建立在分析单因素与经济指标关系的基础上，利用已有的变量来探讨经济指标的走势。因此，在接下来的文章中，我们将用四种方法来探讨多种因素与经济指标之间的关系。

多因素的线性回归：图 14 和表 4 给出多元回归分析给出以下的相关结果。可以得到以下的几点结论：

1) 通过方差齐性检验，F 检验统计量是 1158，自由度 5 和 44，p 值接近为 0，这意味着它是非常显著的。

2) 农村居民消费消费水平和自然增长率是最显著的影响因素，普通本专科毕业生人数所带来的影响不明显。

3) 另外，由于农村居民消费水平、城市居民消费水平、居民消费水平之间有一定的关系，这个现象肯定会影响整体回归结果。站在这个角度来看，直接使用多元线性回归并不是一个好的选择。

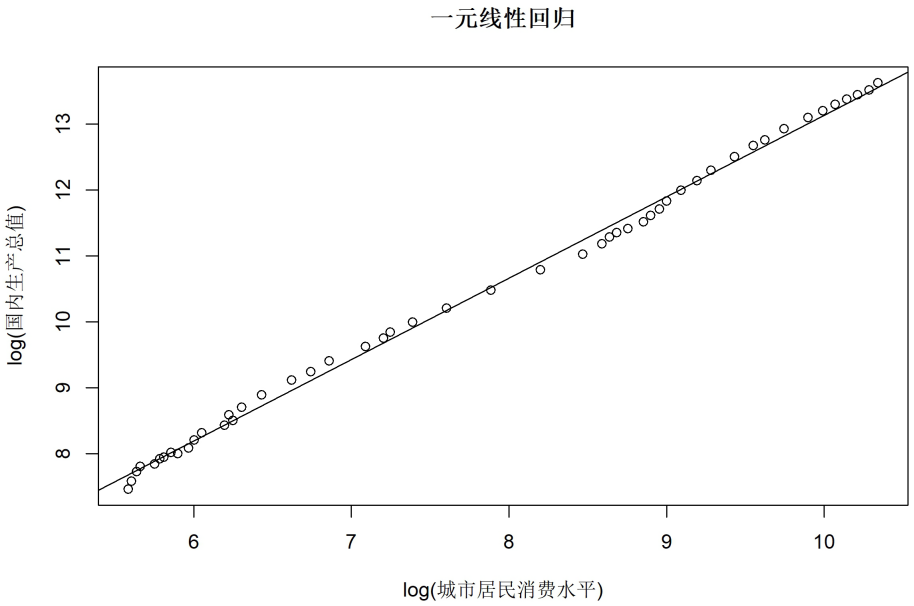

Figure 12. Linear regression of one factor between urban household consumption level and GDP
图 12. 城市居民消费水平与 GDP 之间单因素线性回归

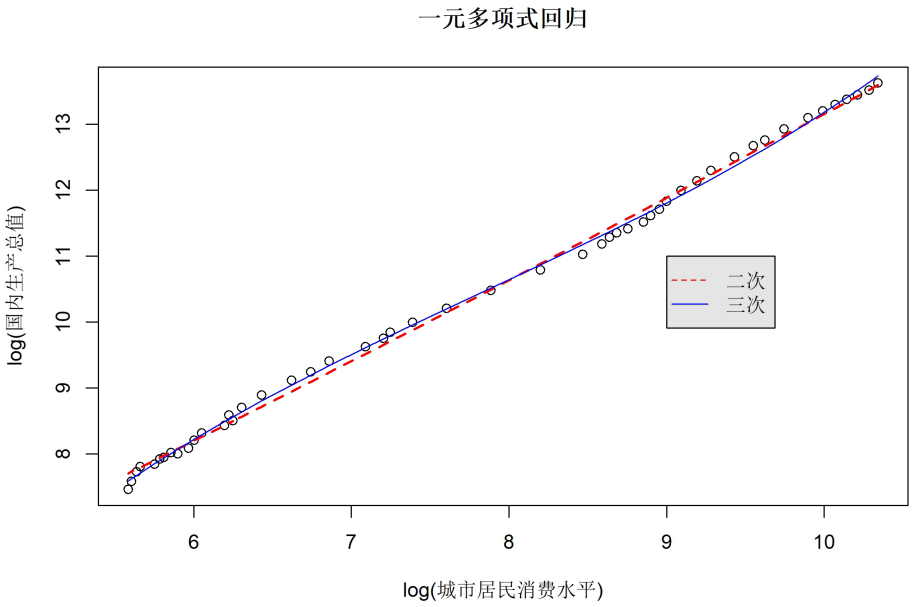

Figure 13. Polynomial regression of one factor between urban household consumption level and GDP
图 13. 城市居民消费水平与 GDP 之间单因子多项式回归

Table 3. Data from regression analysis
表 3. 回归分析的数据

	平均残差绝对值	Rank
单因素线性回归	0.01185898	3
单因素多项式回归(次数 = 2)	0.0114601	2
单因素多项式回归(次数 = 3)	0.006362047	1

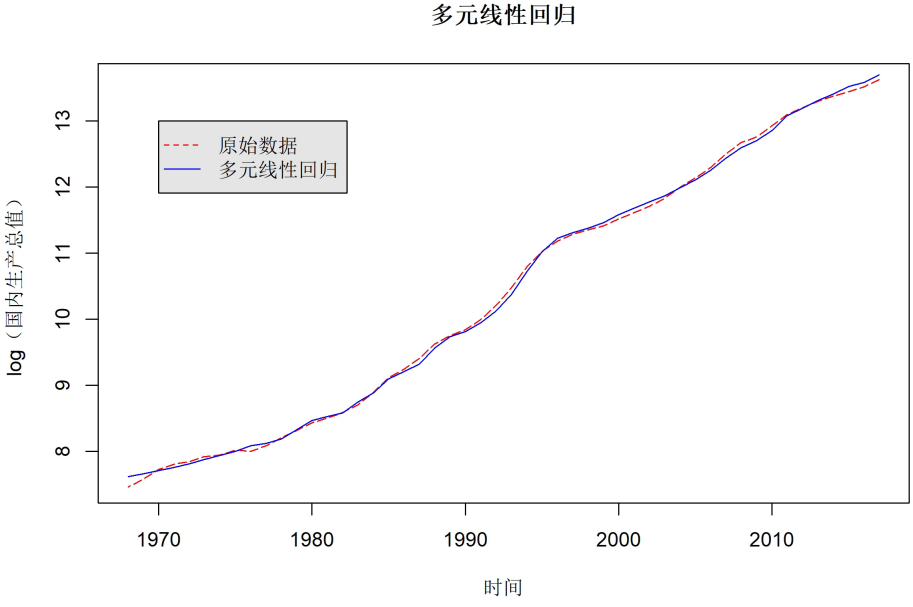

Figure 14. Linear regression of multiple factors
图 14. 多因素的线性回归

4) 再有，注意到学生人数变量对应的 p 值非常的大，意味着显著的认为这个变量不显著。事实上，在现实生活里面，可以说明普通本专科毕业人数对国内生产总值等经济指标的影响会转变成：对三个消费水平和人口自然增长率的影响，从而间接影响国内生产总值。

Table 4. Results of linear regression of multiple factors
表 4. 多因素线性回归的结果

	系数估计值	标准差	T 值	P 值
截距项	2.7984114	0.3135388	8.925	1.97×10^{-11}
城市居民消费水平	0.1399219	1.786	0.0810	\
乡村居民消费水平	0.1875385	2.270	0.0282	\
居民消费水平	0.4784187	0.2683962	1.783	0.0816
学生人数	−0.0004804	0.0148853	−0.032	0.9744
人口自然增长率	−0.2805960	0.0526438	−5.330	3.22×10^{-6}
余项标准差：0.05863 有 44 个自由度				
R 方：0.9992；调整后 R 方：0.9992				
F 值：1.158×10^{6}；P 值：$<2.2 \times 10^{-16}$				

主成分分析 + 多因素线性回归：通过上面的信息，我们知道我需要用一些方法来降低变量共线性问题。经典的操作是主成分分析，即 PCA，它可以建立出一个因素之间互相线性无关的模型，这意味着它在理论中更可靠。其中主成分分析的本质是：一个正交化的线形变化，把数据变换到一个新的坐标系统中，使得这一数据的任何投影的第一大方差在第一个坐标(第一主成分)上，第二大方差在第二个坐标(第二主成分)上，依次类推起到降维的作用：

$$Y^T = X^T W = VB^T W^T W = VB^T$$

其中 X^T 为标准化的数据，W 为一个正交矩阵，V 是矩阵 $\boldsymbol{X^T X}$ 的特征向量矩阵[12]。在 R 语言中，本文利用“pls”包来实现主成分分析，并且在此之后，再一次进行多项式回归。主成分分析后的多元线性回归结果如下：

国内生产水平 = 10.406726 + 0.019648 * 普通本专科毕业人数 + 0.409492 * 城镇居民消费水平
+0.362156 * 农村居民消费水平 + 0.418580 * 居民消费水平 − 0.092223 * 人口自然增长率

数值结果如图 15 和表 5~表 7 所示。我们得到几点结论：

1) 通过方差齐性检验，F 检验统计量是 2278，自由 2 和 47，p 值 = 0，这意味着它是显著的，它比不使用主成分分析的多元线性回归方法好。

2) comp1 和 comp2 是最显著的因素，并且通过累积贡献率来看，这两个因素达到的贡献率为 99.5%，意味着已经可以表示 99.5%的数据信息，所以其他的因素不需要再考虑。

3) 该方法不仅在数据角度上，而且在实际理论分析和现实情况下都有较好的效果。因此用这个方法预测 GDP 的未来，结果也肯定会更加可信。

逐步回归法：该方法将删除一些数据，以便更好地从数据角度，得到 GDP 的变化规律。在 R 中，我决定使用“nutshell”和“MASS”包进行操作。结果如图 16 和表 8 所示。我们也得到几点：

1) 通过方差齐性检验，F 检验统计量是 1481，自由度是 4 和 45，p 值接近于 0，这意味着它是显著的，但是 F 检验统计量没有主成分分析后的多元线性回归大，在这个维度上，结果没有主成分分析后的多元线性回归方法好。

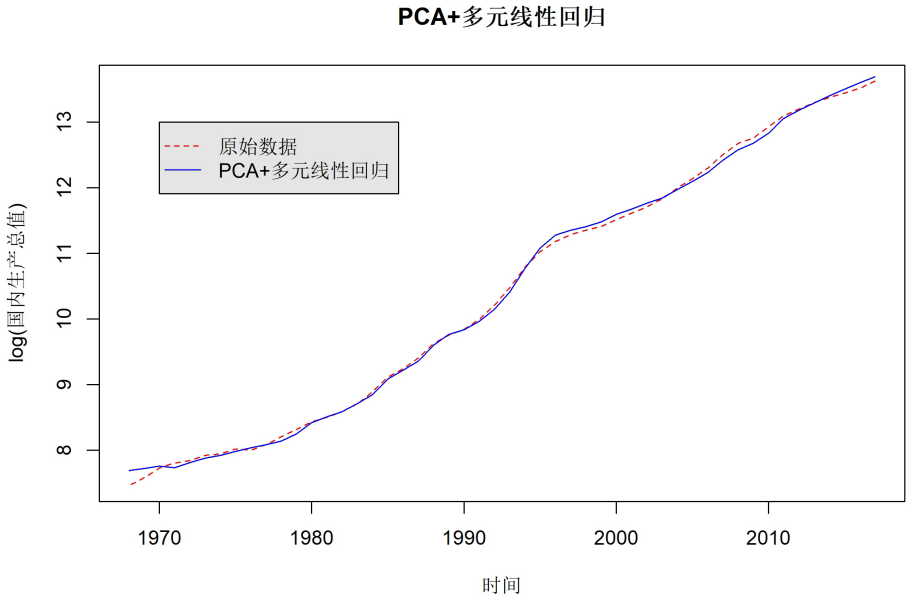

Figure 15. Linear regression of multiple factors with PCA
图 15. 多因素的 PCA + 线性回归

Table 5. Results of PCA
表 5. 主成分分析的数据

	系数估计值	标准差	T 值	P 值(>|t|)
截距项	10.406726	0.009349	1113.19	$<2 \times 10^{-16}$
成分 1	0.610752	0.002873	212.59	$<2 \times 10^{-16}$
成分 2	−0.331471	0.017337	−19.12	$<2 \times 10^{-16}$
余项标准差：0.0661；自由度：47				
R 方：0.999				
调整后 R 方：0.9989				
F 统计量：2.278×10^{4}；自由度 5 和 47				
F 的 P 值：$<2.2 \times 10^{-16}$				

Table 6. Details of two components
表 6. 两个成分的具体信息

	成分 1	成分 2	成分 3	成分 4	成分 5
累计贡献率	0.9691031	0.99571335	0.999415916	0.9999457307	1.000

Table 7. Coefficients of each factor
表 7. 各指标对各成分的系数

	成分 1	成分 2	成分 3	成分 4	成分 5
学生人数	0.500	0.862	\	\	\
城市居民消费水平	0.493	−0.327	0.726	−0.338	\
乡村居民消费水平	0.460	−0.245	0.277	−0.625	−0.512
居民消费水平	0.522	−0.301	−0.114	0.787	\
人口自然增长率	−0.151	0.952	0.258		\

2) 结果表明，通过这种方法的操作，不得不删除变量“普通本专科毕业学生人数”，然而，现实表明：“普通本专科毕业学生人数”对社会和经济有一定的贡献，这意味着这个结果是违背现实的。出现这样的现象的根本原因是：这个方法是通过逐步剔除线性因素，使最终的各个因素都不是线性的，但是这个方法只是单纯从数据角度出发，所以会存在违背现实的情况。因此，本文不认为这是一个好方法。而由于这种方法本身的原因，本文不认为这样的经济指标的预测具有实际意义。

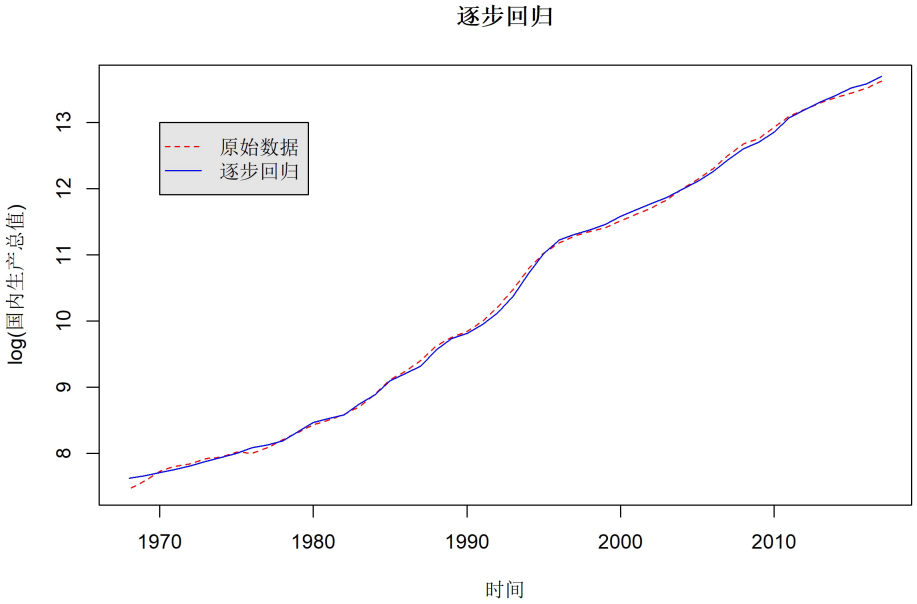

Figure 16. Stepwise regression
图 16. 逐步回归

Table 8. Results of stepwise regression
表 8. 逐步回归的结果

	系数估计值	标准差	T 值	P 值(>|t|)
截距项	2.79217	0.24405	11.441	6.5×10^{-15}
城市居民消费水平	0.25070	0.13608	1.842	0.0720
乡村居民消费水平	0.42463	0.18250	2.327	0.0245
居民消费水平	0.47832	0.26538	1.802	0.0782
人口自然增长率	−0.27999	0.04867	−5.753	7.27×10^{-7}
余项标准差：0.05798；自由度：45				
R 方：0.9992；调整后 R 方:0.9992				
F 统计值 1.481×10^{4}；自由度：4 和 45；P 值 $< 2.2 \times 10^{-16}$				

时间序列分析方法：时间序列分析方法是分析时间序列中观察值之间相关性的一种有效的方法。借助差分运算和平稳时间序列的 ARMA 建模方法，可以构建有效刻画时间序列中的相关性模型。从原始的 GDP 序列看一看出序列是一个有趋势的非平稳时间序列。可以通过一阶差分消除线性增长趋势，然后对差分后的平稳序列进一步采用自回归和移动平均模型，模拟一个较好的模型用于对未来序列观察值的预测。对此序列进行一阶差分，得到的一阶差分序列图如图 17。

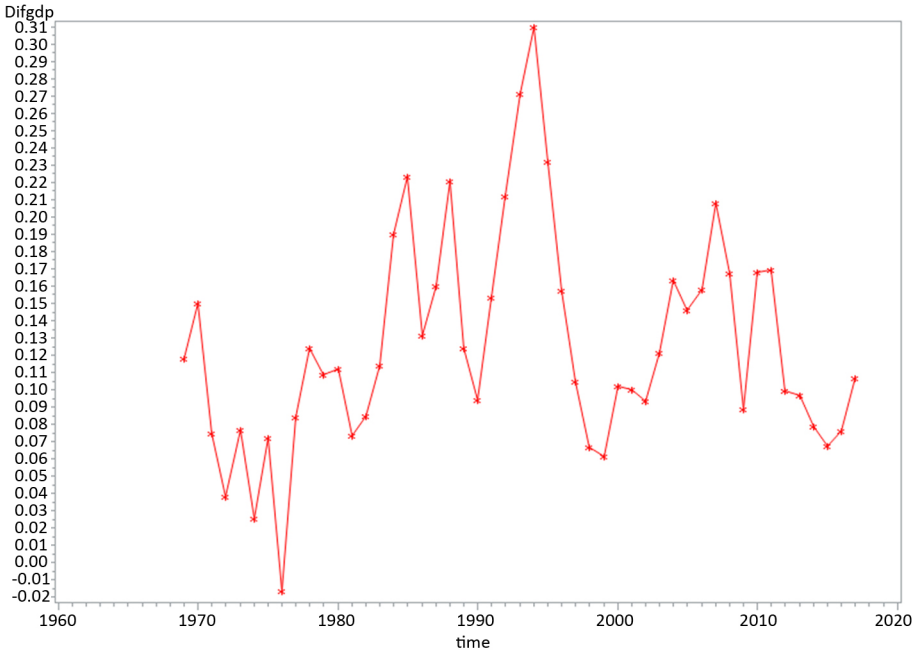

Figure 17. First-order differential sequence diagram
图 17. 一阶差分序列图

观察图 18 可以知道，一阶差分提取了线性增长的趋势，并且时序图已经没有明显的趋势，整个波动比较平稳，所以对差分序列使用 ARIMA 模型。通过差分序列的自相关图和偏自相关图，我们可以发现自相关系数图存在三阶截尾现象，而从偏自相关图可以看出，存在拖尾现象，所以使用 MA(3)模型拟合差分后的序列较为合适。

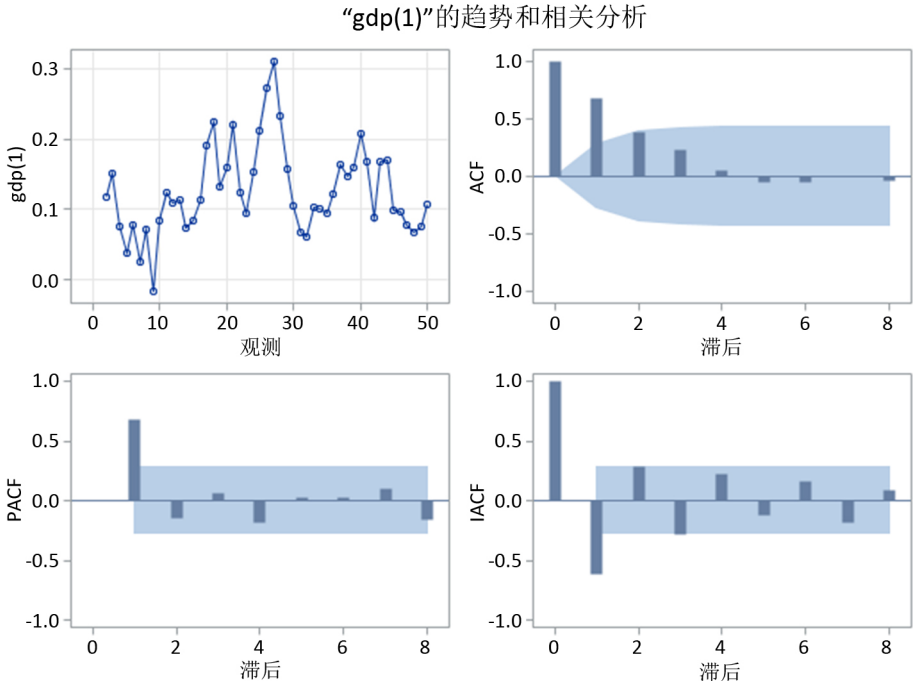

Figure 18. Related diagrams to GDP
图 18. GDP 的相关图表

完成模型定阶后，我们需要检验模型，其中包括两个方面，对模型残差的白噪声检验以及对参数的显著非零检验。对残差进行白噪声检验，看看是否把有效信息提取干净。结果如下：

由于各阶的 P 值都大于 0.05 可以知道，接受 H0 原假设，即认为此序列为白噪声，因此信息提取完毕，残差检验通过。参数的显著性检验也均通过。得到的模型为

$$\nabla_{x_t} = \left(1 + 0.77201B + 0.34853B^2 + 0.30393B^3\right)\varepsilon_t$$

在以预测为目的的回归模型下，如果选用残差平方和作为模型筛选的指标，则只能保证回归的结果整体来看比较靠近真实情况，然而实际上，回归曲线会对数据中间部分进行尽可能靠近的回归，而两边的数据会出现一些偏差，即两边的偏差会有比较大的不确定性，所以在这个情况下，需要使用另外一个模型评判指标。此处使用 AIC，即最小信息量准则[13]。

$$\text{AIC} = -2\ln\left(\text{模型的极大似然函数值}\right) + 2\left(\text{模型中未知的参数个数}\right)$$

结果如表 9。

Table 9. AIC among different models
表 9. 几种模型的 AIC 值比较

	AIC	Rank
多因素线性回归	−134.1414	3
PCA+多因素线性回归	−124.852	4
逐步回归法	−136.1402	2
时间序列法	−157.631	1

表 9 中 AIC 最小的方法效果最佳，因此可以知道使用时间序列是最好的方法。然后是逐步回归，但

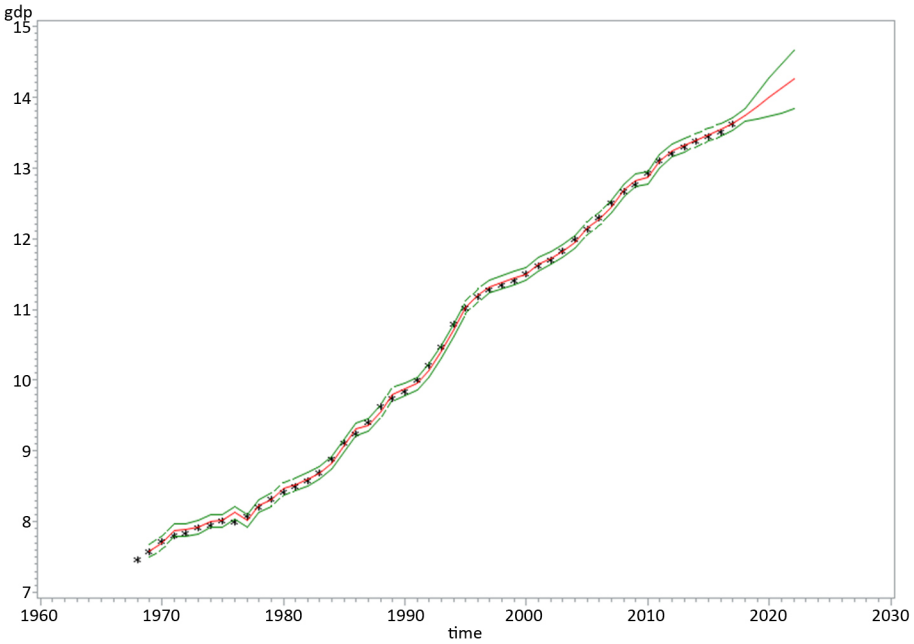

Figure 19. Prediction to GDP
图 19. GDP 预测值

是，如果我们需要使用所有的因素，我们不应该选择逐步回归，因为它会删除一些数据，然而这些被删除的数据在前面已经被证实：它们会对国内生产总值造成影响。至于比较使用主成分分析和没有主成分分析的多元线性回归方法，从 AIC 结果上来看，使用了主成分分析后的结果会较差，经检验，发现前面段的数据偏差比较大，所以出现了 AIC 较大的情况。但实际上，使用主成分分析可以减少共线性问题，让后半段的回归效果变得更好，从而使预测结果更具有信服力。用时间序列分析方法拟合值和 5 期预测值如图 19 所示，其中图中的红色线为 MA(3)时间序列的结果，黑色点为原始时间序列数据，绿色线分别为 95%置信区间的上限。

4. 结论

本文利用相关分析以及线性回归、多项式回归、非线性回归三种方法来探索一个因素与经济指标之间的关系。并通过使用四种方法进行多因素分析，包括直接使用多元线性回归，主成分分析(PCA) + 多元线性回归，逐步回归，时间序列分析，找到一个最为合适的回归方法。得到几点结论：

1) 人口自然增长率与经济指标之间呈多项式线性关系，而且是负相关的关系。这与我们日常认知不相符，主要原因是虽然少儿人口数量呈现下降趋势，但我国由于人口增长所积累的人口红利一直推动着经济的发展。1997~2012 年，我国 15~64 岁年龄段人口数量由 83,448 万人增加至 100,334 万人，其占总人口比例由 67.5%增加至 74.1%。但从 2014 年开始，我国 15~64 岁人口数量开始出现下降，而其占总人口比重则从 2013 年开始出现下降。据预计，劳动人口数量和占比的下降已成趋势。尽管国家目前已经放开了二胎政策，但是未来 15 年内，劳动力下降的整体趋势还不会受二胎政策的影响。未来 10~15 年，我国劳动人数下降、劳动人口占比下降以及劳动人口平均年龄上升将不可避免。为解决此问题，我们需要建立完备的生育制度。由前文的数据可以得出，我国人口经济面临的一个重要问题是生育率的下降。在二孩政策放开后，我国的出生率一直处于稳定，但其中二孩的贡献率占 50%，一孩的贡献率出现下降。当人口红利逐渐消失，我国的经济建设将失去动力。可见，我国需要进一步完善相关的制度鼓励生育。

2) 用以衡量高素质人口的指标——普通本专科毕业人数在主成分分析中可见对经济发展有正向影响作用。因此，提高人口受教育程度应成为长期以来的努力方向，一方面可以引进、吸引受高等教育的人才，另一方面要积极发展本国的高等教育，为经济发展提供人才储备。

3) 与国民消费水平相关的三个变量在方程中的系数相对来说比较高，可以说对经济发展有着极强的正面影响。但事实上消费水平和经济增长可以说是相辅相成的。国家经济能够发展主要依靠着国内的企业的经济效益，而且经济效益提高的同时居民的收入也会随之提高，使得居民消费水平提高，从而扩大居民的内需，进而推动国家的经济发展。可以说在新时代，扩大内需是促进国家经济增长的重要方式。概括来说就是完善消费法和消费政策以促进国民消费。具体做法包括：调整我国经济结构使之符合社会的发展；充分发展社会公益事业和福利事业，减少居民负担，促进居民消费；推动发展社会信用制度以及新型的互联网金融推动国民的消费水平。

4) 时间序列分析和常微分法预测都是通过历史 GDP 趋势进行预测，没有涉及到人口指标。2018 年的 GDP 为 93.18 万亿。而使用常微分法预测 2018 年的值为 96.029 万亿，而真实情况是 90 万亿，可见使用时间序列分析对预测短期 GDP 值可信度较高。因此若要预测 2019 年的数据，为了提高准确性，可以把 2018 年真实数据代进模型进行修正，得出 2019 年的值。

基金项目

广东省大学生创新创业训练项目。

参考文献

[1] 胡鞍钢. 社会与发展——中国社会发展地区差异研究[M]. 杭州: 浙江人民出版社, 2000.

[2] 何晓群, 刘文卿. 应用回归分析[M]. 第 3 版. 北京: 中国人民大学出版社, 2011.

[3] 付凌晖, 王惠文. 多项式回归的建模方法比较研究[J]. 数理统计与管理, 2004, 23(1): 48-52.

[4] Bates, W. 非线性回归分析及其应用[M]. 北京: 中国统计出版社, 1998.

[5] 高慧璇. 应用多元统计分析[M]. 北京: 北京大学出版社, 2005.

[6] 李昕燃, 蒋文浩, 张力文. 基于主成分分析法的宏观经济分析[J]. 现代营销(经营版), 2019(6): 82.

[7] 崔腾飞. 逐步回归分析下山东省税收与 GDP 的关系[J]. 现代商贸工业, 2019, 40(16): 119-120.

[8] 宋平, 邱燕玲. 基于时间序列分析青海省 GDP 预测[J]. 时代金融, 2018(15): 47-48+54.

[9] Hoaglin, D.C., Mosteller, F. and Tukey, J.W. 探索性数据分析[M]. 北京: 中国统计出版社, 2007.

[10] 邓子元. 三年大饥荒原因探析[J]. 农村经济与科技, 2019, 30(1): 55-67.

[11] Pedhazur, E.J. (1982) Multiple Regression in Behavioral Research: Explanation and Prediction. Second Version, Holt, Rinehart and Winston, New York.

[12] Abdi, H. and Williams, L.J. (2010) Principal Component Analysis. *Wiley Interdisciplinary Reviews*: *Computational Statistics*, **2**, 433-459. https://doi.org/10.1002/wics.101

[13] 王燕. 应用时间序列分析[M]. 第 4 版. 北京: 中国人民大学出版社, 2015.