分类号:　　密　　级：公开
学　号：202020118　　单位代码：10764

伊犁师范大学

硕 士 学 位 论 文

基于多重共线性修正下的多元线性回归
——以河南省经济发展影响因素分析为例

Multiple Linear Regression Based on Multiple Cointegration Correction

--Analysis of Economic Development Impact Factors in Henan Province as an Example

学 位 申 请 人	王瑶
指 导 教 师	刘淼 教授
申请学位门类级别	理学硕士
学 科、 专业名称	概率论与数理统计
研 究 方 向	数理统计
所 在 学 院	数学与统计学院

中国·新疆·伊宁
2023 年 5 月

摘 要

河南省 2001～2020 年地区经济发展有了巨大飞跃，经济结构处于高速增长到高质量发展的过渡时期. 在经济转型发展过程中，如何科学分析河南省经济发展影响因素，显得尤为重要. 但针对多个变量建立回归分析模型时，自变量之间容易出现多重共线性现象，从而引起回归系数不稳定、运用普通的最小二乘估计失效等一系列问题，并且筛选后的自变量过少，不能有效解释事物发展的影响因素.

本文通过总结诊断和修正多重共线性的理论方法，提出运用改进的逐步回归、岭回归和主成分-岭估计相关理论能够解决多重共线性引起的回归系数不稳定问题，从而根据回归系数分析河南省地区生产总值的重要影响因素，并建立修正的回归模型预测河南省生产总值.

文章以河南省 2001～2020 年固定生产投资总额等 11 个经济相关的指标数据为例，采用基于t检验的逐步回归筛选自变量，从而达到修正多重共线性的目的；根据修正效果，进一步提出可以运用岭回归和主成分-岭回归进行修正. 首先运用MATLAB定量计算设计矩阵的特征值与特征向量，来具体诊断出自变量间的多重共线性程度；然后借助主成分—岭回归和岭回归的相关理论对地区生产总值的影响因素进行回归分析，得到生产总值关于经济指标的主成分—岭回归方程和岭回归方程，最后根据筛选后的自变量与因变量，分别建立逐步回归方程和岭回归方程，对河南省 2021～2023 年生产总值进行预测，与真实值进行对比，岭回归模型预测精确度最优，研究所得结果为河南省委、省政府以及相关部门制定相应的经济政策和发展策略提供有效的科学依据.

关键词：多重共线性; 改进的逐步回归; 岭回归; 主成分-岭回归; 地区生产总值

Abstract

The economic structure of Henan Province is in a transition period from high growth to high quality development, as the regional economic development has made a great leap from 2001 to 2020. In the process of economic development in transition, it is particularly important to scientifically analyze the factors influencing economic development in Henan Province. However, when establishing regression models between multiple independent variables, multiple covariance between independent variables may occur, resulting in unstable regression coefficients, failure of ordinary least squares estimation, and too few filtered independent variables to effectively explain the influencing factors of development.

This paper summarizes the theoretical methods for diagnosing and correcting multiple covariance and proposes that using improved stepwise regression, ridge regression and principal component-ridge estimation correlation theory can solve the problem of instability of regression coefficients caused by multiple covariance, and according to the regression coefficients can analyze the important influencing factors of regional GDP in Henan Province and correct the multiple covariance, so as to provide help for further analysis and forecasting.

The article takes the data of 11 economic-related indicators in Henan Province from 2001 to 2020 as an example, and screens the variables through stepwise regression based on the t-test, so as to achieve the purpose of correcting the multiple covariance; based on the correction effect, it further proposes to use ridge regression and principal component-ridge regression for correction. The eigenvalues and eigenvectors of the design matrix were first calculated quantitatively using MATLAB to diagnose the degree of multiple covariance; then the regression analysis of the factors influencing the regional GDP was carried out with the help of the relevant theories of principal component-ridge regression and ridge regression. Finally, a stepwise regression equation and a ridge regression equation were established based on the modified independent and dependent variables to forecast the GDP of Henan Province from 2021 to 2023, the results of the study provide an effective scientific basis for the Henan Provincial Party Committee, the provincial government and relevant departments to formulate corresponding economic policies and development strategies.

Keywords: multiple covariance; stepwise regression; ridge estimation; principal component-ridge estimation; regional GDP

目　录

第一章 绪论

1.1 研究背景

多元线性回归模型[1]在经济，农学等众多领域应用广泛，但在实际研究分析中，自变量之间的相关性会造成回归模型不稳定，甚至回归系数的符号不符合实际意义等一些问题.

当自变量之间存在严重多重共线性问题[2]，若直接根据原始数据建立传统回归模型，会由于变量间的多重共线性问题而造成较大误差，或者是在直接使用最小二乘估计时导致精度降低，从而模型的稳定性遭到破坏，不能有效地进行回归分析．因此，研究如何解决多重共线性问题是非常必要的.

到现在为止，修正多重共线性对回归方程影响的理论已经逐步得到发展，并且趋于完善．在这些理论中，有主成分回归，偏最小二乘法等使用较广．在国内外，有不少学者对多重共线性的诊断和处理方法进行了研究，并取得了不错的研究成果．但是，当多个自变量之间存在多重共线性问题，要想消除多重共线性对回归系数和预测的影响，尽可能选择对于因变量影响较大的自变量，还是具有进一步讨论发展的空间.

1.2 研究意义

通过对多重共线性进行修正，可减轻多重共线性给模型带来的危害．本文从常见的逐步回归法入手，针对选择方差扩大因子法诊断多重共线性失效以及筛选变量修正多重共线性不理想问题，提出以有偏估计为基础的岭回归做进一步研究．通过主成分回归法选择影响显著的成分建立模型，提取满足条件的主成分进行回归，处理后得到的主成分之间相互独立，从而消除多重共线性对回归模型的影响；岭回归，是针对设计矩阵呈现病态，运用以无偏估计[3]为本质的逐步回归不能解决较严重的多重共线性问题，牺牲无偏性使方差部分大幅度减小，最终达到降低均方误差的目的[4]．针对主成分回归和岭回归的优点讨论了主成分-岭估计的相关理论知识，并进一步提出对回归模型中多重共线性进行修正步骤及实证过程，从而尽可能保留对因变量影响较大的自变量.

系统总结了特征根判别法，方差扩大因子法，相关阵等相关诊断多重共线性的方法以及通过删除变量，增大样本量，有偏估计等修正多重共线性的方法．实证分析比较关于逐步回归和岭回归模型对于自变量选择的实际应用，为今后学者在实践中使用岭回归法进行实证分析及预测提供参考依据.

借助 MATLAB 等数学统计软件，对河南省经济发展因素进行了排名和对比，根据各个因素结合实际进行分析，并通过回归方程进行回归预测．根据分析结果提出的相关有效建议，所得结果为河南省委、省政府以及相关部门制定相应的经济政策和发展策略提供有效的科学依据.

1.3 国内外研究现状

1934 年 Frisch 首先引入多重共线性概念[5]，在研究多元线性回归模型时，多个自变量之间不可避免会存在近似的线性关系，学者们对其进行过大量的探讨，但也产生了不同的看法．Farrar 和 Glauber[6]对此作了总结，认为多重共线性不影响 LS 估计的 BLUE 性质．而另外一些学者认为多重共线性造成的后果不容忽视，需要修正它．肖霞[7]选择从几何方面解释线性回归中的多重共线性，采用分解定理发现多重共线性一方面是由因变量与自变量之间的整体结构，也可能是样本选择的结果，学者需要明确多重共线性原因及危害之后，再进行多重共线性的修正.

在线性回归中，逐步回归是一种常见的分析方法，当线性回归模型中自变量较多，如何选择变量，从而使回归方程中包含对自变量影响作用显著的解释变量．学者在理论和实际应用中做了相应研究，刘明[8]提出在普通最小二乘估计下，逐步回归的 F 检验和对应自变量的t检验是等价的，运用t检验的显著性完成逐步回归；肖雪梦[9]等人提出分别采用逐步回归，主成分回归以及偏最小二乘回归对同一组数据做建模分析，修正自变量之间的多重共线性问题，并作预测分析.

在运用回归模型对多重共线性的修正过程中，我们把参数估计作为主要研究对象．最小二乘估计具有“一切线性无偏估计中有最小方差”的优良性[10]及其相关的理论研究比较成熟等优势使它在众多估计的研究发展中有了举足轻重的地位，随着科技发展，研究者的深入研究使人们能够处理包含较多自变量的大型回归问题.

因此当自变量较多时，自变量之间不可避免会存在多重共线性问题，从而导致设计矩阵出现病态，显然再用最小二乘估计不是一个最优估计．所以我们需要考虑选择有偏估计来解决．有偏估计中影响最大的有岭估计，偏最小二乘估计和主成分估计，但遗留下来的问题也较多，针对不同的回归模型之间存在的问题，学者们进行了深入的探讨和研究.

主成分回归是指原始数据经过正交旋转处理，把多数原始变量转化为几个综合变量的线性组合，保证各个主成分间相互独立．这样能够将众多的复杂指标转换为少数

几个互不相关的成分，尽可能反映出原来指标信息的综合性指标，从而消除多重共线性的影响．孔朝莉[11]等人运用线性主成分回归模型和对数主成分模型定量分析了六个典型行业在海南省经济发展的关系.

由于多重共线性导致*LS*估计显著性显著下降，霍尔对最小二乘估计进行改进，提出一种有偏估计—岭估计，肯纳德[12]也给予了详细讨论．后人跟随其脚步，做了相关深入研究．尤游[15]等人采用特征值判定法对*LS*估计模型进行多重共线性诊断，再运用岭回归模型和主成分回归模型两种方法进行改进，并将改进后的预测结果进行相对误差分析，得到岭回归模型为较优模型；王飞[14]等人简述了多重共线性的相关理论，并采用岭回归分析法修正实证过程中出现的多重共线性问题；丁先文[15]等人通过最小二乘估计与岭估计讨论了江苏省财政收入及影响因素，通过与普通线性回归分析比较，岭参数估计具有更加准确预测精度；林乐义[16]提出了可以通过岭回归消除多重共线性，解决回归系数与实际不相符问题．岭回归应用下，根据删除后的自变量与因变量进行回归，预测值更加贴合实际值.

综上所述，逐步回归、主成分回归及岭回归能够有效地分析经济社会发展的影响因素．逐步回归是通过删除变量来修正多重共线性，但得到的方程变量较少，不能有效地分析多个影响因素．而主成分回归[17]，在对经济影响因素进行评价分析时，会舍弃贡献率较小的主成分，以含有原始信息15%以下标准，损失原始信息，可能会影响实际问题的分析判断.

因此本文在运用逐步回归的基础上，采用岭回归修正多重共线性，借助岭迹图能够直观且快速判断自变量之间的相关性，且通过岭计算尽可能保留较多对经济影响较大的自变量．并根据主成分回归，岭回归的优缺点，进一步提出能够运用主成分—岭回归修正多重共线性，并总结出相关理论以及实证步骤.

1.4 本文的主要研究思路

本文通过选取河南省2001～2020年经济发展具有代表性的11个经济指标数据（相关数据均来自《河南省统计年鉴》），对河南省经济发展影响因素进行回归建模，运用改进的逐步回归、主成分-岭回归和岭回归理论修正其存在的多重共线性问题，对比得出最佳的修正模型；并根据修正后得到的回归系数，客观分析河南省经济发展主要影响因素，并提出了解决河南省地区经济发展不平衡性的政策建议；运用时间序列模型预测

各个指标值及生产总值，代入岭回归方程，预测河南省生产总值，与筛选的逐步回归对比，得出相对最优的预测模型.

1.5 本文的创新点

（1）本学位论文分别对河南省经济发展影响因素：物质资本投入，人力资源因素，技术进步因素等因素建立主成分-岭回归，通过回归系数方法确定河南省经济发展重要影响因素，并检验模型对多重共线性的修正效果.

（2）根据时间序列对河南省经济发展的多个影响因素分别进行预测，并通过代入岭回归方程来预测河南省生产总值，得到的预测结果与直接运用时间序列相比精确度更高.

第二章 多重共线性问题及其诊断

2.1 多重共线性的含义

建立多元线性回归模型需要满足的充要条件为设计矩阵 X 的秩 $rank(X)=k+1$，即表示样本矩阵 X 的列向量之间线性无关．但是在经济分析中，比如研究经济问题，需要对多个因素进行考虑．但由于事物的复杂性，大多数因素之间存在着一定的相关性．一般地，自变量相关关系较弱的情况被看作是符合回归模型的建模要求；但当自变量间具有较强的相关性，认为回归模型是不符合基本假设，并称该模型存在多重共线性问题[4]．

在实际分析过程中，较常见的是线性关系近似成立的情况，即存在一组常数 $c_0,c_1,c_2,\cdots c_k$ 不全为零，使得

$$c_0+c_1x_{i1}+c_2x_{i2}+\cdots+c_kx_{ik}\approx 0, i=1,2,\cdots,n, \qquad (2.1)$$

当自变量 $x_1,x_2,\cdots,x_k$ 存在式（2.1）所示的关系，称自变量 $x_1,x_2,\cdots,x_k$ 之间存在多重共线性．

通常可以用简单相关系数初步判断归模型自变量存在三种关系[18]：

（1）$r_{x_ix_j}=0$，即 x_i 与 x_j 毫无线性关系．

（2）$r_{x_ix_j}=1$，即 x_i 与 x_j 完全共线．

（3）$0<\left|r_{x_ix_j}\right|<1$，自变量之间存在一定程度的线性关系．

2.2 多重共线性的产生背景

由于事物总是处于相互联系的，通常模型自变量存在多重共线性主要有以下三点原因[19]：

（1）变量代表的实际含义确定了它们之间的相关关系，在对经济问题展开研究时，会因为经济变量随时间的共同变化趋势，而出现共线性问题；

（2）在利用横截面数据建立因变量和自变量的回归模型时，也会出现一些自变量具有高度相关的情况．在这种情况下，与事物发展相关的变量会呈现共同变化的趋势；

（3）样本数据本身的原因．在对影响因素进行综合分析时，为了全面分析事物，所以我们往往会尽可能地多选取有关变量，从而造成严重的多重共线性问题．

2.3 多重共线性的影响及危害

2.3.1 多重共线性的影响

前面我们给出多重共线性的定义，下面具体说明多重共线性对回归方程的影响[20]：

对实际问题进行研究，通常存在$c_0,c_1,c_2,\cdots,c_k$不全为零，使得

$$c_0+c_1x_{i1}+c_2x_{i2}+\cdots c_kx_{ik}\approx 0,i=1,2,\cdots,n, \tag{2.2}$$

即近似共线性的情形，此时设计矩阵X的秩$rank(X)=k+1$虽然成立，但是$|X^TX|\approx 0$.

首先要知道$D(\hat{\beta})$的对角线元素为$Var(\hat{\beta}_0),Var(\hat{\beta}_1),\cdots,Var(\hat{\beta}_k)$，表示回归系数的方差．当$(X^TX)^{-1}$的对角线元素较大时，$D(\hat{\beta})=\sigma^2(X^TX)^{-1}$的对角线很大，从而导致$\beta_0,\beta_1,\beta_2,\cdots,\beta_k$的估计精度降低．这时，用普通最小二乘估计计算估计量$\hat{\beta}$，得到估计值方差很大，这将会使自变量对因变量的作用产生影响，不能合理地对估计量的经济意义做出解释.

2.3.2 多重共线性的危害

由于多重共线性的影响，可能会使得模型中的参数估计变为不定式形式，从而导致参数估计值的方差变成无穷大.

（1）参数估计值的方差增大[20]

我们以二元回归为例，当自变量的相关性从小增大时，估计量的方差增大得很快.

y与x_1,x_2都已经标准化，这时回归常数项为零，我们建立自变量x_1,x_2与因变量y的回归方程为：

$$\hat{y}=\hat{\beta}_1\hat{x}_1+\hat{\beta}_2x_2, \tag{2.3}$$

记$L_{11}=\sum_{i=1}^{n}x_{i1}{}^2$，$L_{12}=\sum_{i=1}^{n}x_{i1}x_{i2}$，$L_{22}=\sum_{i=1}^{n}x_{i2}{}^2$，则$x_1$与$x_2$之间的相关系数为

$r_{12}=\dfrac{L_{12}}{\sqrt{L_{11}}\sqrt{L_{22}}}$.

令$X=\begin{pmatrix}x_{11} & x_{12}\\ x_{21} & x_{22}\\ \vdots & \vdots\\ x_{n1} & x_{n2}\end{pmatrix}$，则

$$X^T X = \begin{pmatrix} x_{11} & x_{21} & \cdots & x_{n1} \\ x_{12} & x_{22} & \cdots & x_{n2} \end{pmatrix} \begin{pmatrix} x_{11} & x_{12} \\ x_{21} & x_{22} \\ \vdots & \vdots \\ x_{n1} & x_{n2} \end{pmatrix}$$

$$= \begin{pmatrix} \sum_{i=1}^{n} x_{i1}^{\ 2} & \sum_{i=1}^{n} x_{i1}x_{i2} \\ \sum_{i=1}^{n} x_{i2}x_{i1} & \sum_{i=1}^{n} x_{i2}^{\ 2} \end{pmatrix} \tag{2.4}$$

$$= \begin{bmatrix} L_{11} & L_{12} \\ L_{12} & L_{22} \end{bmatrix}.$$

根据矩阵的计算得到：

$$\left|X^T X\right| = \sum_{i=1}^{n} x_{i1}^{\ 2} \sum_{i=1}^{n} x_{i2}^{\ 2} - \left(\sum_{i=1}^{n} x_{i1}x_{i2}\right)^2 = L_{11}L_{22} - L_{12}^{\ 2}. \tag{2.5}$$

因此可以得到：

$$\left(X^T X\right)^{-1} = \frac{1}{\left|X^T X\right|}\left(X^T X\right)^* = \frac{1}{L_{11}L_{22} - L_{12}^{\ 2}} \begin{bmatrix} L_{22} & -L_{12} \\ -L_{12} & L_{11} \end{bmatrix}, \tag{2.6}$$

于是能得到回归系数的方差表达式：

$$Var\left(\hat{\beta}_1\right) = \frac{\sum_{i=1}^{n} x_{i2}^{\ 2}}{\sum_{i=1}^{n} x_{i1}^{\ 2} \sum_{i=1}^{n} x_{i2}^{\ 2} - \left(\sum_{i=1}^{n} x_{i1}x_{i2}\right)^2} \sigma^2$$

$$= \frac{L_{22}}{L_{11}L_{22} - L_{12}^{\ 2}} \sigma^2 \tag{2.7}$$

$$= \frac{\sigma^2}{\left(1 - r_{12}^{\ 2}\right) L_{11}}.$$

同理可得，$Var\left(\hat{\beta}_2\right) = \frac{\sigma^2}{\left(1 - r_{12}^{\ 2}\right) L_{22}}$.

从上面两个式子可以看出，当自变量 x_1 与 x_2 之间相关性越高，$\hat{\beta}_1$ 和 $\hat{\beta}_2$ 的方差系数越大．其中，称 $VIF = \frac{1}{1 - r_{12}^{\ 2}}$ 为方差扩大因子系数.

这是二元回归的情况，自变量只有两个，以此类推，能够得到多元线性回归的方差性质.

（2）参数区间估计的置信区间趋于∞.

因为 $\hat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma^2}{\sum_{i=1}^{n} x_{i1}^{\ 2}} VIF\right)$，所以对前面 $\hat{\beta}_1$ 的分布表达式进行正态化，为

$\frac{\hat{\beta}_1-\beta_1}{\sigma/\sqrt{\sum_{i=1}^{n} x_{i1}^{\ 2}\left(1-r_{12}^{\ 2}\right)}} \sim N(0,1)$，于是得到

$$\frac{S_E^{\ 2}}{\sigma^2} \sim \chi^2(n-2). \tag{2.8}$$

我们选取 t 统计量表达式为：

$$t(n-2)=\frac{\frac{\hat{\beta}_1-\beta_1}{\sigma/\sqrt{\sum_{i=1}^{n} x_{i1}^{\ 2}\left(1-r_{12}^{\ 2}\right)}}}{\frac{S_E^{\ 2}}{\sigma^2(n-2)}}=\frac{\hat{\beta}_1-\beta_1}{\hat{\sigma}/\sqrt{\sum_{i=1}^{n} x_{i1}^{\ 2}\left(1-r_{12}^{\ 2}\right)}}. \tag{2.9}$$

这里取置信水平为 $1-\alpha$ 的置信区间，即：

$$P\left(\frac{\left|\hat{\beta}_1-\beta_1\right|}{\hat{\sigma}/\sqrt{\sum_{i=1}^{n} x_{i1}^{\ 2}\left(1-r_{12}^{\ 2}\right)}} < t_{1-\frac{\alpha}{2}}(n-2)\right)=1-\alpha. \tag{2.10}$$

得到 $\hat{\beta}_1$ 的置信区间为：

$$\hat{\beta}_1-\frac{\hat{\sigma}}{\sqrt{\sum_{i=1}^{n} x_{i1}^{\ 2}\left(1-r_{12}^{\ 2}\right)}} t_{1-\frac{\alpha}{2}}(n-2) < \beta_1 < \hat{\beta}_1+\frac{\hat{\sigma}}{\sqrt{\sum_{i=1}^{n} x_{i1}^{\ 2}\left(1-r_{12}^{\ 2}\right)}} t_{1-\frac{\alpha}{2}}(n-2), \tag{2.11}$$

因此，$\hat{\beta}_1$ 的置信区间长度为 $\frac{2\hat{\sigma}}{\sqrt{\sum_{i=1}^{n} x_{i1}^{\ 2}\left(1-r_{12}^{\ 2}\right)}} t_{1-\frac{\alpha}{2}}(n-2)$.

当自变量 x_1 与 x_2 的相关程度增加， r_{12} 增大， r_{12}^2 增大， $1-r_{12}^2$ 减小，则 $\hat{\beta}_1$ 的置信区间长度增大，当 $r \to 1$ 时，置信区间趋于 ∞， $\hat{\beta}_2$ 同理.

当自变量之间存在着严重的多重共线性问题时，假设检验可能会导致错误的判断，造成 R^2 较高；即使参数经过 F 检验，但对各个参数进行 t 检验可能不显著，甚至会出现符号相反的情况.

2.4 多重共线性的诊断

近年来，如何对多重共线性进行诊断，以及计算共线性的严重程度，成为统计学者们讨论的热门问题. 本节系统总结了常见诊断多重共线性的方法，以供做进一步研究[21].

2.4.1 方差扩大因子

对自变量做无量纲化处理，令 $\left(X^*\right)^T X^* = \left(r_{ij}\right)$ 为自变量的相关阵. 记

$$C = \left(c_{ij}\right) = \left(X^{*T} X^*\right)^{-1}, \tag{2.12}$$

称主对角线元素 $VIF = c_{jj}$ 为自变量 x_j 的方差扩大因子.

通过式 $D\left(\hat{\beta}\right) = \sigma^2 \left(X^T X\right)^{-1}$ 可得

$$Var\left(\hat{\beta}_j\right) = C_{jj}\sigma^2 / L_{jj}, j = 1,2,\cdots,k, \tag{2.13}$$

其中 L_{jj} 为 x_j 的离差平方和. 设 R_j^2 为自变量 x_j 对其他变量的复决定系数，记 $c_{jj} = \dfrac{1}{1-R_j^2}$ 为 VIF_j 的定义.

我们通常用 R_j^2 来衡量自变量之间的线性相关程度， $R_j^2 \to 1$ 时， VIF_j 就越大. 从这个角度，我们可以用 VIF_j 值来反映了自变量之间存在多重共线性的程度. 通常情况下， $VIF_j \geq 10$ 表明在最小二乘估计中，各自变量间存在较强的共线性问题，而且会对 LS 估计值产生较大影响.

2.4.2 特征根诊断多重共线性

（1）特征根判别法

当行列式$\left|X^TX\right|\approx 0$时，矩阵$X^TX$对应的行列式至少有一个特征根近似为 0. 即当$\left|X^TX\right|$近似为 0，数据矩阵$X$经过标准化，它的列向量必然存在多重共线性.

记$X=\left(X_0,X_1,\cdots,X_k\right)$，$X_i\left(i=0,1,\cdots,k\right)$为$X$的列向量. 这里$X_0=(1,1,\cdots,1)^T$，$\lambda$为设计矩阵$X^TX$的一个特征根，且近似为 0.

当$\lambda\approx 0$时，其对应的单位特征向量为$C=\left(c_0,c_1,\cdots,c_k\right)^T$，根据矩阵的相关理论，得到$X^TXC=\lambda C\approx 0$，我们在两边分别乘以一个$C^T$，得到$C^TX^TXC\approx 0$，从而有$XC\approx 0$，即

$$c_0+c_1x_{i1}+c_2x_{i2}+\cdots+c_kx_{ik}\approx 0\left(i=1,2,\cdots,n\right), \tag{2.14}$$

对于一个实际问题，获得n组数据，则线性回归模型式（2.1）可以表示为

$$y_i=\beta_0+\beta_1x_{i1}+\beta_2x_{i2}+\cdots+\beta_kx_{ik}+\varepsilon_i. \tag{2.15}$$

我们知道$C=\left(c_0,c_1,\cdots,c_k\right)\neq 0$，可以推出$1,X_1,X_2,\cdots,X_k$线性相关，从而得到$X_1,X_2,\cdots,X_k$线性相关. 这样我们能够诊断出该模型存在多重共线性以及多重共线性的个数.

（2）病态指数（条件数）

通过设计矩阵X^TX特征根接近零的数量，可以判断X中存在的多重共线性关系个数. 下面介绍用条件数来确定特征根近似为零的程度. 记X^TX的最大特征根为λ_m，我们称

$$k_i=\sqrt{\frac{\lambda_m}{\lambda_i}},i=0,1,2,\cdots,k, \tag{2.16}$$

为特征根λ_i的条件数.

条件数度量了矩阵X^TX特征根的散布程度，可以用来判定多元线性回归模型的多重共线性，判定规则如下:

（1）当$0<k<10$时，X不存在共线性问题；

（2）当$10\leq k<100$时，X可能存在较强的共线性问题；

（3）$k\geq 100$时，则X存在严重的多重共线性.

2.5 修正多重共线性问题常见方法

我们通过以上几种诊断方法，结合定量和定性分析，判断出回归方程存在多重共线性．这里从诊断入手，来解决多重共线性问题.

(1)剔除一些不重要的解释变量

针对社会和经济问题进行建模时，由于对其认知程度的限制，会将太多的自变量纳入到模型中．最常见的方法是，做自变量的选元，将一部分不满足显著性条件的自变量舍弃掉．在回归方程的所有自变量都经过了显著性检验之后，如果在回归方程中出现了严重的多重共线性，并且有好几个自变量方差扩大因子大于等于 10，我们将 $k_{\max}$ 的自变量依次剔除掉，然后重新构建回归方程，直到回归方程中不再出现严重的多重共线性问题；或者是按照实际重要性来确定是否保留或删除这些变量.

（2）增大样本量

在建立回归模型时，若收集的样本数据较少，则会引起共线性问题．我们举一个简单例子，假设 x_1 和 x_2 都已经中心化，$\hat{\beta}_1, \hat{\beta}_2$ 的方差计算公式为：

$$Var\left(\hat{\beta}_1\right)=\frac{\sigma^2}{\left(1-r_{12}{}^2\right)L_{11}}, \tag{2.17}$$

$$Var\left(\hat{\beta}_2\right)=\frac{\sigma^2}{\left(1-r_{12}{}^2\right)L_{22}}, \tag{2.18}$$

式中，r_{12} 为 x_1 和 x_2 的相关系数，$L_{11}=\sum_{i=1}^{n} x_{i1}{}^2$，$L_{22}=\sum_{i=1}^{n} x_{i2}{}^2$.

根据上式，当相关系数固定时，样本容量 n 增加，会引起 L_{11} 和 L_{22} 变大，使得回归系数的方差 $Var(\hat{\beta}_1), Var(\hat{\beta}_2)$ 减小，从而削弱了共线性对回归系数的影响．运用回归分析研究经济实际问题时，在条件允许的情况下，可以尽可能使样本个数 n 远大于自变量个数 k，这样一定程度上会减少共线性的程度[4].

（3）回归系数的有偏估计

在回归模型中，消除多重共线性的影响，是统计学家所关心的一个热门问题．除了上述的一些方法得到了广泛地使用之外，统计学家也在努力地对 LS 估计进行改进，运用有偏估计，如岭回归、主成分和偏最小二乘等来增加估计量的稳定性.

第三章 修正多重共线性问题的多元回归模型

考虑线性回归模型：$y=X\beta+e,E(e)=0,Cov(e)=\sigma^2 I_n$.

其中，y 为 $n\times1$ 随机观测向量，X 为 $n\times p$ 的设计矩阵且已中心化和标准化，$rank(X)=k$，β 为 $p\times1$ 的未知参数向量，e 为 $n\times1$ 随机误差向量，I_n 为 n 阶单位矩阵.

3.1 基于筛选变量的改进的逐步回归

3.1.1 逐步回归法概述

在逐步回归法[22]中，引入和删除的自变量显著性水平 α 值不相同，且引入自变量时所需的显著性水平 α_{enter} 应低于删除自变量所需的显著性水平 α_{remoal}.

逐步回归中是通过 F 检验来引入自变量，根据多重共线性问题，逐步回归的 F 检验显著性作用不明显，进一步提出改进的逐步回归[23].

3.1.2 改进的逐步回归法筛选变量

改进的逐步回归是在传统的逐步回归基础上，提出基于 t 检验的逐步回归. 因为在自变量之间存在多重共线性时，仅依靠 R^2，F 是不够的，t 检验能够检验因变量与各个自变量的显著性，并且删除不符合显著性的自变量修正多重共线性[24].

改进的逐步回归法步骤如下：

（1)根据理论与经验分析，自变量有 $X_1,X_2,\cdots,X_m$，Y 是被研究对象的事物，即因变量. 设有 n 个样本观测向量（例如选取 2001～2020 年共 20 年观测值，n=20），m 项观测指标（与研究对象相关的自变量），相关数据来自于河南省统计年鉴[25].

（2）数据的无量纲化处理[26]. 均值处理：

$$y_{ij}=\frac{x_{ij}}{\overline{x_j}},\overline{x_{ij}}=\frac{1}{n}\sum_{i=1}^{n}x_{ij},\frac{1}{n}\sum_{i=1}^{n}y_{ij}=1. \tag{3.1}$$

（3）根据相关系数筛选变量. 通过相关系数矩阵，分析各自变量间的线性相关性，确定自变量之间存在的多重共线性程度，同时删除与因变量相关系数较小的变量.

（4）作 Y 与 $X_1,X_2,\cdots,X_m$ 的多元线性回归方程：

$$\hat{Y}=\hat{c}+\hat{\beta}_1X_1+\hat{\beta}_2X_2+\cdots+\hat{\beta}_mX_m. \tag{3.2}$$

一般 R^2， F 统计量在方程中较容易通过拟合度和显著性检验．主要是需要对回归系数进行 t 检验，删去回归不显著的自变量（ $sig<0.05$ ），依次重新建立回归方程，直到每个自变量都通过 t 检验.

（5）当自变量与因变量线性显著，通过 VIF 检验多重共线性的修正效果．逐步回归经过改进，剔除了模型中影响性较小且相关性严重的变量来消除多重共线性，并且将那些重要且相关关系轻微的变量保留了下来，以便做进一步分析.

3.2 岭回归模型修正多重共线性问题

当自变量之间存在共线性问题，运用无偏估计得到的回归系数 $\hat{\beta}_j$ 不稳定，会导致 $Var(\hat{\beta}_j)=C_{jj}\sigma^2/L_{jj}$ 很大，使得回归精度降低．因此，下面提出运用岭估计这一有偏估计来修正共线性问题.

3.2.1 岭回归基本概念

（1）岭回归概念的提出

当自变量间出现多重共线性，即 $\left|X^TX\right|\approx 0$．我们通过给设计矩阵 X^TX 加上一个矩阵 kI $\left(k>0,I\text{为单位矩阵}\right)$，使得 X^TX+kI 接近奇异的程度比 X^TX 接近奇异的程度小．将设计矩阵标准化，并用 X 表示，岭回归估计[27]为：

$$\hat{\beta}(k)=(X^TX+kI)^{-1}X^TY\text{，}\qquad(3.3)$$

k 称为岭估计参数．标准化方阵 X^TX 为数据矩阵 X 的样本相关阵．如果 Y 也经过标准化处理，式（3.3）称为标准化岭回归估计，岭回归估计 $\hat{\beta}(0)$ 的情况为特例，即指普通最小二乘估计.

$\hat{\beta}(k)$ 实际是回归参数 β 的一个估计族，但需要根据平稳点跟实际情况进行分析判断，确定最优参数.

（2）岭回归的相关性质

性质 3.1[28] $\hat{\beta}(k)$ 是回归参数 β 的有偏估计，即

$$E[\hat{\beta}(k)]=E((X^TX+kI)^{-1}X^Ty)=(X^TX+kI)^{-1}X^TX\beta\text{.}\qquad(3.4)$$

其中， $E(\hat{\beta})=E(\beta+(X^TX)^{-1}X^T\varepsilon)=\beta$ ，且 $\hat{y}=X\hat{\beta}$，从而有

$E(\hat{y})=E(X\hat{\beta})=XE(\hat{\beta})=X\beta=E(y)$. 说明$k=0$时，$E[\hat{\beta}(0)]=\beta$；当$k\neq 0$时，$\hat{\beta}(k)$是$\beta$的有偏估计[29].

性质 3.2[30] 当假定岭参数k为常数，且与y无关，那么$\hat{\beta}(k)=(X^TX+kI)^{-1}X^Ty$可以看作是最小二乘估计$\hat{\beta}$的一个线性变换或者线性函数. 表达式为：

$$\begin{aligned}\hat{\beta}(k)&=(X^TX+kI)^{-1}X^Ty\\&=(X^TX+kI)^{-1}X^TX(X^TX)^{-1}X^Ty\\&=(X^TX+kI)^{-1}X^TX\hat{\beta}.\end{aligned}\tag{3.5}$$

性质 3.3 以MSE表示估计向量的均方误差[30]，则存在$k>0$，使得$MSE[\hat{\beta}(k)]<MSE(\hat{\beta})$，即$\sum_{j=1}^{k}E[\hat{\beta}_j(k)-\beta_j]^2<\sum_{j=1}^{k}D(\hat{\beta}_j)$.

3.2.2 选择岭估计参数K的方法

我们需要选择使$MSE(\hat{\beta}(k))$达到最小的k值，它的最优取值取决于未知参数β和k，所以在实际分析计算中，必须根据样本来确定[31].

（1）岭迹法

岭估计$\hat{\beta}(k)=(X^TX+kI)^{-1}X^TY$的分量$\hat{\beta}_i(k)$作为$k$的函数，取值为$[0,+\infty)$，坐标系下根据$\hat{\beta}_i(k)$取值变化绘制的图形称为岭迹图.

岭参数k值的选择原则：

（i）当回归方程各个回归系数的岭估计基本稳定，趋势变化没有明显的拐点，可以选择此时的岭参数.

（ii）用最小二乘法得到的回归系数与实际意义不一致，而其运用岭估计的正负号趋于合理，基本与其实际意义相符.

（iii）残差平方和增加不太多.

在采用岭迹图法选取最优k值时，如果$\hat{\beta}_i(0)$很大，且随着参数k增大，$\hat{\beta}_i(0)$迅速趋于 0，则该变量影响作用不大，应给予剔除；如果$\hat{\beta}_i(0)$很小，但随着参数k增大，

$|\hat{\beta}_i(0)|$迅速增加，则该变量对因变量Y影响作用较大[32]．采用岭迹法确定k值，有助于实现定性与定量分析的有机结合.

对于岭估计$\hat{\beta}(k)$，它的协方差阵为：

$$\begin{aligned} D(\hat{\beta}(k)) &= \operatorname{cov}(\hat{\beta}(k),\hat{\beta}(k)) \\ &= \operatorname{cov}((X^TX+kI)^{-1}X^Ty,(X^TX+kI)^{-1}X^Ty) \\ &= (X^TX+kI)^{-1}X^T\operatorname{cov}(y,y)X(X^TX+kI)^{-1} \\ &= (X^TX+kI)^{-1}X^T\operatorname{cov}(y,y)X(X^TX+kI)^{-1} \\ &= (X^TX+kI)^{-1}X^T\operatorname{cov}(y,y)X(X^TX+kI)^{-1} \\ &= \sigma^2(X^TX+kI)^{-1}X^TX(X^TX+kI)^{-1} \\ &= \sigma^2c(k). \end{aligned} \tag{3.6}$$

其中$c_{ij}(k)$是岭估计的方差扩大因子，并且会随着k的增大而减小，这个系数可以用来判断多元线性模型的自变量共线性程度．一般的，$c_{ij}(k)\geq 10$，则判定自变量之间具有极为严重的多重共线性，而方差扩大因子就是通过选择参数k，使得$c_{ij}(k)<10$，此时的岭参数k相对来说是较稳定的[33].

（2）控制残差平方和法

控制残差平方和是利用岭估计$\hat{\beta}(k)$的残差平方和不超过$cSSE$来找到最大值.$\left(\text{其中}c>1\text{为指定的常数，SSE为最小二乘估计的残差平方和}\right)$将岭回归的表达式代入上述中，我们可以得到$SSE(k)<cSSE$，即寻找使上式成立的最大的$k$值.

3.2.3 岭回归模型修正多重共线性

在选择变量时，岭回归判定原则[4]是：

（1）岭回归计算是通过中心归一化设计矩阵X，对岭回归系数进行直接对比，将那些相对稳定且绝对值很小的岭回归系数剔除掉.

（2）当k值较小，归一化系数的绝对值较大，但是随着k的增大，回归系数迅速趋于零.

（3）若有几个岭回归系数是不稳定的，需要在剔除某个变量后岭回归的影响程度才能决定[34].

岭回归的基本步骤：

（1）数据标准化处理[26]．计算公式为 $X_{ij}^{*}=\dfrac{X_{ij}-\overline{X_j}}{s_j}$，其中 s_j 为样本标准差.

（2）求标准化的设计矩阵 X^{*}，消除量纲的影响．令设计矩阵 $X=(a_{ij})$，$(i=0,1,2,\cdots n,\ j=1,2,\cdots n)$，其中 $(a_{0j})=(1,1,\cdots,1)$.

（3）用条件数判断存在的多重共线性问题[35]．将原始数据标准化处理得到 X^{*}，利用软件得到 $X^{*T}X^{*}$ 的值 $X^{*T}X^{*}$，求得这个对称方阵 $X^{*T}X^{*}$ 的特征根 $\lambda_1,\lambda_2,\cdots,\lambda_p$ 和特征向量 $\varphi_1,\varphi_2,\cdots,\varphi_p$；然后用特征向量每行数值的平方和除以特征根，最后用得到的每列数据除以每列数据的和，得到方差比例表．这个方差比例表是由对称方阵 $X^{*T}X^{*}$ 的特征根和特征向量来决定，特征值越小，方差就越大，以此分析多重共线性的严重程度.

（4）绘制岭迹图[36]，选取最优 K 值．当自变量的归一化系数趋于平稳时，选取最小 K 值．通常 K 值越小，偏差越小．岭回归分析的重点在于寻找最优的 K 值．岭回归估计可以减少均方误差，优于 LS 估计.

（5）通过确定的 K 值进行回归估计，得到回归系数，建立岭回归方程．将经过时间序列[37]预测的各个指标值，代入建模得到的回归方程，进行因变量的预测值.

3.3 主成分—岭回归模型修正多重共线性问题

3.3.1 主成分-岭回归的提出

考虑线性回归模型为:

$$y=X\beta+\varepsilon,E(\varepsilon)=0,Cov(\varepsilon)=\sigma^2 I_n. \tag{3.7}$$

其中，y 为 $n\times1$ 的观测向量，X 为 $n\times p$ 中心标准化设计矩阵，$rank(X)=p$，β 是 $p\times1$ 的未知参数向量，ε 是 $n\times1$ 随机误差向量，I_n 是单位矩阵.

存在 $p\times p$ 矩阵 Φ，使 $X^TX=\Phi\Lambda\Phi^T$，其中 $\Lambda=diag(\lambda_1,\lambda_2,\cdots,\lambda_p)$，$\Phi=(\varphi_1,\varphi_2,\cdots,\varphi_p)$，$\lambda_1\geq\lambda_2\geq\cdots\geq\lambda_p$ 是设计矩阵 X^TX 的特征值，$\varphi_1,\varphi_2,\cdots,\varphi_p$ 为对应的标准化正交化特征向量.

令 $Z=X\Phi,\alpha=\Phi^T\beta$，则

$$y=Z\alpha+\varepsilon,E(\varepsilon)=0,Cov(\varepsilon)=\sigma^2 I_n. \tag{3.8}$$

得到α的LS估计为$\hat{\alpha}=(Z^TZ)^{-1}Z^Ty=\Lambda^{-1}Z^Ty$；$\beta$的$LS$估计为$\hat{\beta}=\Phi\hat{\alpha}$.

当X^TX存在很小的特征值，模型出现病态，假设$\lambda_{r+1},\cdots,\lambda_p\approx 0$，此时我们能够知道均方误差非常大，即

$$MSE(\hat{\beta})=\frac{\sigma^2\sum_{i=1}^{p}1}{\lambda_i}. \tag{3.9}$$

因而LS估计在均方误差意义下不是一个好的估计．而岭估计是通过均值的有偏性使得方差部分的大幅减小，最终达到降低均方误差的目的．但LS估计变坏的根本原因在于$\lambda_{r+1},\cdots,\lambda_p$很小，岭回归是通过增大$\lambda_{r+1},\cdots,\lambda_p$的值，来修正多重共线性问题．主成分估计，虽然后面$p-r$个主成分对因变量的影响不大，但它仍然是影响$y$的一部分因素，如果被删除，会产生失真的弊端[38]．因此提出将特征值较大的部分运用最小二乘估计方法来估计其参数，而特征值较小（接近0）的部分运用岭回归的方法来估计参数．下面给出主成分-岭回归的定义及步骤.

3.3.2 主成分-岭回归概述

令$\Phi=(\Phi_1,\Phi_2)$，其中Φ_1为$p\times r$矩阵，对应的特征值为$\lambda_1,\lambda_2,\cdots,\lambda_r$，$\Phi_2$为$p\times(p-r)$矩阵，对应的特征值为$\lambda_{r+1},\lambda_{r+2},\cdots,\lambda_p$，于是模型变为[39]：

$$\begin{aligned} y&=X(\Phi_1,\Phi_2)\left(\Phi_1{}^T,\Phi_2{}^T\right)^T\beta+\varepsilon \\ &=Z_1\alpha_1+Z_2\alpha_2+\varepsilon. \end{aligned} \tag{3.10}$$

其中$Z_1=X\Phi_1, Z_2=X\Phi_2$，$\alpha_1=\Phi_1{}^T\beta,\alpha_2=\Phi_2{}^T\beta$．记$c=\frac{\sum_{i=1}^{r}\lambda_i}{\sum_{i=1}^{p}\lambda_i}$，称为前$r$个主成分的贡献率．将模型（3.3）变为：

$$\begin{cases} Y_1=Z_1\alpha_1+\dfrac{\varepsilon}{2}, \\ Y_2=Z_2\alpha_2+\dfrac{\varepsilon}{2}, \end{cases} \tag{3.11}$$

其中，$Y_1=cy$，$Y_2=(1-c)y$，$E(\varepsilon)=0, Cov(\varepsilon)=\sigma^2I_n$，$c$的取值可根据实际需要预先取定.

对模型（3.5）进行参数估计，其中α_1采用最小二乘估计方法，得到模型的参数估计为:

$$\alpha_1 = c((Z_1^T Z_1)^{-1} Z_1^T y)^T = c(\Lambda^{-1} Z_1^T y)^T, \beta_1 = c\Phi(\Lambda^{-1} Z_1^T y)^T = \Phi\alpha_1^T ; \quad (3.12)$$

α_2采用岭估计方法，得到模型的参数估计为：

$$\alpha_2 = (1-c)((Z_2^T Z_2 + kI_{p-r})^{-1} Z_2^T y^T)^T = (1-c)((\Lambda_2 + kI_{p-r})^{-1} Z_2^T y)^T, \quad (3.13)$$

$$\beta_2 = (1-c)\Phi((\Lambda_2 + kI_{p-r})^{-1} Z_2^T y)^T = \Phi\alpha_2^T, \quad (3.14)$$

其中，$\tilde{\Lambda} = \begin{pmatrix} c\Lambda_1^{-1} & 0 \\ 0 & (1-c)(\Lambda_2 + kI_{p-r})^{-1} \end{pmatrix}$.

α_1, α_2的估计$\hat{\alpha}_1, \hat{\alpha}_2(k)$具有下列性质：

（1）$c^T \hat{\alpha}_1$是$c^T\alpha_1$的最佳线性无偏估计；

（2）$\hat{\alpha}_2(k)$是α_2的一个有偏估计；

（3）$Cov(\hat{\alpha}_1) = \frac{\sigma^2\Lambda_1^{-1}}{2}; Cov(\hat{\alpha}_2(k)) = \frac{\sigma^2(\Lambda_2 + kI_{p-r})^{-1}\Lambda_2(\Lambda_2 + kI_{p-r})^{-1}}{2}$.

引理 3.1 岭参数$\hat{\beta}$比$\hat{\beta}(k)$具有更小的偏差.

3.3.3 主成分-岭回归估计修正多重共线性

具体步骤如下：

（1）求特征值.根据模型（3.11），求出设计矩阵$X^T X$的顺序特征值$\lambda_1 \geq \cdots \geq \lambda_{r+1} \geq \cdots \geq \lambda_p$以及对应的特征向量.

（2）模型分解. 根据模型（3.11）和第一步求得的特征值，选取合适的c值，将模型分成两部分，以便进一步参数求解.

（3）根据模型（3.11）对模型中的特征值较大部分采用最小二乘估计方法进行求解.求解得到$\hat{\alpha}_1 = (Z_1^T Z_1)^{-1} Z_1^T Y_1$， $Y_1 = cy$

（4）根据模型（3.11）对模型中特征值较小部分采用岭估计方法进行模型参数求解.求解得到$\hat{\alpha}_2 = (Z_2^T Z_2 + kI_{p-r})^{-1} Z_2^T Y_2$，其中$Y_2 = (1-c)y$.

（5）得到模型（3.8）的参数估计后还原得到原始模型（3.7）的参数估计. 模型（3.8）的参数估计为 $\hat{\alpha}=(\hat{\alpha_1},\hat{\alpha_2})^T$，还原得到原始模型（3.7）的参数估计 $\hat{\beta}=\Phi\hat{\alpha}$，其中 $X^TX=\Phi\Lambda\Phi^T$，$\Lambda=diag(\lambda_1,\lambda_2,\cdots,\lambda_p)$，$\lambda_1\geq\lambda_2\geq\cdots\geq\lambda_p$ 为 X^TX 的特征值.

总结修正多重共线性的回归模型的相关理论与步骤，分析不同的回归模型修正多重共线性的优缺点. 改进的逐步回归通过删除变量来消除多重共线性的影响，比较方便常用，但筛选后可分析的变量较少，不能全面概括事物发展的因素；岭回归通过引入单位矩阵，减少设计矩阵接近奇异的程度，即增大特征值，减小均方误差；考虑主成分舍弃了贡献率较小的部分，我们进一步提出主成分-岭回归修正多重共线性，以便分析事物更加客观. 为了验证模型的可靠性和适用性，提出实证分析，检验各个回归模型下的多重共线性修正效果，以便对河南省经济发展影响因素进行分析和预测.

第四章 基于多重共线性修正的河南省生产总值分析及预测

河南省是人口大省，农业大省，文化大省，新兴工业大省，2001～2020 年河南经济从高速发展阶段过渡到高质量发展阶段，分析河南省经济影响因素，可能为未来的发展提供方向和具体举措．为了检验前文提出的回归模型修正效果，选取相关经济指标进行实证分析，诊断和修正建模过程中存在的多重共线性问题.

4.1 河南省生产总值建模过程的共线性问题诊断和修正

4.1.1 指标选择

本文选取河南省 2001-2020 年的 11 个重要经济增长发展指标，X_1：固定资产投资总额 K(亿元)，X_2:年末就业人口数 L(万人)，X_3:人口增长率 Gr(‰)，X_4:各级教育平均在校人数 S(万人)，X_5:历年科学技术支出 $R\&D$(亿元)，X_6:专利申请量 P，X_7:城镇化率 U(%)，X_8:二三产业就业人数占比 St，X_9:出口总额(万元) Tr，X_{10}:历年小麦产量(万吨) G，X_{11}:历年能源消耗总数(万吨) M．数据来源于河南省 2021 年统计年鉴[25]，具体如表所列：

表 4-1 河南省 2001～2020 年指标原始数据

指标 年份	生产总值 y	固定资产投资总额 x1	年末就业人口数 x2	人口增长率 x3	教育平均在校人数 x4	科学技术支出 x5
2001	5533.01	1627.99	5516.59	6.94	1070.39	7.25
2002	6035.48	1820.45	5522	6.03	1135.32	7.95
2003	6942.41	2310.54	5535.67	5.64	1132.54	9.06
2004	8411.19	3099.38	5587	5.2	1133.95	10.4
2005	10243.47	4378.69	5662	5.25	1139.09	13.85
2006	11977.87	5907.74	5719	5.32	1158.49	17.37
2007	14824.49	8010.11	5773	4.9	1176.98	25.23
2008	17735.93	10490.65	5835	4.97	1190.49	30.44
2009	19181	13704.65	5948.78	4.99	1206.97	35.52
2010	22655.02	14124.69	5156	4.95	1230.96	44.67
2011	26318.68	17770.51	5129	4.94	1240.29	56.59
2012	28961.92	21449.99	5110	5.16	1223.96	69.64
2013	31632.5	26087.45	5094	5.51	1090.27	80
2014	34574.76	30782.17	5082	5.78	1092.38	81.25

2015	37084.1	35660.34	5075	5.65	1102.32	83.25
2016	40249.34	40415.09	5052	6.15	1127.4	96.1
2017	44824.92	44496.93	5029	5.98	1171.01	137.94
2018	49935.9	48101.18	4992	4.92	1217.87	155.67
2019	53717.75	51949.28	4934	4.18	1259.58	211.07
2020	54997.07	54183.09	4884	2.09	1292.59	254.28

指标 年份	专利 申请量 x6	城镇化率 x7	二三产业 人数占比 x8	出口 总额 x9	小麦 产量 x10	能源 消耗总数 x11
2001	4092	24.43	36.96	1419864.3	2299.71	8367
2002	4441	25.8	38.46	1754333	2248.39	9005
2003	5261	27.2	39.81	2467779.5	2292.5	10595
2004	6318	28.9	41.92	3457810.8	2480.93	13074
2005	8981	30.65	44.56	4131243.2	2577.69	14625
2006	11538	32.5	46.67	5289240.3	2936.5	16234
2007	14916	34.34	49.42	6424771	2958.31	17837.79
2008	19090	36.03	51.21	7504743.5	3036.2	18976.36
2009	19589	37.7	53.52	5018380.5	3092.2	19751.24
2010	25149	38.82	55.11	7131309.5	3121	18964
2011	34076	40.47	56.91	12208344	3144.9	20462.41
2012	43442	41.99	58.2	18697083	3223.07	20919.96
2013	55920	43.6	59.88	22312067	3266.33	21909.1
2014	62434	45.05	62.68	24188066	3385.2	22889.93
2015	74373	47.02	66.13	26840255	3526.9	22343
2016	94669	48.78	69.4	28353441	3618.62	22323
2017	119243	50.56	72.66	31718144	3705.21	22162
2018	154381	52.24	73.9	35789861	3602.85	22659
2019	144010	54.01	74.64	37546362	3741.77	22300
2020	186369	55.43	74.96	40750000	3753.13	22752

注:该数据来源于河南统计年鉴

通过对原始数据分析，经济指标大部分呈上升趋势，其中就业人口和人口增长率受非抗力因素，数据呈波动下降的趋势．其他变量数据均呈明显的上升趋势．这种情况是符合多重共线性原因分析中提到的一种，与经济变量有共同变化趋势的自变量间

容易存在共线性问题．所以，采用删除对因变量影响作用较小，但多重共线性问题较严重的变量.

4.1.2 改进的逐步回归修正多重共线性

当模型自变量之间存在多重共线性问题，回归方程就会变得不稳定．对原始数据方程两边取对数，消除可能存在的异方差特征，但不影响变量之间的关系.

通过相关系数矩阵表 4-2，可以看到地区生产总值 y 与自变量 x_1，x_2，x_5，x_6，x_7，x_8，x_9，之间的相关系数接近 1，由此分析固定资产总额等自变量与地区生产总值之间存在某种很强的相关关系，删除与地区生产总值相关系数较小的变量 x_3, x_4，使模型符合实际要求.

表 4-2 相关系数矩阵

		y	z1	z2	z3	z4	z5	z6	z7	z8	z9	z10	z11
相关性	y	1.000	0.995	-0.742	-0.363	0.382	0.963	0.968	0.998	0.993	0.980	0.982	0.938
	z1	0.995	1.000	-0.789	-0.308	0.336	0.959	0.968	0.996	0.994	0.984	0.970	0.905
	z2	-0.742	-0.789	1.000	0.008	-0.114	-0.707	-0.717	-0.738	-0.741	-0.776	-0.652	-0.527
	z3	-0.363	-0.308	0.008	1.000	-0.803	-0.317	-0.369	-0.363	-0.331	-0.303	-0.356	-0.472
	z4	0.382	0.336	-0.114	-0.803	1.000	0.432	0.365	0.376	0.366	0.276	0.375	0.405
	z5	0.963	0.959	-0.707	-0.317	0.432	1.000	0.918	0.965	0.969	0.917	0.953	0.884
	z6	0.968	0.968	-0.717	-0.369	0.365	0.918	1.000	0.973	0.971	0.966	0.941	0.891
	z7	0.998	0.996	-0.738	-0.363	0.376	0.965	0.973	1.000	0.998	0.981	0.982	0.930
	z8	0.993	0.994	-0.741	-0.331	0.366	0.969	0.971	0.998	1.000	0.976	0.980	0.913
	z9	0.980	0.984	-0.776	-0.303	0.276	0.917	0.966	0.981	0.976	1.000	0.957	0.912
	z10	0.982	0.970	-0.652	-0.356	0.375	0.953	0.941	0.982	0.980	0.957	1.000	0.951
	z11	0.938	0.905	-0.527	-0.472	0.405	0.884	0.891	0.930	0.913	0.912	0.951	1.000

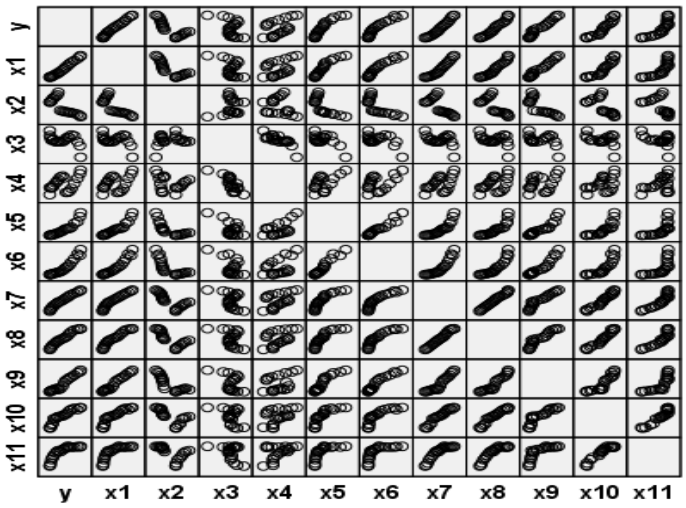

图 4-1 相关系数矩阵散点图

我们通过相关系数矩阵散点图，更加直观清晰看出自变量与因变量的线性相关关系．但也可以看出存在大部分自变量之间相关程度较高的现象，反映出模型各自变量间可能存在多重共线性.

根据自变量建立多元线性回归模型，进行相关的显著性检验，F 检验均显著，但部分自变量 t 检验不显著，且出现回归系数与现实意义相违背的情况．根据t检验选择依次删除不显著的变量$x_3, x_4, x_7, x_8, x_5, x_2, x_{10}, x_6, \lambda$来修正多重共线性，删除后的变量为$x_1, x_{11}$．对删除后得到的变量进行 t 检验并做共线性检验，给出方差因子表，如表 4-3 所示：

表 4-3 方差因子表

模型		非标准化系数		标准系数	t	Sig.	共线性统计量	
		B	标准 误差	试用版			容差	VIF
1	（常量）	-1.697	0.329		-5.153	0.000		
	z1	0.766	0.020	0.818	39.207	0.000	0.187	5.358
	z11	0.467	0.049	0.197	9.467	0.000	0.187	5.358

a. 因变量：y

可以看到，方差扩大因子法分析自变量，VIF 不超过 10 ，但是在利用条件数对模型的多重共线性问题做出诊断，给出共线性诊断表 4-4：

表 4-4 共线性诊断表

模型 维数	特征值	条件索引	方差比例		
			（常量）	z1	z11

	1	2.996	1.000	0.00	0.00	0.00
1	2	0.004	28.323	0.04	0.21	0.00
	3	0.000	157.144	0.96	0.79	1.00

a. 因变量：y

在多重共线性诊断表中，自变量x_1, x_{11}的条件数分别为 28.322，157.144，依据条件数$0<k<10$判断：回归模型的自变量之间仍然存在较严重的共线性问题.

通过逐步回归法删除回归方程中的变量，能够消除多重共线性影响，但也会删除掉我们对因变量影响作用较大的自变量；而且即使方程中只保留了少数自变量，共线性问题也仍然很严重.

4.1.3 岭回归模型修正多重共线性问题

为了提高回归模型解释的科学性，对河南省 2001 ～2020 年指标数据进行标准化. 当自变量存在严重的多重共线性问题，使得设计矩阵退化，从而导致LS估计稳定性变差，这里我们提出采用岭估计进行回归分析.

（1）设计矩阵标准化

求标准化的设计矩阵X^*，消除量纲的影响. 计算得出标准化的设计矩阵列向量如表 5 所示：

表 4-5 设计矩阵列向量表

x0*	x1*	x2*	x3*	x4*	x5*
0.224	0.038	0.231	0.292	0.204	0.016
0.224	0.043	0.231	0.254	0.217	0.018
0.224	0.047	0.232	0.238	0.216	0.020
0.224	0.054	0.234	0.219	0.216	0.024
0.224	0.063	0.237	0.221	0.217	0.031
0.224	0.073	0.239	0.224	0.221	0.039
0.224	0.088	0.242	0.206	0.225	0.057
0.224	0.104	0.244	0.209	0.227	0.069
0.224	0.116	0.249	0.21	0.23	0.08
0.224	0.143	0.216	0.208	0.235	0.101
0.224	0.17	0.215	0.208	0.237	0.128
0.224	0.195	0.214	0.217	0.234	0.158
0.224	0.222	0.213	0.232	0.208	0.181

0.224	0.247	0.213	0.243	0.209	0.184
0.224	0.275	0.212	0.238	0.210	0.188
0.224	0.302	0.211	0.259	0.215	0.217
0.224	0.333	0.210	0.252	0.224	0.312
0.224	0.367	0.209	0.207	0.233	0.352
0.224	0.403	0.207	0.176	0.24	0.478
0.224	0.421	0.204	0.088	0.247	0.575
x6*	x7*	x8*	x9*	x10*	x11*
0.012	0.134	0.143	0.015	0.164	0.098
0.013	0.141	0.149	0.019	0.160	0.106
0.015	0.149	0.154	0.026	0.163	0.125
0.018	0.158	0.162	0.037	0.177	0.154
0.026	0.168	0.173	0.044	0.184	0.172
0.033	0.178	0.181	0.057	0.209	0.191
0.043	0.188	0.192	0.069	0.211	0.210
0.055	0.197	0.198	0.08	0.216	0.223
0.057	0.206	0.207	0.054	0.220	0.232
0.073	0.212	0.214	0.076	0.222	0.223
0.098	0.221	0.221	0.131	0.224	0.241
0.125	0.230	0.226	0.200	0.23	0.246
0.161	0.238	0.232	0.239	0.233	0.258
0.18	0.246	0.243	0.259	0.241	0.269
0.215	0.257	0.256	0.288	0.251	0.263
0.273	0.267	0.269	0.304	0.258	0.263
0.344	0.277	0.282	0.34	0.264	0.261
0.446	0.286	0.286	0.384	0.257	0.267
0.416	0.295	0.289	0.402	0.267	0.262
0.538	0.303	0.291	0.437	0.267	0.268

消除量纲对数据影响后，进行线性回归过程中自变量的相关性，可能会出现多重共线性，运用条件数来进一步诊断自变量间的共线性问题，

（2）条件数诊断多重共线性个数

采用条件数判断存在多重共线性，需要求得对称方阵 $X^{*T}X^{*}$ 的特征根与特征向量. 根据软件计算 $X^{*T}X^{*}$ 的特征根和特征向量，运用公式计算条件数如表 4-6 所示. 但大部分特征根接近于零，说明该模型存在较严重的共线性问题. 这时需要用条件数 k_i 进一步确定存在的多重共线性个数，运用条件数公式 $k_i=\sqrt{\frac{\lambda_m}{\lambda_i}}$，认为当 $k\geq 10$ 时，量纲处理后的指标矩阵 X 没有多重共线性.

表 4-6 对称方阵的特征根和特征向量

特征根	特征向量					
	x0*	x1*	x2*	x3*	x4*	x5*
10.81212	-0.23516	-0.10932	0.76575	0.136908	-0.02966	-0.24521
1.09757	-0.12736	-0.41074	-0.01346	0.213051	0.485148	-0.28052
0.04638	0.035286	-0.07692	-0.3141	0.249246	0.558773	0.340077
0.0315	0.041839	-0.01171	-0.08662	0.086752	-0.28517	0.28128
0.00696	0.017646	-0.17559	-0.48359	-0.08576	-0.32352	-0.61326
0.0033	-0.01257	0.082866	-0.00247	0.038596	-0.15136	0.295264
0.00123	0.019204	-0.04781	0.05499	0.011919	-0.16251	0.079254
0.00044	0.836809	0.299202	0.194846	0.079905	0.076682	-0.12681
0.0003	-0.45272	0.765956	-0.07653	-0.06292	0.149437	-0.10298
0.00015	0.007955	0.079917	-0.12502	-0.12986	-0.04331	-0.1354
0.00005	-0.00098	-0.23761	0.100799	-0.84769	0.114799	0.234316
0.00001	-0.13921	-0.18507	-0.02807	0.333833	-0.41587	0.311348
条件数	x6*	x7*	x8*	x9*	x10*	x11*
1	-0.24234	-0.06835	-0.0903	-0.20843	0.25708	0.292461
3.138	0.440657	-0.12046	0.00889	0.280053	-0.29119	0.288887
4.86	-0.36585	0.112463	-0.04234	-0.28573	0.302462	0.287696
18.52	0.298487	-0.2138	-0.65982	0.234851	0.344133	0.280878
39.42	-0.003	0.001536	0.06667	-0.31841	0.233652	0.294021
57.27	0.121404	-0.55402	0.123655	-0.52557	-0.44403	0.266518
93.93	0.042457	0.721577	-0.32267	-0.21272	-0.46943	0.263218
156.93	0.109839	0.029972	0.171119	0.083215	0.051498	0.303474
189.62	0.20203	0.10588	0.101674	0.081861	0.067846	0.303213

268.60	-0.66803	-0.23267	-0.18424	0.439754	-0.36809	0.278626
471.64	0.083827	0.048991	0.165045	0.046628	0.131367	0.301027
1115.36	-0.03778	0.161864	0.57442	0.330582	0.084131	0.300502

通过表 4-6 分析，只有 3.13862，4.86450 是小于 10 的，于是我们可以得到只有两个变量间不存在多重共线性关系，且最大特征值与最小特征值的比值为 1115，根据特征根与条件数判断该模型具有严重的多重共线性问题.

（3）绘制岭迹图，选择 K 值

绘制岭迹图，进行岭迹分析. 把 11 个回归系数的岭迹绘成图，代码和岭迹图如下：

```
%% 岭回归(Ridge Regression)
%先画出岭迹图，以便选取合适的岭参数
%k=0:1e-3:10;%岭参数
k=0:5e-2:1;%岭参数
%k=0.1:0.02:0.3;%岭参数
b=ridge(Y,X,k);%回归系数

%岭迹图，一般选取开始平稳的“拐点”处的k值
figure(1)
plot(k,b)
xlabel('k')
ylabel('β')
title('岭迹图')
%legend('2001','2002','2003','2004','2005','2006','2007','2008','2009','2
legend('x1','x2','x3','x4','x5','x6','x7','x8','x9','x10','x11')
```

图 4-2 岭迹图绘制代码

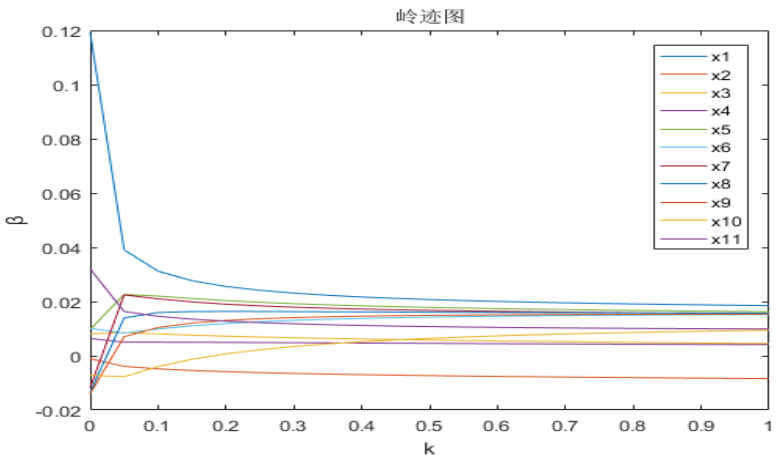

图 4-3 0-1 岭迹图

根据岭迹图分析，当 k 与 0 略有偏离时，$\hat{\beta}(k)$ 与 $\hat{\beta}=\hat{\beta}(0)$ 就有较大的差距，这说明了 LS 估计的稳定性较差. 运用软件设置 K 的范围是[0, 1]，每次增加步长为 0.05. 把 11 个 K 值算出 $\hat{\beta}(k)$，得到岭迹图.

具体对标准化岭回归系数进行分析，原来普通 LS 估计为负数的 x_7, x_8, x_9，，其标准化回归系数 $\hat{\beta}_7(k), \hat{\beta}_8(k), \hat{\beta}_9(k)$ 迅速从负值变为正值，而原先回归系数为较大正值的 $\hat{\beta}_1(k)$ 迅速减少，岭迹图在 $k=0$ 到 $k=0.3$ 之间达到稳定． 把岭参数取值范围改为 $[0, 0.3]$，步长改为 0.02，重新做岭回归，岭迹分析图如图所示：

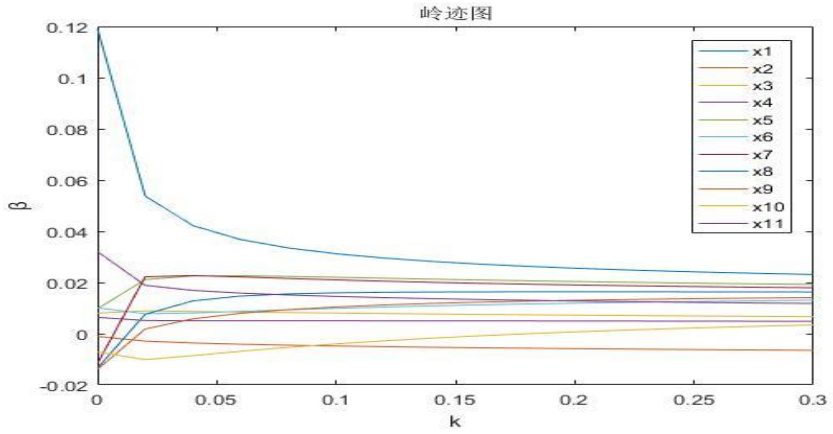

图 4-4 0-0.3 岭迹分析图

通过岭迹分析图看到，当 $k=0.10$ 左右，岭迹大体上达到稳定.

为了确定岭参数 k，需要具体分析不同 k 值下的岭回归系数情况，如表 4-7 所示：

表 4-7 岭回归系数估计值

K	RSQ	x1	x2	x3	x4	x5
0.00	0.99973	0.99569	-0.00894	0.06737	0.05328	0.08110
0.02	0.99878	0.18331	-0.05736	0.05250	0.03960	0.15520
0.04	0.99850	0.16125	-0.06659	0.04221	0.03613	0.14087
0.06	0.99830	0.15209	-0.07201	0.03541	0.03356	0.13406
0.08	0.99813	0.14674	-0.07585	0.03021	0.03162	0.12985
0.10	0.99796	0.14309	-0.07879	0.02597	0.03011	0.12686
0.12	0.99779	0.14034	-0.08113	0.02239	0.02893	0.12458
0.14	0.99762	0.13815	-0.08304	0.01930	0.02798	0.12273
0.16	0.99744	0.13633	-0.08461	0.01658	0.02722	0.12118
0.18	0.99726	0.13478	-0.08593	0.01417	0.02661	0.11984
0.20	0.99707	0.13342	-0.08705	0.01200	0.02613	0.11867

K	x6	x7	x8	x9	x10	x11
0.00	0.08436	-0.09767	-0.11241	-0.11659	-0.05990	0.26784
0.02	0.11427	0.14506	0.13551	0.12149	0.04093	0.09445

0.04	0.12490	0.13366	0.13273	0.12767	0.07027	0.08498
0.06	0.12800	0.12884	0.13086	0.12889	0.08241	0.08195
0.08	0.12884	0.12608	0.12940	0.12901	0.08893	0.08091
0.10	0.12879	0.12422	0.12818	0.12876	0.09292	0.08066
0.12	0.12834	0.12284	0.12712	0.12835	0.09557	0.08076
0.14	0.12770	0.12176	0.12617	0.12787	0.09742	0.08102
0.16	0.12697	0.12086	0.12532	0.12736	0.09876	0.08135
0.18	0.12620	0.12009	0.12453	0.12684	0.09976	0.08170
0.20	0.12543	0.11941	0.12381	0.12632	0.10051	0.08205

然后通过参照复决定系数以及残差平方和综合分析，当$k=0.17$时，调整后的$R^2=0.998$，符合模型总体的显著性要求.当$k=0.17$时，残差平方和趋于稳定的最小值，因此运用软件得到$k=0.17$时的岭回归系数表 4-8.

表 4-8 岭估计系数检验表

k = 0.17 ******				
Mult R	.99739			
Square	.99739			
Adj R Squ	.99382			
SE	1295.3366			
	ANOVA table			
	Df	SS	MS	F value
Regress	11.000	5.15E+009	467919091	278.87236
Residual	8.000	13423176	1677897	
	B	SE(B)	Beta	B/SE(B)
x1	0.123	0.006707	0.137	18.373231
x2	-4.385	0.942877	-0.091	-4.650242
x3	2056.681	3047.53	0.012	0.674867
x4	7.912	4.761259	0.030	1.661709
x5	28.173	2.846575	0.1198	9.897240
x6	0.036	0.003662	0.123	9.813992
x7	202.488	8.063093	0.1193	25.112955
x8	159.088	8.718341	0.1227	18.247520
x9	1.535	0.118174	0.126	12.986039

x10	3.223	0.390008	0.098	8.264423
x11	0.284	0.050103	0.083	5.661517
Constant	-2027.108	8513.326	0.000	-.238110

4.1.4 主成分-岭回归修正多重共线性问题

同样是将自变量数据标准化，然后用 MATLAB 计算得到设计矩阵 $X^{*T}X^{*}$ 的特征根和特征向量如表 4-6 所示. 运用主成分回归，根据贡献率公式算得 $c=90\%$，根据前五个变量的占有率达到 90%，将前两个主成分作为第一个模型，后面九个主成分作为第二个模型，其对应的主成分特征向量如表 4-9 所示：

表 4-9 主成分-岭回归的主成分特征向量

	第一个模型		第二个模型		
	1	2	3	4	5
x1	0.297509	-0.64434	-0.36828	0.414497	-0.03477
x2	-0.02972	0.026415	-0.05438	0.016059	0.12431
x3	-0.02978	-0.0094	0.161074	-0.02473	0.188131
x4	0.011368	-0.05438	-0.03955	-0.08508	0.123733
x5	0.056522	0.154287	0.366565	0.189268	0.418304
x6	-0.06747	-0.09489	0.531168	-0.21947	-0.37241
x7	-0.85796	0.052701	-0.13179	0.312613	0.015437
x8	0.363352	0.700291	-0.16608	0.325469	-0.22964
x9	0.004344	0.124539	-0.43243	-0.57326	0.484739
x10	0.000629	-0.09571	-0.08454	-0.45075	-0.45634
x11	0.183595	-0.1744	0.422257	0.025093	0.35039

第二个模型					
6	7	8	9	10	11
-0.03014	-0.16934	0.141118	-0.0984	0.128429	0.334642
-0.02775	-0.39977	0.713439	0.407242	-0.26238	-0.27473
0.042484	-0.62915	-0.25958	0.094663	0.652876	-0.19198
-0.07497	-0.49537	-0.50387	0.056909	-0.65646	0.183632
0.600864	-0.07594	0.206641	-0.32357	-0.0864	0.325829
-0.43038	-0.23442	0.271643	-0.31162	0.014611	0.325823
-0.05544	-0.01233	0.003321	0.152026	0.071938	0.338044
-0.15911	-0.13906	0.003666	0.153805	0.100445	0.336017

-0.22194	0.009009	0.16173	-0.15729	0.154885	0.331135
0.561194	-0.06379	0.013657	0.375597	0.082305	0.32439
-0.22814	0.293135	-0.08348	0.630784	0.061364	0.29593

然后，经由特征向量矩阵转换之后，多元线性回归模型转化为

$$y_1 = \alpha_1 x_1 + \alpha_2 x_2, \tag{4.1}$$

$$y_2 = \alpha_3 x_3 + \alpha_4 x_4 + \alpha_5 x_5 + \alpha_6 x_6 + \alpha_7 x_7 + \alpha_8 x_8 + \alpha_9 x_9 + \alpha_{10} x_{10} + \alpha_{11} x_{11}, \tag{4.2}$$

对(4.1)式进行最小二乘回归，得到的参数估计为：

$$\alpha_1 = (Z_1^T Z_1)^{-1} Z_1^T Y_1,$$

其中 $Z_1 = X\Phi_1$， $Y_1 = cy = 0.9c$. 代入 $\alpha_1 = (Z_1^T Z_1)^{-1} Z_1^T Y_1$，求得式（4.1）的参数估计为

$$\hat{\alpha_1} = (0.3054, 0.0755).$$

对式（4.2）进行岭估计，参数估计为 $\alpha_2 = (Z_2^T Z_2 + kI_{n-r})^{-1} Z_2^T Y_2$，其中

$$Z_2 = X\Phi_2, Y_2 = (1-c)y = 0.1y.$$

采用岭迹法确定 $\alpha_2 = (Z_2^T Z_2 + kI_{n-r})^{-1} Z_2^T Y_2$ 中的 k 值，对数据进行主成分-岭回归，MATLAB 程序如下：

```
%%  迭代计算
k_Range = 0 : 0.01 :0.2;
n = 1;
alpha1 = pinv(Z1' * Z1) * Z1' * Y1;
for k = k_Range
    alpha2 = pinv(Z2' * Z2 + k * ones(size(Z2, 2))) * Z2' * Y2;
    R = A * [alpha1; alpha2];
    Table_alpha2(:, n) = alpha2;
    Result(n).k = k;
    Result(n).alpha1 = alpha1;
    Result(n).alpha2 = alpha2;
    Result(n).R = R;
    n = n + 1;
end
plot(Table_alpha2');xlim([1 size(X, 1)])
```

图 4-5 主成分-岭回归中岭回归代码

根据计算，得到第二主成分中岭回归系数（即 $\alpha_3, \alpha_4, \alpha_5, \alpha_6, \alpha_7, \alpha_8, \alpha_9, \alpha_{10}, \alpha_{11}$）随 k 值变化的相关数据如表所示：

表 4-10 主成分-岭回归岭迹回归系数表

k	α3	α4	α5	α6
0.00	-0.00243	0.001706	-0.00918	-0.00286
0.01	-0.00242	0.001711	-0.00917	-0.00278
0.02	-0.00242	0.001711	-0.00917	-0.00277

0.03	-0.00242	0.001711	-0.00917	-0.00277
0.04	-0.00242	0.001711	-0.00917	-0.00277
0.05	-0.00242	0.001711	-0.00917	-0.00277
0.06	-0.00242	0.001712	-0.00917	-0.00277
0.07	-0.00242	0.001712	-0.00917	-0.00277
0.08	-0.00242	0.001712	-0.00917	-0.00277
0.09	-0.00242	0.001712	-0.00917	-0.00277
0.10	-0.00242	0.001712	-0.00917	-0.00277
α7	α8	α9	α10	α11
0.022	0.028899	0.003015	-0.03616	-0.01825
0.022	0.029095	0.003284	-0.03403	-0.01016
0.022	0.029115	0.003311	-0.03381	-0.00934
0.022	0.029123	0.003322	-0.03373	-0.00902
0.022	0.029127	0.003327	-0.03369	-0.00886
0.022	0.029129	0.003331	-0.03366	-0.00876
0.022	0.029131	0.003333	-0.03364	-0.00869
0.022	0.029132	0.003335	-0.03363	-0.00864
0.022	0.029133	0.003336	-0.03362	-0.0086
0.022	0.029134	0.003337	-0.03361	-0.00857
0.022	0.029134	0.003338	-0.0336	-0.00855

通过软件绘制回归系数α_2随k值变化的岭迹图，如图 4-6 所示：

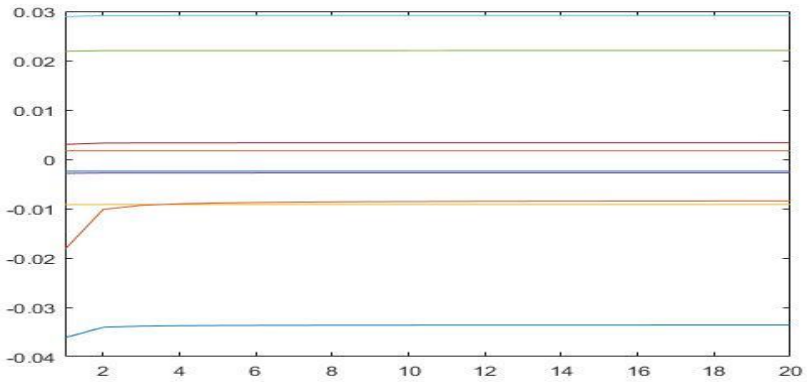

图 4-6 主成分-岭回归 0-0.2 岭迹图

通过岭迹图可以看出，岭参数在$k=0.07$(横轴为第八个点)趋于稳定，所以选择$k=0.07$. 再将k值代入参数估计式$\alpha_2=(Z_2{}^TZ_2+kI_{n-r})^{-1}Z_2{}^TY_2$，求得（4.2）的估计参数：

$\hat{\alpha}_2 = (-0.0024, 0.0017, -0.009, -0.0027, 0.022, 0.029, 0.003, -0.034, -0.008)$.

最后，将两个模型的参数估计结果合并，并且将特征向量矩阵进行还原得到主成分—岭回归模型的参数估计：

$\hat{\beta} = (0.143, -0.097, 0.0002, 0.012, 0.103, 0.096, 0.123, 0.089, 0.102, 0.083, 0.105)$.

4.2 河南省生产总值影响因素分析及预测

4.2.1 改进的逐步回归预测河南省生产总值

根据上节的逐步回归修正结果，将因变量 y 与删除后的自变量 x_1, x_{11} 之间建立回归方程为：

$$y = 65.083 + 0.791x_1 + 0.487x_{11}.$$

选取 2001～2020 年的固定资产投资总额以及能源消费总额数据，根据时间序列预测 2021～2023 年的固定资产总额和能源消费总额，代入回归方程，预测 2021～2023 年的生产总值.得到结果如表所示：

表 4-11 逐步回归预测值

年份	真实值	时间序列预测值 x1	时间序列预测值 x11	逐步回归	精确度
2021	58887.41	56530.31	22929.57	55947.26	5.26%
2022	61345.05	58990.94	23136.02	57994.16	5.78%
2023		61564.97	23342.48	60130.76	

小结：从预测结果来看，通过建立改进的逐步回归方程，预测得到未来三年河南省生产总值分别为 55947.2 亿元，57994.1 亿元，60130.8 亿元，2021 年，2022 年平均预测精度为 5%左右，预测效果一般.

逐步回归通过筛选变量修正了自变量间的共线性问题，但保留较少的变量预测效果较差，有理由怀疑该模型不能有效分析河南省经济发展影响因素，选择主成分—岭回归模型和岭回归模型尽可能保留相关经济变量，并做进一步的分析.

4.2.2 基于主成分—岭回归和岭回归的河南省生产总值影响因素分析

（1）基于主成分-岭回归的河南省经济发展影响因素分析

根据前面主成分—岭回归修正多重共线性得到的回归系数，对河南省经济发展影响指标进行排名，可以分析得到近阶段河南省的生产总值的主要影响因素，以及未来的发展方向的调整，列表如下：

表 4-12 主成分—岭回归系数排序

经济发展指标	主成分—岭回归系数	排名
x1：固定资产投资总额	0.14320181	1
x7：城镇化率	0.12323017	2
x11：能源消费总量	0.10542699	3
x5：科学技术支出	0.10344341	4
x9：出口总额	0.10233106	5
x6：专利申请量	0.09612963	6
x8：二三产业占比	0.08874394	7
x10：小麦产量	0.0831859	8
x4：教育在校人数	0.01212889	9
x3：人口增长率	0.00015665	10
x2：年末就业人口数	-0.0973497	11

数据经过标准化处理，得到 y 对 $x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8, x_9, x_{10}, x_{11}$ 的标准化主成分—岭回归方程为：

$$\hat{y} = 0.143x_1 - 0.097x_2 + 0.0002x_3 + 0.012x_4 + 0.103x_5 + 0.096x_6$$
$$+0.123x_7 + 0.089x_8 + 0.102x_9 + 0.083x_{10} + 0.105x_{11} .$$

从方程可知，就业人口与地区生产总值呈现负相关关系，其他自变量与河南省生产总值呈现正相关关系．当上述变量发生变动时会引起河南地区生产总值同方向的变动．根据回归系数对河南省经济发展影响因素排名，其中与河南省生产总值呈正相关关系的主要影响因素为：固定资产投资总额，城镇化率，能源消费总量．从回归结果可以分析：在其他变量一定的情况下，固定资产投资总额每增加 0.143 亿元，河南省生产总值增加 1 亿元；城镇化率每提高 0.123%，河南省生产总值增加一亿元；能源消费总量每增加 0.105 万吨，河南省生产总值增加一亿元.

（2）基于岭回归的河南省经济发展影响因素分析

对河南省 2001 ～2020 年指标数据进行标准化．为了提高回归模型解释的科学性，针对自变量存在多重共线性造成设计矩阵退化，从而使得最小二乘估计效果明显变差问题，这里我们提出采用岭估计进行回归分析.

借助标准化岭回归参数，对数据指标进行排序如表 4-13：

表 4-13 岭回归系数排序

经济发展指标	岭回归系数	排名
x1：固定资产投资总额	0.137	1
x9：出口总额	0.126	2
X8：二三产业占比	0.1198	3
X6：专利申请量	0.1233	4
X5：科学技术支出	0.1198	5
x7：城镇化率	0.1193	6
x10：小麦产量	0.098	7
x11：能源消耗总数	0.083	8
x4：教育在校人数	0.030	9
x3：人口增长率	0.012	10
x2：年末就业人口数	-0.091	11

y 对 $x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8, x_9, x_{10}, x_{11}$ 的标准化岭回归方程为：

$$\hat{y} = 0.137x_1 - 0.091x_2 + 0.012x_3 + 0.030x_4 + 0.120x_5 + 0.123x_6$$
$$+0.124x_7 + 0.123x_8 + 0.126x_9 + 0.098x_{10} + 0.083x_{11}.$$

经过标准化处理后，模型中的截距项无空值，因此从方程可知，就业人口与地区生产总值呈现负相关关系，其他自变量与河南省生产总值呈现正相关关系．当上述变量发生变动时会引起河南地区生产总值同方向的变动，其中与河南省生产总值呈正相关关系的主要影响因素为：固定资产投资总额，出口总额，二三产业占比．从回归结果可以分析：在其他变量一定的情况下，固定资产投资总额每增加 0.137 亿元，河南省生产总值增加 1 亿元；出口总额每增加 0.126 亿元，河南省生产总值增加一亿元；二三产业占比每增加 0.1198%，河南省生产总值增加一亿元.

4.2.3 基于岭回归模型的河南省预测

通过变量代换，得到原始变量，即未标准化岭回归方程为：

$$\hat{y} = -2027.108 + 0.123x_1 - 4.385x_2 + 2056.681x_3 + 7.912x_4 + 28.173x_5$$
$$+202.488x_7 + 159.088x_8 + 1.535x_9 + 3.223x_{10} + 0.284x_{11}.$$

我们通过对 2001～2020 年河南经济指标数据分析，大部分自变量整体呈上升趋势，对数据做差分处理，运用时间序列预测自变量的值，具体预测值如表 4-14 所示：

表 4-14 未来三年河南经济发展指标数据

	x1	x2	x3	x4	x5	x6
2021	56530.31	4887.28	0.35526258	1304.28	297.94	205591
2022	58990.94	4887.28	-0.720660942	1315.98	341.73	229484
2023	61564.97	4887.28	-1.28331915	1327.67	385.51	253377
	x6	x7	x8	x9	x10	x11
2021	205591	57.06	75．28	43766967.44	3829.63	22929.57
2022	229484	58.69	75．60	46790758.82	3906.12	23136.02
2023	253377	60.32	75．92	49814550.21	3982.62	23342.48

要说明的是：由于疫情影响较大，河南省就业人口数预测结果与实际可能有出入，尽量选择与 20 年齐平数据.

我们通过代入原始变量的岭回归方程得出河南省未来三年的生产总值如表所示：

表 4-15 河南省 2021～2023 年生产总值预测

	真实值	时间序列	精确度	岭回归	精确度
2021	58887.41	56276.53	4.64%	59443.90	-0.94%
2022	61345.05	57555.98	6.58%	60870.41	0.78%
2023		58835.43		63366.17	

小结：岭回归预测得到河南省生产总值 2021 年为 59443.90 亿元，2022 年为 60870.41 亿元，通过与实际值对比计算得到，预测值的精确度分别达到-0.94%和 0.78%，预测效果较好，实际意义下，预期河南省生产总值在 2023 年能达到 6.3 万亿元. 与上节提出的改进的逐步回归，以及时间序列相比，岭回归能够考虑多个变量影响，做出的回归值更准确，可以作为预测模型的一种优选.

第五章 结论与展望

5.1 结论与建议

5.1.1 结论

(1)运用改进的逐步回归、岭回归和主成分-岭回归作对比，检验多重共线性修正效果，得出岭回归和主成分-岭回归在自变量的筛选中能尽可能保留较多变量的结论.

(2)运用统计软件建立时间序列模型、改进的逐步回归以及岭回归，对河南省生产总值预测值进行精确度对比，通过选择合适的岭参数 K 的岭回归模型预测效果较好.

(3)河南省地区经济发展的影响因素进行分析研究. 结合经济学有关理论分析结果如下:

对于河南省地区经济发展，三个主要影响因素为物质资本投入(固定资产投资总额)，技术进步水平(历年科学技术支出和专利申请量)，以及经济结构因素(城镇化率和二三产业占就业人数的比重).

分析说明近二十年技术发展水平，以及经济结构因素中的城镇化率，二三产业就业人数占总就业人数的比重这三个方面对地区生产总值影响日益显著，极大地促进河南省地区经济发展，可以继续加强城镇化水平，实现城镇一体化.

值得注意的是，人口增长率和就业人口数排名都处于末位，表明河南人口增长率对经济发展起阻滞作用，河南省经济增长和发展成果都被过量的新增人口消耗掉了，这是符合河南省实际情况. 岭回归中就业人口回归系数为负值说明对于就业问题的解决迫在眉睫. 可以加强对人才的重视和分配，特别是高层次人才的引进. 人力资源对河南省地区生产总值影响较小，并不代表不重要，考虑到河南是人口大省，如何实现把人口阻滞作用转化为人口优势，实现高质量发展，需要政策关注.

5.1.2 建议

为此，针对河南经济和社会实现高质量发展，提出以下几点对策：要维持在一个高质量的发展层次上，要从健全人力资源交易市场入手，拓宽劳动力的就业途径，重点解决劳动力的就业途径，并要加速对人才的培育，尤其是对高级人才的培育；持续推进我国城市化进程，深化区域协作，以“建设黄河流域开放门户”为核心，实现我国经济社会发展和社会经济发展的“三足鼎立”.

5.2 本文不足和局限

(1)由于时间和水平的限制，参考的外国文献并不多，所以导致国外文献没有国内文献丰富.

(2)本人只提出了主成分-岭回归的建模步骤和修正过程，因为软件量纲转化问题，并没有给出因变量的预测过程.

5.3 未来研究方向

(1)可以研究岭回归中运用岭计算法和岭迹图法确定岭系数的区别以及适用情况.

(2)可进一步运用主成分-岭回归对近几年地区生产总值经济相关的影响因素进行分析预测.

参考文献

[1] 黄文珂. 多元回归建模过程中共线性的诊断与解决方法[D]. 哈尔滨工业大学, 2012.

[2] 朱钰, 郑屹然, 尹默. 统计学意义下的多重共线性检验方法[J]. 统计与决策, 2020, 36(07): 34-36.

[3] 茆诗松, 程依明, 濮晓龙. 概率论与数理统计[M].2 版.北京: 高等教育出版社, 2011.

[4] 何晓群, 刘文卿. 应用回归分析(第 5 版) [M]. 北京: 中国人民大学出版社, 2019.

[5] D. E Farrar, R. R Glauber. Multicollinearity in Regression Analysis: The Problem Revisited [J]. Rev. Econ. Stat. 1967(01): 92-107.

[6] C. Mason, D. J Perreault. Collinearity, Power, and Interpretation of Multiple Regression Analysis [J]. J. Mar. Res. 1991, 28(03): 268-280.

[7] 肖霞, 伍兴国. 线性回归中多重共线性的几何解释[J]. 统计与决策, 2021, 37(21): 46-51.

[8] 刘明, 王仁曾. 基于 t 检验的逐步回归的改进[J]. 统计与决策, 2012, 28(06): 16-19.

[9] 肖雪梦, 张应应. 三种回归方法在消除多重共线性及预测结果的比较[J]. 统计与决策, 2015, 31(24): 75-77.

[10] 林石莲. 多重共线性修正方法的比较与应用研究[D]. 广东财经大学, 2016.

[11] 孔朝莉. 基于主成分回归的海南旅游业影响因素分析[J].统计与管理, 2019, 34(01): 111-114.

[12] A. E Hoerl, R.W Kennard. Ridge regression: Biased estimation for nonorth-ogonal problems [J]. Technometrics. 1970, 12(01): 55-88.

[13] 尤游, 刘苏兵. 岭回归和主成分回归下的芜湖市社会消费品零售总额实证研究[J]. 信阳农林学院学报, 2020, 30(03): 28-30.

[14] 王飞, 孙嘉聪. 多重共线性问题的岭回归实例[J]. 数学学习与研究, 2019, 38(20): 132-134.

[15] 丁先文, 袁红. 基于岭回归的江苏省财政收入估计[J]. 江苏理工学院学报, 2020, 12(06): 5-7.

[16] 林乐义. 岭回归在消除多重共线性中的应用[J]. 辽东学院学报(自然科学版), 2020, 27(04): 274-278.

[17] 孔朝莉, 李国徽. 基于 GM(1, 1)与主成分回归的海南 GDP 预测及其影响因素分析[J]. 数学的实践与认识, 2016, 46(17): 66-80.

[18] 蔡素丽. 多元线性回归模型应用实证分析[J]. 廊坊师范学院学报(自然科学版), 2017, 17(04): 5-8.

[19] 刘芳, 董奋义. 计量经济学中多重共线性的诊断及处理方法研究[J]. 中原工学院学报, 2020, 31(1): 44-45.

[20] 魏红燕. 回归分析中多重共线性的诊断与处理[J]. 周口师范学院学报, 2019, 36(02): 11-15.

[21] 杨楠. 岭回归分析在解决多重共线性问题中的独特作用[J]. 统计与决策, 2004(03): 14-15.

[22] 刘立祥. 线性回归模型中自变量的选择与逐步回归方法[J]. 统计与决策, 2015, 12(21): 80-83.

[23] 游士兵, 严研. 逐步回归分析法及其应用[J]. 统计与决策, 2017(14): 31-35.

[24] 李珊珊, 刘越. 基于 MATLAB 的逐步回归分析在体育研究中的应用[J]. 长春大学学报, 2020, 30(12): 25-30.

[25] 河南省统计局, 国家统计局调查总队. 河南统计年鉴[M]. 郑州: 中国统计出版社, 2021.

[26] 高晓红, 李兴奇. 多元线性回归模型中无量纲化方法比较[J]. 统计与决策, 2022, 38(06): 5-8.

[27] 王琪, 冷林峰. 改进岭回归与主成分回归的股指跟踪研究[J]. 重庆理工大学学报(自然科学), 2018, 32(01): 212-221.

[28] 尹康. 常用统计软件关于岭回归计算原理的比较分析[J]. 统计研究, 2013, 30(02): 109-110.

[29] 朱海龙, 李萍萍. 基于岭回归和 LASSO 回归的安徽省财政收入影响因素分析[J]. 江西理工大学学报, 2022, 43(01): 59-65.

[30] 耿建军. 基于岭回归的组合预测包容检验[J]. 山西师范大学学报(自然科学版), 2016, 30(03): 24-28.

[31] 王纯杰, 温丽男. 基于岭回归和 Lasso 回归的螺纹钢期货价格实证分析[J]. 吉林师范大学学报(自然科学版), 2020, 41(01): 36-41.

[32] 孙嘉聪. 岭估计法解决线性回归模型的多重共线性问题[D]. 渤海大学, 2020.

[33] 谢婼青, 姜国麟. 运用重复交易模型编制综合指数中的多重共线性问题: 基于参数改进的解决方法[J]. 系统工程理论与实践, 2022, 42(06): 1434-1447.

[34] O. Harrison, R. Etaga. Effect of Multicollinearity on Variable Selection in Multiple Regression[J]. Sci. J. Appl. Math. Stat. 2021, 9(06): 153.

[35] M. J Ray, K. Nimon. Using commonality analysis in multiple regressions: a tool to decompose regression effects in the face of multicollinearity [J]. Methods. Ecol. Evol. 2014, 5(04): 320-328.

[36] I. Pardoe . Applied regression model [M]. John Wiley & Sons, 2020.

[37] 赵囡. 2018 年～2020 年河南省 GDP 预测研究——基于 ARIMA 模型[J]. 洛阳理工学院学报(社会科学版), 2018, 33(03): 24-29.

[38] 董小刚, 刁亚静. 岭回归、LASSO 回归和 Adaptive-LASSO 回归下的财政收入因素分析[J]. 吉林师范大学学报(自然科学版), 2018, 39(02): 45-53.

[39] 熊幼林. 病态线性回归模型系数的主成分——岭估计[J]. 数学学习与研究, 2014, 33(09): 121-122.