Advances in Social Sciences 社会科学前沿, 2017, 6(4), 441-451
Published Online April 2017 in Hans. http://www.hanspub.org/journal/ass
https://doi.org/10.12677/ass.2017.64061

Hans汉斯

Analysis the Influence Factors of Steel Demand in China

Xiaojun Sun

School of Statistics and Mathematics, Yunnan University of Finance and Economics, Kunming Yunnan
Email: sxj920709@163.com

Received: Apr. 7th, 2017; accepted: Apr. 21st, 2017; published: Apr. 30th, 2017

Abstract

Steel industry plays an important role in the national economy, which is widely used in economic construction, national defense construction, and other aspects of social development, but with the development of the national economy and the adjustment of industrial structure, long-term and extensive development pattern in China made great impact on the steel industry, fine direction so that China's steel industry development is imminent. This article uses the nine factors, including crude oil output, output of raw coal, natural gas production, cement production, pig iron production, power generation, the whole society fixed assets investment, consumer and government spending, to demand for steel products in China were analyzed, and establish three models, including the general linear regression model, the stepwise regression model and three Lasso regression model regression model. By comparison, find the Lasso regression model prediction effect is best, and according to the given model for the development of steel are proposed.

Keywords

Steel Demand, Multiple Regression, Stepwise Regression, Lasso Regression

影响我国钢材需求量的因素分析

孙小军

云南财经大学统计与数学学院，云南 昆明
Email: sxj920709@163.com

收稿日期：2017年4月7日；录用日期：2017年4月21日；发布日期：2017年4月30日

摘　要

钢材工业在国民经济中起着举足轻重的作用，它被广泛用于经济建设、国防建设、社会发展等方面。但

文章引用: 孙小军. 影响我国钢材需求量的因素分析[J]. 社会科学前沿, 2017, 6(4): 441-451.
https://doi.org/10.12677/ass.2017.64061

随着我国国民经济的快速发展和产业结构的调整，我国长期粗放发展模式使钢材工业受到很大冲击，使我国钢材工业向着精细化方向发展已迫在眉睫。本文使用原油产量、原煤产量、天然气产量、水泥产量、生铁产量、发电量、全社会固定资产投资额、居民消费、政府消费等9个因素，对我国成品钢材需求量进行分析，建立一般线性回归模型、逐步回归模型和Lasso回归模型等三个回归模型，通过比较，发现Lasso回归模型的预测效果最好，并根据所给模型对今后钢材发展提出了建议。

关键词

成品钢材需求量，多元线性回归，逐步回归，Lasso回归

1. 引言

这几年以来，随着我国工业的快速发展、产业结构的调整和我国国民经济的快速发展，我国钢铁工业不仅在数量上取得了较快增长，而且在质量、技术经济、准备水平、节能与环保等诸方面也取得了巨大成就，钢铁工业成为了非常具有竞争力的工业，为我国国民经济的快速发展做出了突出贡献。钢铁行业的稳定发展是实现我国新型工业化战略目标的关键一环，其发展水平的高低是衡量我国工业化水平和综合国力高低的重要标志。

《钢铁产业调整和振兴规划》(2009)表明，我国是钢铁生产和消费大国，粗钢产量连续 13 年位居世界第一，钢铁产业涉及面广、产业关联度高、消费拉动大，在经济建设、社会发展、国防建设、财政税收以及稳定就业等诸多方面发挥着至关重要的作用。21 世纪以来，我国粗钢产量年均以 21.1%的速度迅速增长。2007 年，我国钢铁规模以上企业完成工业增加值为 9936 亿元，占我国 GDP 的 4%，实现利润 2436 亿元，占我国工业利润总额的 9%；到 2008 年，国内粗钢消费量为 4.53 亿吨，直接出口钢材产品折合为粗钢 6000 万吨，占世界钢铁贸易总量的 15%。钢材产品大体上可以满足国内需要，对保障国民经济又好又快发展做出了突出贡献。

但是，我国钢铁工业形势依然严峻，存在以下几个问题：一是投资盲目，产能过剩；二是企业分布不合理；三是资源控制力弱；四是流通方向混乱；五是创新能力不强；六是产业不集中，粗钢生产企业平均规模不足 100 万吨，排名前 5 的企业钢材产量仅占全国钢材总量的 28.5%。

为了解决上述的一些问题，学者们对钢铁需求量进行了深入研究：宝良，郗维强等人[1]运用计量经济模型，就相关的宏观因素对钢材的需求影响程度进行分析，得出了钢材消费与工业增加值、建筑业竣工面积之间的弹性关系，并根据模型进行了简单的预测；刘铁敏，周伟，王青等人[2]建立计量经济模型，从钢材历史统计数据分析相关因素对我国钢材需求量的影响，得到钢材需求量与相关因素的模型，并对模型进行预测；李凯、代丽华、韩爽等人[3]应用生长曲线模型，预测中国钢铁到达峰值点的时间以及拐点时间，并从不同的指标角度对结果进行分析；吴文东、吴刚、魏一鸣和范英等人[4]采用基于相关系数的组合预测方法对我国未来的成品钢材需求量进行预测；赵月红[5]通过对影响钢铁需求量的变量进行协整检验，说明它们之间存在长期的协整关系，建立了误差修正模型，并对钢材需求量进行了预测。通过查阅外文文献，发现国外预测钢材需求量的方法大概可以分为三种：第一种方法[6]是将钢材需求量看作是工业产值或者其他宏观经济变量的函数，建立模型研究钢材需求量；第二种方法[7]是采用向量自回归

模型预测钢材需求量；第三种方法[8]是利用使用强度技术预测钢材需求量。

基于以上背景，为解决我国钢材需求量存在的国内供需平衡基础不平衡、生产成本高、对环境破坏大等诸多问题，本文将对我国 1999 年到 2014 年的钢材需求量及相关因素进行分析，找到未来几年我国钢材需求量呈现何种趋势，预测下一年，甚至未来几年内我国钢材需求量，做到未雨绸缪，防患于未然。

2. 数据分析

2.1. 数据来源及变量

通过中华人民共和国国家统计局[9]，中国统计年鉴查阅到 1999 年到 2014 年 16 年间我国成品钢材需求量、原油产量、原煤产量、天然气产量、生铁产量、发电量、水泥产量、全社会固定资产投资额、居民消费、政府消费的数据，变量名称如表 1 所示。

2.2. 数据预处理

对于上述给定的数据，为了能更好的建立回归模型，首先需要对数据进行简单分析，从因变量对自变量的影响和样本之间的相关系数等方面来分析数据各自的变化情况以及它们相互之间的关系。

2.2.1. 因变量 y 对自变量影响分析

为了观察成品钢材与其他变量之间的关系，从而建立合适的模型，因此对成品钢材与每个自变量作了散点图，具体结果如图 1 所示。

从图 1 第一行中可以看出，每一个自变量对因变量 y 都存在一定的线性关系，并且线性关系较强，因此对它们建立多元线性回归模型是合适的。

2.2.2. 样本相关系数

为了进一步刻画各变量之间线性关系的强弱，给出了各变量之间的相关系数，相关系数(记为 $Corr$)不同的值代表的相关程度不同：

1) 若 $Corr=0$，表示没有线性关系；

2) 若 $Corr=1$，称为完全正相关，$Corr=-1$，称为完全负相关；

3) 若 $0<|Corr|<1$，则称有“一定程度”的线性关系，$|Corr|$ 越接近于 1，则线性相关程度越高，越接近于 0，则线性相关程度越低。

对于本文使用的数据，给出了各变量之间的样本相关系数，结果如表 2 所示。

Table 1. Information about each variable
表 1. 各变量说明

变量	变量说明(单位)
y	我国成品钢材需求量(万吨)
x_1	原油产量(万吨)
x_2	原煤产量(亿吨)
x_3	天然气产量(亿立方米)
x_4	生铁产量(万吨)
x_5	发电量(亿千瓦时)
x_6	水泥产量(万吨)
x_7	全社会固定资产投资额(亿元)
x_8	居民消费(亿元)
x_9	政府消费(亿元)

Table 2. The sample correlation coefficient
表 2. 样本相关系数

	y	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9
y	1	0.987**	0.984**	0.997**	0.991**	0.999**	0.998**	0.980**	0.983**	0.992**
x_1	0.987**	1	0.991**	0.985**	0.988**	0.990**	0.985**	0.951**	0.956**	0.972**
x_2	0.984**	0.991**	1	0.984**	0.992**	0.988**	0.981**	0.937**	0.947**	0.961**
x_3	0.997**	0.985**	0.984**	1	0.991**	0.996**	0.992**	0.973**	0.977**	0.988**
x_4	0.991**	0.988**	0.992**	0.991**	1	0.991**	0.984**	0.945**	0.952**	0.970**
x_5	0.999**	0.990**	0.988**	0.996**	0.991**	1	0.997**	0.975**	0.982**	0.990**
x_6	0.998**	0.985**	0.981**	0.992**	0.984**	0.997**	1	0.984**	0.987**	0.992**
x_7	0.980**	0.951**	0.937**	0.973**	0.945**	0.975**	0.984**	1	0.996**	0.993**
x_8	0.983**	0.956**	0.947**	0.977**	0.952**	0.982**	0.987**	0.996**	1	0.996**
x_9	0.992**	0.972**	0.961**	0.988**	0.970**	0.990**	0.992**	0.993**	0.996**	1

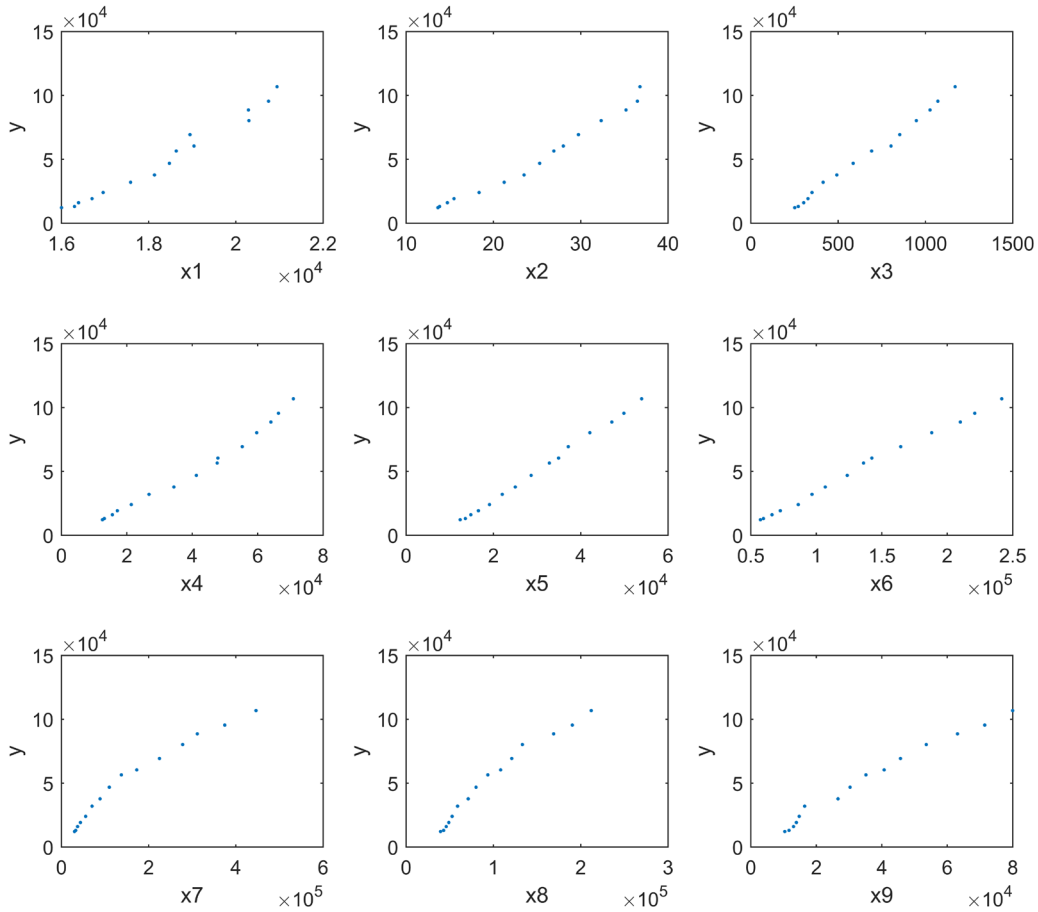

Figure 1. Scatter plot between the variables
图 1. 各变量之间散点图

从样本的相关系数表表 2 可以看出，各变量的相关系数都在 0.9 以上，根据相关系数的判别，说明成品钢材与自变量有着高度的线性相关性，适合做 y 与 9 个自变量的多元线性回归。

3. 模型构建

3.1. 多元线性回归模型

在上述问题中，中国成品钢材的需求量 y 的影响因素有原油产量(x_1)、原煤产量(x_2)、天然气产量(x_3)、生铁产量(x_4)、发电量(x_5)、水泥产量(x_6)、固定资产投资额(x_7)、居民消费(x_8)和政府消费(x_9)等，因此，可以采用多元线性回归进行问题的分析。

多元线性回归模型的基本形式[10]：设因变量 y 与自变量 $x_1, x_2, \cdots, x_p$ 的理论线性回归模型为：

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \varepsilon$$

其中， $\beta_0, \beta_1, \cdots, \beta_p$ 是 $p+1$ 个未知参数， β_0 称为回归常数， $\beta_1, \cdots, \beta_p$ 称为回归系数。 y 称为被解释变量(因变量)，而 $x_1, x_2, \cdots, x_p$ 是 p 个可以精确测量并可控制的一般变量，称为解释变量(自变量)。 ε 是随机误差，与一元线性回归一样，对随机误差项我们常假定其满足如下假设：

$$\begin{cases} E(\varepsilon) = 0 \\ \mathrm{var}(\varepsilon) = \sigma^2 \end{cases}$$

称

$$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \varepsilon$$

为理论回归方程。

写成矩阵形式为：

$$Y = X\beta + \varepsilon$$

式中

$$Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, X = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1p} \\ 1 & x_{21} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{bmatrix}, \beta = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_n \end{bmatrix}, \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

为了方便对多元回归方程模型进行参数估计，对回归方程有如下的一些基本假设：

1) 解释变量 $x_1, x_2, \cdots, x_p$ 是确定性变量，不是随机变量，且要求 $rank(X) = p+1 < n$ 。

2) 随机误差项具有零均值和等方差，即：

$$\begin{cases} E(\varepsilon_i) = 0, i = 1, 2, \cdots, n \\ \mathrm{cov}(\varepsilon_i, \varepsilon_j) = \begin{cases} \sigma^2, i = j \\ 0, i \neq j \end{cases}, i, j = 1, 2, \cdots, n \end{cases}$$

这个假定常称为高斯-马尔柯夫条件。

3) 正态分布的假定条件为：

$$\begin{cases} \varepsilon_i \sim N(0, \sigma^2), i = 1, 2, \cdots, n \\ \varepsilon_1, \varepsilon_2, \cdots \varepsilon_n \text{相互独立} \end{cases}$$

对于多元线性方程的未知参数，采用最小二乘估计方法，经整理后得到如下形式：

$$X'X\hat{\beta} = X'Y$$

基于上述的分析，对本文使用的变量建立如下多元线性回归模型：

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6 + \beta_7 x_7 + \beta_8 x_8 + \beta_9 x_9 + \varepsilon$$

使用 R 软件，编写相应的程序代码。为了便于后面的比较，这里的回归系数给出的是标准化后的系数，具体结果如表 3 所示：

根据表 3 给出的结果，得到回归方程为：

$$\hat{y}_{\text{linear}} = 0.079x_1 - 0.142x_2 + 0.058x_3 + 0.453x_4 + 0.390x_5 + 0.022x_6 + 0.291x_7 + 0.162x_8 - 0.303x_9$$

调整后的决定系数 $R^2 = 0.9998$，由决定系数看，回归方程高度显著；p值 $= 7.894 \times 10^{-10}$，表明回归方程高度显著，说明自变量 $x_1, x_2, \cdots, x_9$ 整体上对 y 有高度显著的线性影响。但是从回归系数的显著性检验看，在 0.05 的显著性水平下，只有 x_4, x_7 对 y 是显著的，其他的自变量都不显著。造成这种现象的原因可能是自变量 $x_1, x_2, \cdots, x_9$ 之间存在多重共线性，由于多重共线性的存在，利用普通最小二乘估计得到的回归参数估计值很不稳定，回归系数的方差随着多重共线性强度的增加而加速增长，就会造成回归方程高度显著的情况下，有些回归系数通不过显著性检验，甚至导致回归系数的正负号得不到合理的经济解释。

3.2. 多重共线性诊断

当回归方程的解释变量之间存在很强的线性关系，回归方程的检验高度显著时，有些与因变量 y 的简单相关系数绝对值很大的自变量，其回归系数不能通过显著性检验，甚至出现有的回归系数所带符号与实际经济意义不符，这时就认为变量间存在多重共线性。近年来，统计学家提出了许多判断多重共线性的方法，本文主要使用方差膨胀因子法来判断九个自变量之间是否存在多重共线性。

对自变量作中心标准化[10]，则 $X^{*\prime}X^{*} = (r_{ij})$ 为自变量的相关阵。记：

$$C = (c_{ij}) = \left(X^{*\prime}X^{*}\right)^{-1}$$

Table 3. Coefficient of linear regression model
表 3. 线性回归模型系数

Coefficients:				
	Estimate	Std. error	t value	Pr(>|t|)
x_1	0.079	1.807e+00	0.827	0.4460
x_2	−0.142	4.199e+02	−1.290	0.2535
x_3	0.058	6.146e+00	0.935	0.3925
x_4	0.453	2.198e−01	3.146	0.0255 *
x_5	0.390	5.117e−01	1.756	0.1395
x_6	0.022	5.737e−02	0.198	0.8510
x_7	0.291	1.888e−02	3.629	0.0151 *
x_8	0.162	1.257e−01	0.727	0.5001
x_9	−0.303	2.387e−01	−1.745	0.1415

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1. Residual standard error: 450.1 on 5 degrees of freedom Multiple R-squared: 0.9999, Adjusted R-squared: 0.9998. F-statistic: 7831 on 9 and 5 DF, p-value: 7.894e−10.

称其为主对角线元素 $VIF_j = c_{jj}$ 为自变量 x_j 的方差膨胀因子，则有：

$$\mathrm{var}\left(\hat{\beta}_j\right) = c_{jj}\sigma^2 / L_{jj},\ j = 1,2,\cdots,p$$

式中，L_{jj} 为 x_j 的离差平方和，由上式可知，用 c_{jj} 作为衡量自变量 x_j 的方差膨胀程度的因子是恰如其分的。记 R_j^2 为自变量 x_j 对其余 $p-1$ 个自变量的决定系数，则方差膨胀因子可以表示为：

$$VIF_j = c_{jj} = \frac{1}{1-R_j^2}$$

R_j^2 度量了自变量 x_j 与其余 $p-1$ 个自变量的线性相关程度，这种相关程度越强，说明自变量之间的多重共线性越严重，R_j^2 越接近 1，VIF_j 就越大。经验表明，当 $VIF_j \geq 10$ 时，就说明自变量 x_j 与其余自变量之间有严重的多重共线性，且这种多重共线性可能会过度地影响最小二乘估计值。

对本文的 9 个自变量，进行多重共线性检验，计算自变量 $x_1, x_2, \cdots, x_9$ 的方差膨胀因子，结果如表 4 所示。

从表 4 可看出，每个自变量的 VIF 值都很大，最小的方差膨胀因子 $VIF_3 = 269.26$，远远超过了 10，说明 9 个自变量之间存在严重的多重共线性，如果还按照一般的线性回归方法进行建模，所得到的预测结果会很不理想，为此需要先消除自变量之间的多重共线性，再对变量进行建模。

3.3. 消除多重共线性方法

通过上述的分析可知，9 个自变量之间存在严重的多重共线性，这会对回归预测产生严重的影响，为此在建模前需要消除变量之间的多重共线性。消除多重共线性的方法有很多，例如可以剔除一些不重要的解释变量、增大样本量、回归系数的有偏估计、逐步回归法、岭回归法、主成分回归法、偏最小二乘回归法等，本文主要通过逐步回归法和 Lasso 回归法来消除多重共线性的影响。

3.3.1. 逐步回归

逐步回归法[10] [11]是一种选择自变量最优子集的方法，该方法的基本思想是有进有出，具体做法是将变量一个个引入，每引入一个自变量后，对已选入的变量要进行逐个检验，当原引入的变量由于后面变量的引入而变得不再显著时，要将其剔除。引入一个变量或从回归方程中剔除一个变量，为逐步回归的一步，每一步都要进行 F 检验，以确保每次引入新的变量之前回归方程中只包含显著的变量，这个过程反复进行，直到既无显著的自变量选入回归方程，也无不显著的自变量从回归方程中剔除为止。

对本文的 9 个自变量，通过使用逐步回归法，来选择对成品钢材都显著的变量，从而建立回归模型，为了便于后面的比较，给出了标准化后的回归系数。逐步回归的结果如表 5 所示。

由表 5 可知，经过逐步回归后得到的回归方程为：

$$\hat{y}_{\text{step}} = -0.090x_2 + 0.387x_4 + 0.569x_5 + 0.326x_7 - 0.184x_9$$

由回归方程可以看出，剔除的变量有原油产量(x_1)、天然气产量(x_3)、水泥产量(x_6)、居民消费(x_8)等四个变量，对成品钢材需求量有显著性影响的自变量是原煤产量(x_2)、生铁产量(x_4)、发电量(x_5)、固定资产投资(x_7)和政府消费(x_9)，在其他条件不变的情况下，当原煤产量每增加一个单位，我国成品

Table 4. Variance inflation factor of independent variable
表 4. 自变量的方差膨胀因子

x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9
642.10	847.98	269.26	1460.47	3482.17	862.26	454.10	3490.11	2125.11

Table 5. The results of stepwise regression
表 5. 逐步回归结果

Step: AIC = 182.43 $y \sim x_2 + x_4 + x_5 + x_7 + x_9$				
	Estimate	Std. error	t value	Pr(>|t|)
x_2	−0.090	1.735e+02	−1.996	0.077077 .
x_4	0.387	5.744e−02	10.284	2.83e−06***
x_5	0.569	1.700e−01	7.710	2.97e−05***
x_7	0.326	8.401e−03	9.134	7.56e−06***
x_9	−0.184	6.519e−02	−3.872	0.003776**

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1. Residual standard error: 378.5 on 9 degrees of freedom. Multiple R-squared: 0.9999, Adjusted R-squared: 0.9999. F-statistic: 1.992e+04 on 5 and 9 DF, p-value: < 2.2e−16. Shapiro-Wilk normality test. data: b$res. W = 0.97312, p-value = 0.9013.

钢材需求量就会减少 0.090 个单位；当生铁产量每增加一个单位，成品钢材需求量增加 0.387 个单位；其他变量的解释也如此，变量的解释符合经济规律，说明该模型建立符合常理。

从表 5 中可以看出调整后的决定系数 $R^2 = 0.9999$，说明回归方程是显著的，在 0.05 的显著性水平下，除了 x_2，其他剩余的变量都通过了显著性检验，说明建立该模型比较合理。关于残差的 Shapiro-Wilk 正态性检验的 p 值为 0.9013，可以认为在 0.05 的显著性水平下不能拒绝残差来自正态总体的假定。说明该模型的建立符合线性回归模型的一般条件，可以用该模型来进行预测。

3.3.2. Lasso 回归

多元线性回归模型的矩阵形式为：$y = X\beta + \varepsilon$，参数 β 的普通最小二乘估计(OLS)为 $\hat{\beta} = (X'X)^{-1}X'y$，当自变量 x_j 与其余自变量间存在多重共线性时，$\mathrm{var}(\hat{\beta}_j) = c_{jj}\sigma^2 / L_{jj}$ 很大，$\hat{\beta}_j$ 就很不稳定在具体取值上与真值有较大的偏差，有时甚至会出现与实际经济意义不符的结果。针对出现多重共线性时，普通最小二乘法明显效果变坏的问题，学者们提出了岭回归法、Lasso 回归法等消除多重共线性的方法。

Lasso 回归[12]的基本思想如下：

假定自变量的数据矩阵 $X = \{x_{ij}\}$ 为 $n \times p$ 的，OLS 估计寻求那些使得残差平方和最小的系数 β，即：

$$\left(\hat{\alpha}^{(\mathrm{ols})}, \hat{\beta}^{(\mathrm{ols})}\right) = \arg\min_{(\alpha,\beta)} \sum_{i=1}^{n}\left(y_i - \alpha - \sum_{j=1}^{p} x_{ij}\beta_j\right)^2$$

Lasso 回归则需要一个惩罚项来约束系数的大小，在原理上和岭回归的想法有些类似，但 Lasso 回归法在惩罚项中添加的不是系数的平方而是其绝对值，即在约束条件 $\sum_{j=1}^{p}|\beta_j| \le s$ 下，系数需要满足下面的条件：

$$\left(\hat{\alpha}^{(\mathrm{lasso})}, \hat{\beta}^{(\mathrm{lasso})}\right) = \arg\min_{(\alpha,\beta)} \sum_{i=1}^{n}\left(y_i - \alpha - \sum_{j=1}^{p} x_{ij}\beta_j\right)^2$$

出于绝对值的特点，Lasso 回归的做法是筛选掉一些系数。对于回归系数的选择，本文使用 C_p 统计量，如果从 k 个自变量中选取 p 个（$k > p$）参与回归，那么 C_p 统计量的定义为：

$$C_p = \frac{SSE_p}{S^2} - n + 2p;\ SSE_p = \sum_{i=1}^{n}\left(Y_i - Y_{pi}\right)^2$$

据此，选择 C_p 最小的模型即为 Lasso 回归的最终模型。

对本文数据进行 Lasso 回归，所得结果如表 6 和表 7 所示。

由表 6 的 C_p 变化结果可知，最小的 $C_p = 6.7411$，故选择使 C_p 最小步的系数，即选择第 6 步的回归系数，第 6 步回归系数结果如表 7 所示。

由表 7，可以建立 Lasso 回归方程：

$$\hat{y}_{\text{lasso}} = 0.7849x_3 + 0.4506x_4 + 0.7821x_5 + 0.0477x_6 + 0.0474x_7$$

从回归方程可以看出，Lasso 回归剔除的变量有原油产量(x_1)、原煤产量(x_2)、居民消费(x_8)和政府消费(x_9)等四个变量，使用天然气产量(x_3)、生铁产量(x_4)、发电量(x_5)、水泥产量(x_6)和固定资产投资(x_7)五个变量来对我国成品钢材需求量进行建模。

4. 模型效果比较

根据第三节的分析，本文建立了三个回归模型，它们分别是：

$$\hat{y}_{\text{linear}} = 0.079x_1 - 0.142x_2 + 0.058x_3 + 0.453x_4 + 0.390x_5 + 0.022x_6 + 0.291x_7 + 0.162x_8 - 0.303x_9$$

$$\hat{y}_{\text{step}} = -0.090x_2 + 0.387x_4 + 0.569x_5 + 0.326x_7 - 0.184x_9$$

$$\hat{y}_{\text{lasso}} = 0.7849x_3 + 0.4506x_4 + 0.7821x_5 + 0.0477x_6 + 0.0474x_7$$

为了比较三个模型拟合效果的好坏，给出了三个模型对 1999 年-2014 年成品钢材需求量进行预测，预测结果和原始成品钢材需求量的值如图 2 所示。

从图 2 中可以看出，预测效果最好的是 Lasso 回归，和真实数据值比较接近，效果最差的是一般的线性回归模型，和真实值的偏差比较大。

Table 6. The change of C_p value in the Lasso regression
表 6. Lasso 回归中 C_p 值的变化情况

step	Df	Rss	C_p
0	1	1.4276e+10	70467.4485
1	2	4.6493e+09	22942.3026
2	3	3.2065e+09	15821.0543
3	4	2.4687e+08	1211.7678
4	5	5.5538e+07	269.1838
5	6	3.2116e+06	12.8551
6	7	1.5680e+06	6.7411
7	8	1.4191e+06	8.0061
8	9	1.1358e+06	8.6074
9	10	1.0128e+06	10.0000

Table 7. Regression coefficient of Lasso regression
表 7. Lasso 回归的回归系数

x_3	x_4	x_5	x_6	x_7
0.7849	0.4506	0.7821	0.0477	0.0474

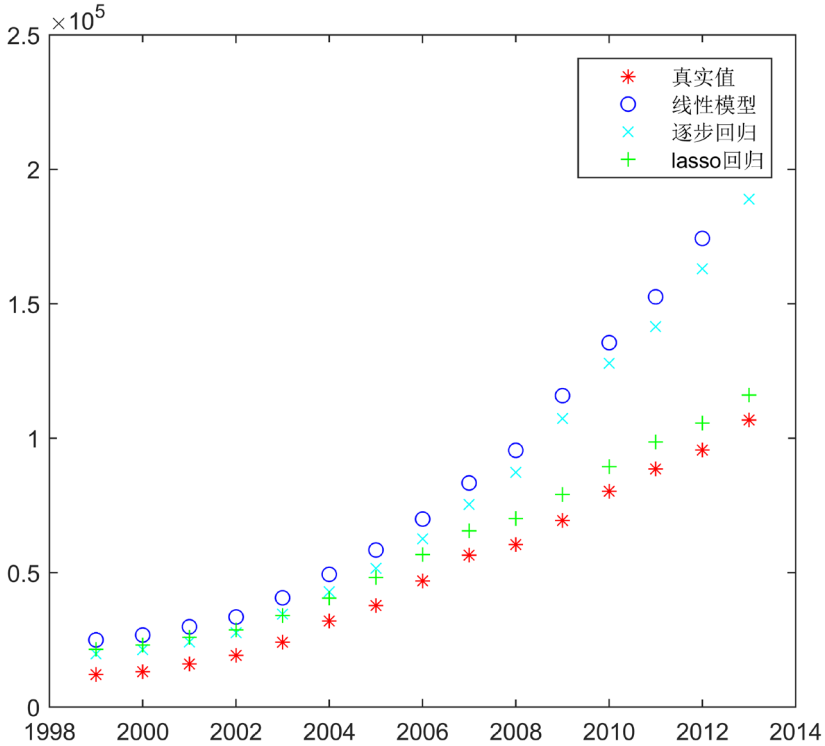

Figure 2. Comparison on steel demand forecast for three models
图 2. 三个模型对成品钢材需求量的预测值比较

因此对于我国成品钢材需求量的分析，可以通过天然气产量(x_3)、生铁产量(x_4)、发电量(x_5)、水泥产量(x_6)和固定资产投资(x_7)五个变量来建立模型，并由该模型来预测今后几年的成品钢材需求量。

5. 结论与建议

5.1. 结论

本文通过问题分析及假设建立了一般线性回归模型、逐步回归模型和 Lasso 回归模型，发现使用天然气产量、生铁产量、发电量、水泥产量和固定资产投资这五个变量就能对我国钢材需求量进行很好地分析。以此同时，为完成对未来我国钢材需求量的预测，需要对其相关变量的未来可能变化趋势进行深入分析，并以此为依据运用本模型进行钢材需求量的预测，变量的未来取值可以根据变量增长情况结合我国实际情况进行估计。

5.2. 建议

1) 在结构调整方面，应通过依法强制性推行节能减排目标和严格税收征管，加快淘汰落后产能，加速推进钢铁产业装备和技术结构升级，推进产业结构升级，在依靠法律手段的同时还要体现市场竞争的作用，避免钢铁产业越淘汰产能越大，而是要依靠竞争让企业体会到产品结构升级的好处。此外还要通过加快钢铁企业联合重组实现结构调整、提高我国钢铁行业自主创新能力和行业自律协调能力。

2) 在节能环保方面，应制定科学的管理细则，提高废铁使用比例，减少能耗，同时引进新技术提高生产率，加大对自然环境造成影响的企业的处罚，同时建立各种市场的预警、预测系统，使市场参与者能科学地判断市场运行效率和风险。

参考文献 (References)

[1] 宝良, 郗维强. 我国钢材需求分析研究[J]. 内蒙古科技与经济, 2006(4): 2-5.

[2] 刘铁敏, 周伟, 王青. 依据钢铁生产过程的中国铁矿石需求预测模型[J]. 金属矿山, 2007(2): 2-4.

[3] 李凯, 代丽华, 韩爽. 应用生长曲线模型预测中国钢铁工业的峰值点[J]. 冶金经济与管理, 2005(2): 1-3.

[4] 吴文东, 吴刚, 魏一鸣. 基于相关系数的钢材需求量组合预测[J]. 中国管理科学, 2008(16): 45-49.

[5] 赵月红. 中国钢材需求量的预测研究[J]. 内蒙古科技与经济, 2008(6): 320-323.

[6] Ghosh, S. (2006) Steel Consumption and Economic Growth: Evidence from India. *Resources Policy*, **1**, 7-11.

[7] Yu, B., Qiao, Y., Li, X., *et al*. (2014) Low-Carbon Transition of Iron and Steel Industry in China: Carbon Intensity, Economic Growth and Policy Intervention. *Journal of Environmental Sciences*, **16**, 10-16.

[8] Crompton, P. (2009) Future Trends in Japanese Steel Consumption. *Resources Policy*, **22**, 103-104.

[9] http://www.stats.gov.cn/tjsj/ndsj/

[10] 何晓群, 刘文卿. 应用回归分析(第二版)[M]. 北京: 中国人民大学出社, 2007: 18-92.

[11] 吕海燕, 李海燕, 李武林. 基于逐步回归分析的粮食产量因素研究[J]. 河南科学, 2013(12): 1-3.

[12] 吴喜之. 复杂数据统计方法——基于 R 的应用[M]. 北京: 中国人民大学出版社, 2015.