Statistics and Application 统计学与应用, 2023, 12(2), 543-549
Published Online April 2023 in Hans. https://www.hanspub.org/journal/sa
https://doi.org/10.12677/sa.2023.122058

Hans汉斯

基于多元回归分析的成品钢材需求量影响因素实证分析

严　玮

华南师范大学数学科学学院，广东 广州

收稿日期：2023年3月27日；录用日期：2023年4月17日；发布日期：2023年4月29日

摘　要

本文引入了中国统计年鉴的2000~2021年各年份的我国钢材需求量、原油产量、生铁产量、原煤产量、发电量、铁路货运量、固定资产投资额、居民消费、政府消费、GDP、工业增加值11个不同的量来进行回归分析。通过建立回归模型充分说明成品钢材需求量与其他10个变量的关系，建立了多个回归模型，再选择相对最优模型，最后通过所建立的最优模型分析影响成品钢材需求量的因素。

关键词

多元线性回归，成品钢材，多元加权最小二乘估计，异方差性，多重共线性

Empirical Analysis of Influencing Factors of Finished Steel Demand Based on Multiple Regression Analysis

Wei Yan

Institute of Mathematical Sciences, South China Normal University, Guangzhou Guangdong

Received: Mar. 27th, 2023; accepted: Apr. 17th, 2023; published: Apr. 29th, 2023

Abstract

In this paper, 11 different quantities of China's steel demand, crude oil output, pig iron output, raw coal output, power generation, railway freight volume, fixed asset investment, resident consumption, government consumption, GDP and industrial added value from 2000 to 2021 are introduced in the China Statistical Yearbook for regression analysis. By establishing a regression model to

文章引用: 严玮. 基于多元回归分析的成品钢材需求量影响因素实证分析[J]. 统计学与应用, 2023, 12(2): 543-549.
DOI: 10.12677/sa.2023.122058

fully explain the relationship between finished steel demand and other 10 variables, a number of regression models are established, and then the relative optimal model is selected. Finally, through the established optimal model it analyzes the factors affecting the finished steel demand.

Keywords

Multiple Linear Regression, Finished Steel, Multivariate Weighted Least Squares Estimate, Multicollinearity, Heteroscedasticity

1. 引言

钢铁在各行各业都被广泛使用，也是迄今为止，在世界范围内金属材料中被使用最多的一种。一个国家钢铁工业的水平直接对一个国家的工业化基础有重要的影响和决定作用。作为基础工业之一的钢铁工业，对于一个国家的国民经济来说，是极其重要的支撑，无论是在一些相关的基础建设的使用中，还是在与我们息息相关的日常生活中，都离不开钢铁。在过去的十年中，我国钢铁业迅速发展，钢铁已经成为了我国所需要的重要工业基础物资。根据官方统计报告显示，我国钢铁出口量在 2017 年已经达到了全球第一名，进口量达到了全球第十一名，由此可见，钢铁行业的生产、销售及进口、出口，都密切关系着我国经济水平及发展情况。

随着全球经济一体化和全世界工业化进程提速，钢铁的重要性愈来愈重要尤其对于重工业较为落后的国家，自身生产的钢铁量不足以满足本国基础建设的需要，因此不得不选择从国外进口，这就对全球的贸易往来产生了积极作用，推动全球经济一体化的进程不断加快。无论是经济处于领先地位的发达国家，还是正在快速发展的发展中国家，都存在对钢铁相关的研究，其中包括对钢铁价格需求影响因素的研究，也包括对整个钢铁市场的研究。

在模型预测方面，多元回归模型的分析与挖掘作用被极其广泛地运用在各个主题。王春辉，周生路[1]以江苏省为例，运用多元回归方法对江苏省的粮食产量进行了预测，得出了合理的结果。Taylor G. Don [2]对全国卡车货运量进行预测时，使用了经济指标作为影响变量建立回归模型。

在钢材需求量的预测的实证上，王志孟、陶雪良[3]从人均消耗量、钢铁积蓄量、部门钢材消费量以及居住条件的改善潜力四个测度，利用三种不同的方法，分别在主观和模型层次预测了我国 2000 年的钢材需求量。万洁雯[4]利用现期可以获取到的 2000 年至 2020 年上半年的以季度为单位的数据，以我国钢材销售量为核心分别构建多元回归模型、ARIMA 模型，对我国 2020 年钢材需求量前两季度数据进行预测。

从现有的文献来看，国内对于我国钢材需求量的研究方向主要是钢材需求量的预测，较少有研究钢材需求量的影响因素，其中主要采用回归分析、ARIMA 模型进行钢材需求量的预测。本文将采用回归分析研究钢材需求量的影响因素。

钢材是工业建设和经济发展不可或缺的重要物资。本文选取了 2000 年至 2021 年的钢材需求量以及对钢材需求量可能存在显著影响的多个变量的数据作为主要研究对象，对影响中国成品钢材需求量进行回归分析，并且建立多个回归方程，再选择相对最优模型，最后分析影响成品钢材的因素。

2. 变量设置

理论上认为影响成品钢材的需求量的因素主要有经济发展水平、收入水平、产业发展、人民生活水平提高、能源转换技术等因素。为此，本文收集了我国成品钢材的需求量，选择与其相关的十个因素：原油产量 x_1 (万吨)、生铁产量 x_2 (万吨)、原煤产量 x_3 (亿吨)、发电量 x_4 (亿千瓦时)、铁路货运量 x_5 (万吨)、固定资产投资额 x_6 (亿元)、居民消费 x_7 (亿元)、政府消费 x_8 (亿元)、GDPx_9 (亿元)、工业增加值 x_{10} (亿元)作为解释变量，旨在通过建立这些经济变量的线性模型来说明影响成品钢材需求量的原因。

x_1 是指还没有经过加工处理的石油产量。因为目前国内钢材严重依赖国外矿石进口，而进口矿石运输主要依赖海运，海运的最重要费用就是产生在轮船耗油问题，原油涨价，汽油，柴油等相关轮船用油必然上涨，进而带动海运费用上涨，然后带动进口矿石成本增加，从而导致钢厂成本增加，钢材生产成本增加势必对钢材价格带来一定影响，进而影响钢材需求量，故本文选取原油产量探索对钢材需求量的影响。

x_2 是指含碳量大于 2%的铁碳合金的产量。

x_3 是指包括无烟煤、烟煤、褐煤，不包括石煤的产量。

x_4 是指发电机进行能量转换产出的电能数量。

x_5 的全程是铁路货物运输量，是指铁路货物运输量。

x_6 是以货币表现的建造和购置固定资产的工作量以及与此有关的费用的总称。反映固定资产投资规模、速度和投资比例关系的综合性指标。国家规定投资计划和控制投资规模的重要依据。

x_7 是指常住住户对货物和服务的全部最终消费支出。

x_8 是指政府部门为全社会提供公共服务的消费支出以及免费或以较低价格向住户提供的货物和服务的净支出。

x_9 是指国内生产总值 GDP，也就是我国所有常驻单位在一个季度内的生产活动按照我国市场价格测算的最终成果。在过去四十多年的时间里，我国 GDP 增长迅速，同时，我国工业化水平也随着 GDP 的增长不断提高，GDP 可以反映经济大形势的好坏程度，GDP 的增长与钢材产量息息相关，前者对后者有较大的冲击效应，前者的增长也对后者的增长有较大的积极作用和贡献度，加入该指标旨在探索经济大背景对钢材需求量的影响。

x_{10} 是指工业企业在生产活动中的成果扣除消耗和损失的部分后的价值。钢材消耗与工业生产密切相关，所以加入该指标希望从亲近指标测度上探索钢材需求量的变化趋势。

本文从国家统计年鉴上获取了 2000 至 2021 年共 22 年我国钢材需求量、原油产量、生铁产量、原煤产量、发电量、铁路货运量、固定资产投资额、居民消费、政府消费、GDP、工业增加值的年度数据。本文收集的数据均为定量变量，其符号和经济意义如下表 1 所示。

Table 1. Symbol description
表 1. 符号说明

符号	变量	代表意义	单位
y	中国成品钢材的需求量	成品钢材需求总量	万吨
x_1	原油产量	原油工业发展水平	万吨
x_2	生铁产量	生铁工业发展水平	万吨
x_3	原煤产量	原煤工业发展水平	亿吨
x_4	发电量	发电技术水平	亿千瓦时

Continued

x_5	铁路货运量	运输产业水平	万吨
x_6	固定资产投资额	固定资产支出水平	亿元
x_7	居民消费	居民支出水平	亿元
x_8	政府消费	政府支出水平	亿元
x_9	GDP	国内生产水平	亿元
x_{10}	工业增加值	工业生产水平	亿元

3. 多元回归模型的建立

3.1. 数据预处理

为了观察中国成品钢材需求量与已选取的其他变量之间的关系，从而建立出更合适的模型，因此先分别作出钢材需求量和每个已选取的自变量的曲线统计图和 Pearson 相关系数及其检验结果分别如图 1 和表 2 所示。

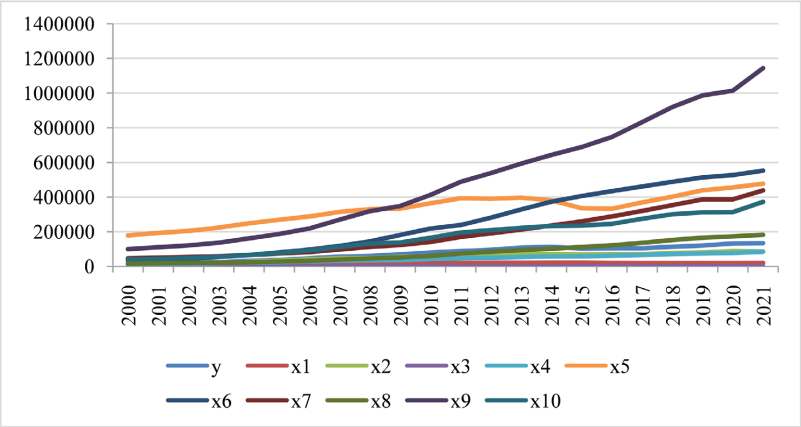

Figure 1. Statistical chart of curves of variables from 2000 to 2021
图 1. 2000 年至 2021 年各变量曲线统计图

Table 2. Phase relation table
表 2. 相关系数表

	y	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}
y	1	0.821	0.993	0.963	0.978	0.954	0.962	0.93	0.94	0.952	0.976

由图 1 可以大致的来看，$x_2, x_4, x_8, x_6, x_7, x_{10}$ 和因变量 y 在 2000 年到 2003 年的增长速度都相对平稳没有明显的增势；从 2003 年到 2008 年，个别变量开始缓慢增长；从 2008 年到 2019 年中旬，增长的幅度开始加大了；在 2019 年中旬到 2020 年中旬，由于受新冠疫情的影响，大多数变量都是平缓甚至降低的，但在 2020 年中旬开始又有了明显的增势。x_1, x_3 的曲线近似为一条水平直线，这两个变量分别表示原油和原煤的量，可能受到资源和政策的限制，因而增长的速度非常缓慢。从图 1 中可以明显看到，随着年限的增加，除了在 2020 年至 2021 年受新冠疫情的影响下部分量没有增长外，我国的各种产业和支出水平都随之逐渐增长。

由表 1 相关系数表可知，y 与除 x_1 以外的 9 个自变量 $x_2,x_3,x_4,x_5,x_6,x_7,x_8,x_9,x_{10}$ 的相关系数都在 0.9 以上，说明所选取的这 9 个自变量是与 y 高度线性相关的，且 y 与 x_1 的相关系数大于 0.8，说明 y 与 x_1 也高度线性相关，用 y 与自变量做多元线性回归是合适的。

3.2. 回归模型的初步建立

将原始数据导入到 spss 的数据框中，然后用 spss 软件回归线性分析可以得到 y 对 10 各自变量的线性回归方程为方程(1)。

$$\begin{aligned}\hat{y} = &-81090.351 + 4.304x_1 - 0.048x_2 - 915.505x_3 + 1.249x_4 \\ &+ 0.141x_5 + 0.151x_6 - 0.460x_7 + 0.544x_8 + 0.041x_{10}\end{aligned} \tag{1}$$

从回归方程中可以看出 $x_1,x_4,x_5,x_6,x_8,x_{10}$ 对成品钢材需求量起正影响，x_2,x_3,x_7 对成品钢材需求量起负影响。从实际社会生活来看，生铁生产水平、原煤生产水平和居民的消费水平提高，都会促进成品钢材的需求量，应该和成品钢材的需求量成正相关，这与定性分析的结果不一致。为此，本文对它进行更深层次的分析。

3.3. 回归拟合优度诊断

拟合优度可用于检验回归方程对样本观测值的拟合程度。回归方程(1)的复相关系数 $R = 0.999$，决定系数 $R^2 = 0.998 = 99.8\%$，由决定系数可知回归方程高度显著。

通过方差分析可知，F 检验值为 762.62，P 值等于 0.000，表明回归方程高度显著，说明 $x_1,x_2,x_3,x_4,x_5,x_6,x_7,x_8,x_9,x_{10}$ 整体上对 y 有高度显著的线性影响。

通过 T 检验可知，当显著性水平 $\alpha = 0.05$ 时，只有变量 x_1,x_6,x_7 的 P 值小于 0.05，通过了显著性检验。虽然自变量 $x_1,x_2,x_3,x_4,x_5,x_6,x_7,x_8,x_9,x_{10}$ 整体上对 y 有显著影响，但每个自变量对 y 的显著性却较差。其中 x_2 的 $P = 0.902$ 最大，不显著。x_3,x_4,x_{10} 的 P 值分别是 0.303，0.285，0.710，也不显著。x_5,x_8 的 P 值在 0.05~0.10 之间，也只是弱显著。由此可见，在多元线性回归中，虽然回归方程整体的显著性很强但是并不意味着每个自变量都显著。

另外，每个自变量的显著性和这些自变量与因变量 y 两两之间的简单相关系数的大小并不一致，产生这个问题的原因是自变量之间存在共线性。其中 x_2,x_3,x_7 的偏回归系数是负数，而因变量 y 与这三个自变量却是高度正相关，这也是共线性带来的问题。为此，在本文的后面还需对共线性问题进行分析与消除。

为了尽可能的保留合理变量，本文就针对逐个变量给以 T 检验分析，逐步剔除 P 值最大的不合理变量，使回归模型更完善。由此可以得到 y 对 4 个自变量的线性回归方程为公式(2)。由回归方程(2)中可以看到，对成品钢材需求量起正影响，对成品钢材需求量起负影响。此时回归方程虽然通过了 F，T 检验，但是增加了不合理变量所占回归的比重，这不符合社会实际。

$$\hat{y} = -67490.008 + 2.851x_1 + 0.181x_5 + 0.220x_6 - 0.148x_7 \tag{2}$$

3.4. 异方差性的检验

首先，本文分别以回归标准化残差和因变量 y 来绘制残差图分析模型是否存在异方差。

我们可以残差图中看出，回归的标准化残差在一开始随因变量 y 的变大，回归的标准化残差以 standardized Residual = 0 为轴对称向外变大，呈现喇叭口形状，因此我们可以初步判定初等回归方程(1)可能存在异方差。

其次，计算残差绝对值与自变量 x_i 的相关性时采用 Spearman 等级相关系数，而不采 Pearson 简单相关系数。这是因为级相关系数可以反映线性相关的情况，而简单相关系数不能如实反映非线性相关的情

况。计算残差绝对值与 x_i 的等级相关系数可知，等级相关系数 $r_{e1}=0.474$，P 值等于 0.026，认为残差绝对值与自变量 x_i 显著相关，存在异方差性。

3.5. 多元加权最小二乘估计

由于一般的多元线性回归方程出现异方差，故我们需要消除异方差性的影响，本文将使用应用较广泛的加权最小二乘法。首先，本文先选取权函数自变量。我们计算出普通残差的绝对值 $\mathrm{ABSE}=e_i$ 与 $x_1,x_2,x_3,x_4,x_5,x_6,x_7,x_8,x_9,x_{10}$ 的等级相关系数，其中残差绝对值与自变量 x_1 的相关系数为 $r_{e1}=0.474$，由于 x_1 的相关系数比其他自变量的都要大，因此我们选 x_1 构造权函数。其次，确定 Weight Eetlmate 估计幂指数 m，最后得到 m 的最优解为 $m=5$。最后，进行加权最小二乘估计拟合，加权最小二乘的回归方程为方程(3)。对回归方程(3)进行方差分析，有决定系数 $R^2=0.999$，$F=1156.949$，普通最小二乘的回归方程为方程(1)，y 有 $R^2=0.998$，$F=762.626$，通过对比两者的 R^2 和 F 可以说明加权最小二乘估计拟合效果略好于普通最小二乘的效果。因此选用加权最小二乘估计是正确合理的，但是 x_3,x_7 的系数都是负数，说明因变量 y 与自变量 x_3,x_7 呈负相关的关系，这与实际意义不符合。

$$\begin{aligned}\hat{y}=&-81979.390+4.606x_1+0.153x_2-904.965x_3+0.704x_4\\&+0.127x_5+0.135x_6-0.389x_7+0.532x_8+0.067x_{10}\end{aligned}\tag{3}$$

3.6. 自相关性的检验

对于自相关性我们用 DW 检验来判断，已知回归估计式的残差来定义 DW 统计量，假设有原假设 H_0 为 $\rho=0$，通过化简后 DW 值与 $\hat{\rho}$ 的关系式为 $\mathrm{DW}\approx 2(1-\hat{\rho})$。有 $\mathrm{DW}\approx 2(1-\hat{\rho})=2.467$，因而可以近似的计算出 $\hat{\rho}=-0.2335$，通过查表可以判断出误差项的自相关性呈轻微的负自相关。由于自相关性不是很明显，所以本文将不做处理。

3.7. 多重共线性的诊断

首先，本文先通过方差扩大因子 VIF 对方程(1)进行多重共线性的诊断。由表 3 可知 $x_2,x_3,x_4,x_6,x_7,x_8,x_9,x_{10}$ 的方差扩大因子 VIF 都很大，远远的超过了 10，说明成品钢材需求量的回归方程(1)存在着严重的多重共线性。又因为 $x_1,x_2,x_3,x_4,x_5,x_6,x_7,x_8,x_9,x_{10}$ 的方差扩大因子都是大于 10 的，说明回归方程的多重共线性就是由自变量间的多重共线性引起的。

Table 3. Variance enlargement factor VIF table
表 3. 方差扩大因子 VIF 表

	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_{10}
VIF	32.689	360.374	258.385	2645.576	164.911	680.100	2507.571	1200.501	517.291

接下来，为了消除方程的多重共线性，本文通过逐步剔除最大的 VIF 的自变量以消除方程的多重共线性。最后，得到剔除了自变量 $x_2,x_3,x_4,x_6,x_7,x_9,x_{10}$ 的新回归方程(4)，剩下的自变量 x_1,x_5,x_8 的方差扩大因子分别为 $\mathrm{VIF}_1=2.854,\mathrm{VIF}_5=7.710,\mathrm{VIF}_8=4.594$，都是小于 10 的，且回归系数也都有合理的社会经济解释，说明回归模型不存在强的多重共线性了，可以作为最终的回归模型。建立 y 与 x_1,x_5,x_8 的回归方程为方程(4)所示。

$$\hat{y}=-145918.473+8.229x_1+0.099x_5+0.406x_8\tag{4}$$

标准化的回归方程为方程(5)所示。

$$\hat{y} = 0.316x_1 + 0.211x_5 + 0.567x_8 \quad (5)$$

由标准化的回归方程(5)我们可以看到，对成品钢材需求量影响较大的是原油产量、铁路货运量和政府消费，其中政府消费的系数较大，影响也就较大。从整体上来看，消除多重共线性影响后得回归方程更较为符合社会实际。对方程(4)进行方差分析，有 $F = 1016.619$，P 值为 0.000，可知此回归方程仍然具有高度的显著。从方程(5)的样本决定系数 $R^2 = 0.994$，调整的样本决定系数 $R_a^2 = 0.993$，而方程(1)的样本决定系数 $R^2 = 0.998$，调整的样本决定系数 $R_a^2 = 0.997$，与方程(1)相比的方程(4)拟合优度仍然很高，并且回归系数有合理的经济解释。

4. 总结

随着社会经济的不断发展，科学技术的不断进步，统计方法越来越成为人们必不可少的工具手段。应用回归分析是其中的一个重要分支，本着国家经济水平的不断提高，我们采用回归分析的方法对我国成品钢材的需求量进行分析应用。我们首先建立了初等回归方程，对于初等回归模型是否违背原假设我们做了异方差性检验，自相关性检验，检验出了初等回归模型具有异方差性，故接下来使用了多元加权最小二乘估计消除异方差性。在分析中我们发现了变量之间存在共线性。因此我们对多重共线性进行了诊断，然后又通过逐步剔除最大的 VIF 的自变量对多重共线性给予了消除，重新建立了线性回归方程。我们建立了多个回归模型，最后发现通过逐步剔除最大的 VIF 的自变量的方式建立的多元回归方程 $\hat{y} = -145918.473 + 8.229x_1 + 0.099x_5 + 0.406x_8$ 是最合理的，由方程可知对成品钢材需求量影响较大的是原油产量、铁路货运量和政府消费，其中政府消费对钢材需求量影响最大。

参考文献

[1] 王春辉, 周生路, 吴绍华, 吴滢滢. 基于多元线性回归模型和灰色关联分析的江苏省粮食产量预测[J]. 南京师大学报(自然科学版), 2014(4): 105-109.

[2] Stock, J.R. (1993) Marketing Intelligence & Planning. *International Journal of Physical Distribution & Logistics Management*, **11**, 13-15. https://doi.org/10.1108/EUM0000000001124

[3] 王志孟, 陶雪良. 2000 年我国钢材需求量预测[J]. 冶金经济与管理, 1993(2): 9-11.

[4] 万洁雯. 后疫情时代我国钢材需求预测[D]: [硕士学位论文]. 南昌: 江西财经大学, 2021. https://doi.org/10.27175/d.cnki.gjxcu.2021.000757