

基于线性回归模型的粮食产量影响因素分析 ——以山东省为例

孙钦瑶, 牟唯嫣

北京建筑大学理学院, 北京

收稿日期: 2025年9月28日; 录用日期: 2025年10月22日; 发布日期: 2025年10月29日

摘 要

文章在考察山东省2000~2022年粮食总产量变化影响因素的基础上, 基于最小二乘线性回归模型分析了对粮食总产量有显著影响的因素, 结果认为居民人均消费与农业生产资料价格总指数两个解释变量是对粮食总产量具有显著影响的变量。另外对数据的复共线性分析结果, 考虑了主成分回归的线性回归方法, 验证了LASSO回归下变量选取结果, 对不同的回归结果做比较, 主成分分析回归的误差相对较小。

关键词

粮食产量, 线性模型, 最小二乘回归, 主成分回归

Analysis of Grain Yield Influencing Factors Based on Linear Regression Models —Taking Shandong Province as an Example

Qinyao Sun, Weiyan Mu

School of Science, Beijing University of Civil Engineering and Architecture, Beijing

Received: September 28, 2025; accepted: October 22, 2025; published: October 29, 2025

Abstract

Based on the examination of the influencing factors of the total grain output in Shandong Province from 2000 to 2022, this article analyzed the factors with significant influence on the total grain output using the least squares linear regression model. The results suggest that the per capita consumption of residents and the total index of agricultural production materials prices are the two explanatory variables that have a significant impact on the total grain output. Additionally, consid-

ering the multicollinearity analysis of the data, the principal component regression method was adopted. The variable selection results under LASSO regression were verified. By comparing different regression results, it was found that the error of the principal component analysis regression was relatively small.

Keywords

Grain Yield, Linear Model, Least Squares Regression, Principal Component Regression

Copyright © 2025 by author(s) and Hans Publishers Inc.
This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).
<http://creativecommons.org/licenses/by/4.0/>



1. 引言

党的十八大以来，习近平总书记高度重视国家粮食安全，始终坚持把解决好十几亿人口的吃饭问题作为治国理政的头等大事。粮食是关系到城乡每个人基本生活的特殊商品，在整个物价体系中具有基础地位，因此对人民生活影响最直接、最明显。我国是一个发展中国家而且是一个人口大国，粮食产量直接影响着国家的粮食安全问题。

在经济学中，Solow (1957)与 Hayami & Ruttan (1985)指出粮食生产常被纳入农业生产函数的分析框架(如 Cobb-Douglas 生产函数)，其产出受土地、劳动、资本、技术等多重因素影响[1][2]。国内外学者围绕粮食产量影响因素开展了大量实证研究。例如，王学真等(2006)指出山东省粮食生产波动受播种面积、单产水平和农业政策影响显著[3]；孙宏业等(2020)通过对内蒙古农产品价格的分析，强调了价格指数对农业生产的重要调节作用[4]；唐军(2023)则从玉米价格波动角度揭示了市场机制对粮食供给的影响[5]。

本文基于生产函数理论，结合已有研究，从生产要素投入、市场价格、消费需求等多个维度选取解释变量。并通过构建线性回归模型，旨在识别出对山东省粮食总产量具有显著影响的关键因素，为粮食安全政策提供量化依据。

2. 变量选择与解释

2.1. 变量选择与数据来源

本文根据研究目的，在已有研究基础上，确定响应变量为粮食总产量(万吨)，解释变量为居民消费价格总指数(上年 = 100)、居民人均消费(元)、商品零售价格总指数(上年 = 100)、农业生产资料价格总指数(上年 = 100)、农业总产值(亿元)、粮食单产(千克/公顷)、农作物播种面积(公顷)、农业机械年末总动力(万千瓦)。

自《山东省统计年鉴 2001~2023》中提取对应数据详细数据见表 1。

Table 1. Data
表 1. 数据

年份	粮食总产量 (万吨)	居民消费价格总指数 (上年 = 100)	居民人均消费(元)	商品零售价格总指数(上年 = 100)	农业生产资料价格总指数(上年 = 100)	农业总产值 (亿元)	粮食单产(千克/公顷)
表示	y	x_1	x_2	x_3	x_4	x_5	x_6
2000	3838	100	2615	99	99	1300	4938

续表

2001	3721	102	2798	100	102	1401	5201
2002	3293	99	2989	99	100	1421	4763
2003	3436	101	3288	100	102	1599	5355
2004	3517	104	3678	103	110	1892	5570
2005	3917	102	4228	101	106	2034	5837
2006	4093	101	4841	101	103	2283	5848
2007	4108	104	5646	104	107	2589	5896
2008	4354	105	6436	105	119	2863	6086
2009	4443	100	6978	99	96	3170	6088
2010	4503	103	7809	103	103	3588	6043
2011	4701	105	9057	105	111	3737	6172
2012	4816	102	10,100	102	106	3829	6214
2013	4883	102	11,074	101	101	4336	6099
2014	5038	102	12,517	101	100	4556	6087
2015	5147	101	14,065	100	99	4663	6123
2016	5332	102	15,392	101	99	4388	6261
2017	5374	102	16,732	101	101	4403	6356
2018	5320	103	18,161	102	108	4678	6329
2019	5357	103	19,511	102	106	4914	6444
2020	5447	103	20,032	102	102	5168	6577
2021	5501	101	23,900	101	103	5815	6584
2022	5544	102	23,638	102	104	6207	6620
$\max_j(x_{ij})$	5544	105	23,900	105	119	6207	5544
$\min_j(x_{ij})$	3293	99	2615	99	96	1300	3293

2.2. 数据预处理

2.2.1. 数据插补

《年鉴》指出，农业生产资料价格总指数数据自 2020 年起不再计入年鉴统计范围，因而要对数据进行插补。为增强插补结果的稳健性，本文尝试了两种常用插补方法：回归插补法与时间序列移动平均预测法。通过比较两种方法在 2010~2019 年数据的回代均方误差 MSE (表 2)，移动平均法表现最优，且考虑宏观经济周期性与数据的实际观察值性质，考虑用移动平均的方式进行数据插补。

Table 2. Comparison of interpolation methods MSE

表 2. 插补方法 MSE 对比

方法	回归插补法	时间序列移动平均预测法
MSE	44.84	21.48

插补公式为：

$$x_{k+1} = \frac{\sum_{i=k-N+1}^k x_i}{N}$$

插补后 2020 年~2022 年数据见表 1。

2.2.2. 数据标准化

由于选择变量单位不一, 量级不同, 直接进行回归会对参数有较大的影响, 甚至可能直接影响回归结果。为保证线性回归模型的精度与说服力, 需要对数据进行标准化处理, 去除量纲的影响。

对数据进行标准化的方式有很多, 王松桂等[6]在《线性模型引论》中提出的中心标准化的方式为

$$z_{ij} = \frac{x_{ij} - \bar{x}_j}{\sqrt{\sum_{i=1}^n (x_{ij} - \bar{x}_j)^2}}$$

然而观察原始数据, 解释变量农作物播种面积(公顷)数据量级过大、极差过大, 如果运用此种方式进行中心标准化会令分母部分过于大而导致结果过于小, 不利于回归分析。综合考虑后, 本文数据适合运用极值标准化的方式处理, 具体方式为

$$z_{ij} = \frac{x_{ij} - \min_j(x_{ij})}{\max_j(x_{ij}) - \min_j(x_{ij})}$$

3. 模型建立

3.1. 线性模型介绍

线性模型是一类统计模型的总称, 包括线性回归模型、方差分析模型、协方差模型和线性混合效应模型等, 通常用于表示一个或者多个解释变量与响应变量之间的线性关系, 是研究相关关系的重要工具, 本文所用模型即为线性模型中的线性回归模型。

含有个解释变量的线性回归模型的一般形式为

$$Y = \beta_0 + \beta_1 X_1 + \dots + \beta_{p-1} X_{p-1} + e$$

本文选用共 8 个解释变量, 即, 选用 2000~2022 共 23 组观测值, 对模型进行实证分析。

3.2. 线性模型的检验

模型的建立需要满足线性回归模型的基本假设, 包括解释变量之间相互独立, 残差之间相互独立且满足正态分布, 残差需保证方差齐性。

3.2.1. 独立性检验

德宾 - 沃森检验(Durbin-Watson Test)是一种针对数据自相关性程度进行的检验, 验证数据之间是否存在自相关关系。德宾 - 沃森指数(Durbin-Watson Statistic)的取值范围为 0~4。当德宾 - 沃森指数小于 2 时, 认为样本数据存在正自相关性; 当德宾 - 沃森指数大于 2 时, 认为样本数据存在负自相关性; 德宾 - 沃森指数值在 1.5~2.5 时, 可以认为样本数据不存在自相关性, 尤其德宾 - 沃森指数等于 2 时, 认为样本数据不存在自相关性, 即样本数据是相互独立的。

借助 SPSS 软件对数据进行线性回归计算, 由表 3, 计算结果中的 DW 检验值为 1.848, 介于 1.5 与 2.5 之间, 可以认为样本数据通过独立性检验。

Table 3. Model testing

表 3. 模型检验

模型	R	R方	调整后R方	标准估算的错误	R方变化量	F变化量	显著性F变化量	德宾 - 沃森
1	0.983 ^a	0.965	0.946	0.0580512	0.965	48.935	0.000	1.848

3.2.2. 变量的共线性检验

方差膨胀因子(VIF)是检验变量之间复共线性程度的一个重要参考，表达式为

$$VIF_i = \frac{1}{1 - R_i^2}$$

其中 R_i^2 变量与其他解释变量间的复相关系数。如果复共线性程度较高，说明解释变量之间具有较高的相关性，而较高的复共线性会影响模型的效果，干扰模型的判断。解释的一般规则是， $VIF < 3$ 则认为该变量与其他变量之间不存在复共线性， VIF 在 3~10 认为该变量与其他变量之间存在中等复共线性， VIF 大于 10 时认为该变量与其他变量之间存在潜在的严重复共线性。

借助 SPSS 软件对观测数据进行共线性诊断可以看出(表 4)，居民消费价格总指数 x_1 、农业生产资料价格总指数 x_4 、农业总产值 x_5 、粮食单产 x_6 、农业机械年末总动力 x_8 四个解释变量与其他解释变量之间都存在严重的复共线性。居民人均消费 x_2 、商品零售价格总指数 x_3 、农作物播种面积 x_7 三个变量与其他解释变量之间存在中等程度的复共线性。

Table 4. Collinearity test

表 4. 共线性检验

模型	共线性统计	
	容差	VIF
(常量)	0.007	138.017
居民消费价格总指数 x_1	0.100	9.986
居民人均消费 x_2	0.126	7.930
商品零售价格总指数 x_3	0.004	249.163
农业生产资料价格总指数 x_4	0.017	57.894
农业总产值 x_5	0.024	41.347
粮食单产 x_6	0.129	7.723
农作物播种面积 x_7	0.044	22.525

3.3. 线性模型的建立

在前文的检验中得出，解释变量之间的复共线性程度较为严重，故而在线性回归模型中可能不适宜用最小二乘法进行参数估计。因而在建立模型时，分别建立最小二乘回归模型与更针对复共线性解释变量数据的有偏估计方法，主成分回归法进行参数估计，并对建立的两个模型进行检验与比较。

3.3.1. 最小二乘回归线性模型

最小二乘估计(Least Square Estimation)在线性模型参数估计理论与方法中占有中心的基础地位，在线性模型无偏估计类中占据最优的地位。

对于线性模型

$$y = X\beta + e$$

其中 y 是 $n \times 1$ 观测向量， X 为 $n \times p$ 的设计矩阵， β 为 $p \times 1$ 未知参数向量，在使误差向量 $e = y - X\beta$ 达到最小的思想下来求得 β 的估计

$$\hat{\beta} = (X'X)^{-1} X'y$$

即为最小二乘估计(LS 估计)。

● 模型建立

运用 SPSS 软件对数据进行最小二乘回归计算。首先需要对回归方程的显著性进行检验，即验证原假设 $H_0: \beta_1 = \beta_2 = \dots = \beta_p = 0$ ，如果这个假设被拒绝，认为至少存在一个 $\beta_i \neq 0$ ，也就是响应变量存在一个可以显著依赖的解释变量。计算结果见表 5，可以看出回归方差检验结果的 $P = 0.00 < 0.05$ ，拒绝原假设，认为模型中至少存在一个可以解释响应变量的解释变量。

Table 5. Regression equation test
表 5. 回归方程检验

模型		平方和	自由度	均方	F	显著性
1	回归	1.319	8	0.165	48.935	0.000 ^b
	残差	0.047	14	0.003		
	总计	1.366	22			

接下来对模型系数进行检验，结果如表 6，可以看出在 0.05 的置信水平下， x_1 、 x_3 、 x_5 、 x_6 、 x_7 、 x_8 没有通过显著性检验，没有拒绝原假设 $H_0: \beta_i = 0$ ，认为这几个解释变量对响应变量没有显著性的解释能力。

Table 6. Regression coefficient test
表 6. 回归系数检验

模型	系数B	标准错误	t	显著性
(常量)	-0.066	0.103	-0.641	0.532
居民消费价格总指数 x_1	0.833	0.443	1.878	0.081
居民人均消费 x_2	1.253	0.150	8.350	0.000
商品零售价格总指数 x_3	-0.277	0.160	-1.736	0.105
农业生产资料价格总指数 x_4	-1.759	0.645	-2.726	0.016
农业总产值 x_5	0.209	0.288	0.726	0.480
粮食单产 x_6	0.193	0.292	0.659	0.520
农作物播种面积 x_7	0.186	0.148	1.256	0.230
农业机械年末总动力 x_8	0.446	0.213	2.095	0.055

把这几个解释变量去掉之后重新进行最小二乘回归，首先在回归方程的检验中，由表 7，结果显著。

Table 7. Regression equation test
表 7. 回归方程检验

模型		平方和	自由度	均方	F	显著性
1	回归	1.267	2	0.634	127.923	0.000 ^b
	残差	0.099	20	0.005		
	总计	1.366	22			

接下来对系数进行检验，由表 8，两个解释变量的结果显著性水平 $P < 0.05$ ，即这两个解释变量对响应变量对模型都具有显著的解释作用。

Table 8. Regression coefficient test**表 8.** 回归系数检验

模型	系数B	标准错误	t	显著性
(常量)	0.097	0.033	2.899	0.009
居民人均消费 x_2	0.959	0.061	15.781	0.000
农业生产资料价格总指数 x_4	-0.134	0.052	-2.563	0.019

根据最小二乘回归线性模型的结果, 所得线性回归模型为

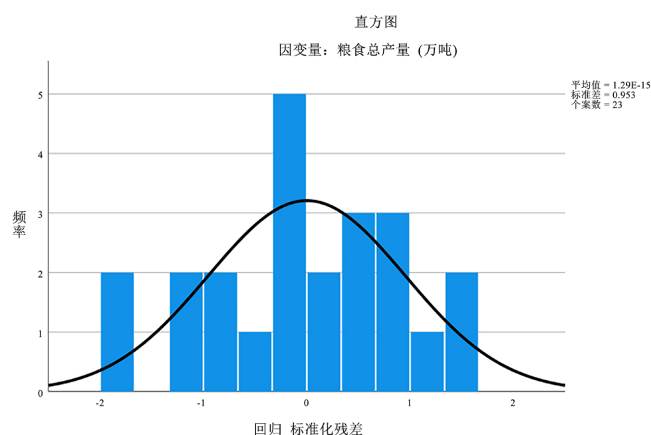
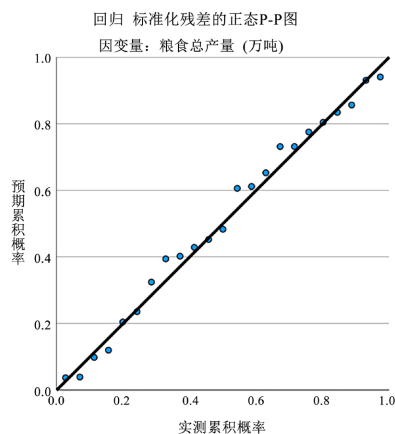
$$\hat{y} = 0.97 + 0.959 \frac{x_2 - \min_{j=2}(x_{ij})}{\max_{j=2}(x_{ij}) - \min_{j=2}(x_{ij})} - 0.134 \frac{x_4 - \min_{j=4}(x_{ij})}{\max_{j=4}(x_{ij}) - \min_{j=4}(x_{ij})}$$

计算得

$$\hat{y} = 1.413 + 0.959x_2 - 0.134x_4$$

● 回归诊断

由线性模型的基本假设, 需要回归方差对残差进行检验。

**Figure 1.** Histogram of standardized residual distribution**图 1.** 标准化残差分布直方图**Figure 2.** Normal P-P Plot**图 2.** 正态 P-P 图

由图 1 可以看到，残差大致呈正态分布的形状，且图 2 中的散点聚集在对角线附近，说明残差满足通过检验，模型成立。

3.3.2. 主成分回归法

主成分回归法是针对变量间复共线性程度严重情况下的经典方法，是主成分分析与线性回归相结合的方法，在有偏估计方法中具有相对广泛的应用[6]。主成分回归分为两个阶段：主成分分析法降维与多元线性模型回归。

主成分分析即在损失少量信息的前提下，将解释变量的信息通过正交旋转变换的方式，提取出 m 个互不相关的主成分 $z_1, z_2, \dots, z_m$ ，使每个主成分都是 p 个原解释变量 $x_1, \dots, x_p$ 的线性组合。这样可以在保留大量原始信息的前提下降维，简化问题，提高模型效率。

● 模型建立

Table 9. KMO and Bartlett’s test
表 9. KMO 和巴特利特检验

KMO取样适切性量数		0.534
巴特利特球形度检验	近似卡方	268.905
	自由度	28
	显著性	0.000

根据图 3 可以看出，在提取 3 个因子后，特征值下降幅度变缓，因而选择 3 个主成分即可保留足够的信息(表 9)。

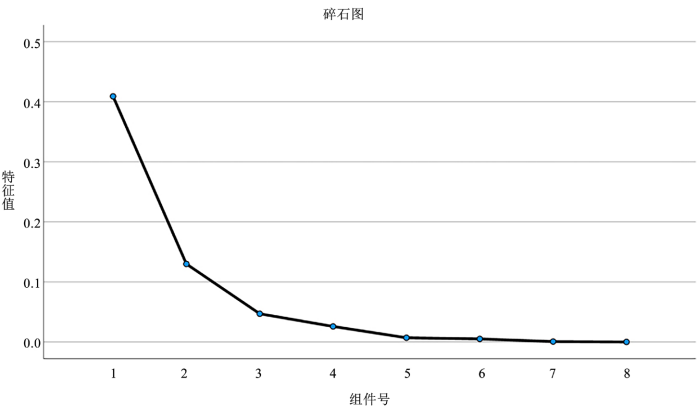


Figure3. Gravel diagram
图 3. 碎石图

故而在确定保留 3 个因子的原则性进行计算，得到表 10，可以看出，在提取 3 个主成分后，方法贡献率已经达到 93.674%，模型保留了足够的信息。

Table 10. Total variance explained
表 10. 总方差解释

成分	初始特征值 ^a			提取载荷平方和		
	总计	方差百分比	累积%	总计	方差百分比	累积%
1	0.409	65.366	65.366	0.409	65.366	65.366

续表

2	0.130	20.773	86.139	0.130	20.773	86.139
3	0.047	7.535	93.674	0.047	7.535	93.674
4	0.026	4.156	97.829			
5	0.007	1.151	98.980			
6	0.005	0.842	99.822			
7	0.001	0.141	99.963			
8	0.000	0.037	100.000			

根据成分得分系数矩阵表 11，可以得到三个主成分分别为

$$z_1 = 0.244x_1 + 0.7x_2 + 0.005x_3 + 0.221x_4 + 0.253x_5 + 0.173x_6 - 0.22x_7 + 0.144x_8$$

$$z_2 = -0.194x_1 + 0.437x_2 + 0.32x_3 - 0.087x_4 - 0.154x_5 + 0.101x_6 - 0.315x_7 + 0.087x_8$$

$$z_3 = 0.487x_1 + 0.351x_2 + 0.411x_3 + 0.008x_4 + 0.163x_5 - 0.068x_6 + 0.487x_7 - 0.824x_8$$

Table 11. Component score coefficient matrix table

表 11. 成分得分系数矩阵表

	成分		
	1	2	3
居民消费价格总指数 x_1	0.244	-0.194	0.487
居民人均消费 x_2	0.070	0.437	0.351
商品零售价格总指数 x_3	0.005	0.320	0.411
农业生产资料价格总指数 x_4	0.221	-0.087	0.008
农业总产值 x_5	0.253	-0.154	0.163
粮食单产 x_6	0.173	0.101	-0.068
农作物播种面积 x_7	-0.022	-0.315	0.487
农业机械年末总动力 x_8	0.144	0.087	-0.824

以主成分分析所得三个因子为解释变量，以粮食总产量 y 为响应变量进行最小二乘回归分析，建立多元回归模型。

由表 12 可以看出，回归方差显著性通过检验，即至少存在一个解释变量能对响应变量进行显著解释。

Table 12. Regression equation test

表 12. 回归模型检验

模型	平方和	自由度	均方	F	显著性
1 回归	1.128	3	0.376	29.890	0.000 ^b
残差	0.239	19	0.013		
总计	1.366	22			

对系数进行检验，由表 13 可以看出三个解释变量都通过了系数检验。

Table 13. Regression coefficient test

表 13. 回归系数检验

模型	B	标准错误	t	显著性
(常量)	0.470	0.023	20.083	0.000

续表

z_1	0.068	0.024	2.865	0.010
z_2	0.204	0.024	8.550	0.000
z_3	0.069	0.024	2.891	0.009

因此, 此多元性线性回归模型公式为

$$y = 0.47 + 0.68z_1 + 0.204z_2 + 0.069z_3$$

从而去标准化后主成分回归模型公式为

$$\hat{y} = 0.47 + 0.16x_1 + 0.16x_2 + 0.09x_3 + 0.13x_4 + 0.15x_5 + 0.13x_6 - 0.04x_7 + 0.05x_8$$

3.4. Lasso 回归方法对参数选取的验证

为进一步检验变量选择的稳健性, 本文引入 Lasso (Least Absolute Shrinkage and Selection Operator) 回归方法。Lasso 通过对回归系数施加 L1 惩罚项, 实现变量选择与系数收缩, 有效处理多重共线性问题(张兆鹏, 2024) [7]。其目标函数为:

$$\min_{\beta} \left\{ \frac{1}{2n} \sum_{i=1}^n \left(y_i - \beta_0 - \sum_{j=1}^p \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^p |\beta_j| \right\}$$

其中, λ 为正则化参数, 控制惩罚项的强度。当 λ 足够大时, 部分系数会被压缩至零, 从而实现变量选择。这一特性使 Lasso 特别适用于处理高维数据和多重共线性问题。

Table 14. Model testing

表 14. 模型检验

	平方和	df	均方	F	p 值
回归	1.182	8	0.148	11.210	0.000
残差	0.185	14	0.013		
总计	1.366	22			

λ 值在大于 0.01 后, 自变量的标准化回归系数趋于稳定确定, 故而 λ 值确定为 0.01。根据表 14 可以看到, 此时确定的模型通过 F 检验, 即认为模型中至少存在一个可以解释响应变量的解释变量。

Table 15. LASSO regression coefficient test

表 15. LASSO 回归系数检验

	非标准化系数			t	p
	B	标准误	Beta		
常数	0.144	0.139	-	1.036	0.318
居民消费价格总指数	0	0.216	-0.045	0	1
居民人均消费	0.721	0.214	1.055	3.367	0.005 ^b
商品零售价格总指数	0	0.229	-0.031	0	1
农业生产资料价格总指数	0	0.274	-0.522	0	1
农业总产值	0	0.299	0.224	0	1
粮食单产(千克/公顷)	0	0.325	0.153	0	1
农作物播种面积(公顷)	0	0.182	0.144	0	1
农业机械年末总动力(万千瓦)	0	0.153	0.058	0	1

由表 15 可以看到，LASSO 回归的系数检验结果中仅有解释变量居民人均消费是对响应变量粮食总产量有着显著影响的，结果与最小二乘法相近，可以说明最小二乘选择变量的结果是稳定的。但 LASSO 回归模型计算结果中显著的变量过少，考虑到模型的稳健性，此方法不予考虑。

3.5. 回归方法对比

为衡量回归模型对数据的拟合程度，以均方误差为标准对两种方法的回归效果进行评价，其中均方误差计算为

$$MSE = \frac{1}{n} \sum_{j=1}^n (y_j - \hat{y}_j)^2$$

计算得两种方法的效果见表 16，可以明显看出主成分回归拟合效果优于最小二乘回归，因而对于复共线性程度较高的数据应优先考虑主成分回归

Table 16. Effect comparison

表 16. 效果对比

MSE	主成分回归	最小二乘回归
	8.44	26.32

4. 分析与结论

通过对三种模型的效果对比可知，主成分回归模型

$$\hat{y} = 0.47 + 0.16x_1 + 0.16x_2 + 0.09x_3 + 0.13x_4 + 0.15x_5 + 0.13x_6 - 0.04x_7 + 0.05x_8$$

对原始数据的解释程度相对较高，模型更加精确。

结合自变量的含义，可以得到粮食生产总量与居民消费价格指数、居民人均消费、上铺零售价格总指数、农业生产资料价格总指数、农业总产值、粮食单产、农业器械总动力呈正相关关系，这与常识相符。与农作物播种面积呈负相关关系，这似乎与常识不符，但随着技术的升级，即农业器械总动力呈升级趋势，山东省农作物播种面积并不需要出现上升趋势也可以实现粮食生产总量是的上升，故而在数据量较少的情况下计算得到负相关关系也是合理的。

参考文献

- [1] Solow, R.M. (1957) Technical Change and the Aggregate Production Function. *The Review of Economics and Statistics*, 39, 312-320. <https://doi.org/10.2307/1926047>
- [2] Hayami, Y. and Ruttan, V.W. (1985) *Agricultural Development: An International Perspective*. 2nd Edition, Johns Hopkins University Press.
- [3] 王学真, 石雪莲, 高峰. 山东省粮食生产波动影响因素分析[J]. 山东理工大学学报(社会科学版), 2006(6): 18-22.
- [4] 孙宏业, 柳宜可. 内蒙古农产品价格影响因素实证分析[J]. 内蒙古财经大学学报, 2020, 18(5): 139-142.
- [5] 唐军. 山东省玉米价格波动影响因素研究及政策优化[D]: [硕士学位论文]. 贵阳: 贵州大学, 2023.
- [6] 王松桂. 线性模型引论[M]. 北京: 科学出版社, 2004.
- [7] 张兆鹏. 商业银行理财产品实际年化收益的影响研究——基于 Lasso 和主成分回归的实证分析[J]. 中国证券期货, 2024(3): 54-63.