Computing, Performance and Communication Systems (2023)
Clausius Scientific Press, Canada

DOI: 10.23977/cpcs.2023.070106
ISSN 2371-8889 Vol. 7 Num. 1

Multiple Regression Analysis of Factors Influencing Grain Yield

Jingjun Li

Social Statistics & Demography, University of Southampton, Southampton, SO17 1BJ, UK

Keywords: Annual grain yield; Regression analysis; Stepwise regression; Ridge regression

Abstract: Since ancient times, China has been a major agricultural country and is now the world's top food producer. With the second largest population in the world, it is of great relevance to investigate the factors influencing grain production. In this paper, we look at the main factors that affect grain yield. A simple multiple linear regression analysis was used to develop a model with fertiliser application, sown area, flooded area, farm machinery power and agricultural labour as independent variables and total annual grain production as the dependent variable. The resulting model fitted well and the observations were independent of each other. However, there was serious covariance between the variables, so we tested the model and concluded that the model satisfied the chi-squaredness, but there was more serious covariance between the variables, which affected the model building and could cause model distortion. So finally we build stepwise regression and ridge regression models respectively to eliminate the multicollinearity among the variables in order to optimise the model.

1. Design questions

Table 1: Summary of data relating to grain production

Year	Grain production (million tonnes) y	Fertilizer application (million kg) x_1	Sown area (thousand hectares) x_2	Damaged area (Thousands of hectares) x_3	Agricultural machinery power (million kilowatts) x_4	Agricultural labour force (ten thousand people) x_5
2000	38728	1659.8	114047	16209.3	18022	31645.1
2001	40731	1739.8	112884	15264	19497	31685
2002	37911	1775.8	108845	22705.3	20913	30351.5
2003	39151	1930.6	110933	23656	22950	30467
2004	40298	1999.3	111268	20393.7	24836	30870
2005	39408	2141.5	110123	23944.7	26575	31455.7
2006	40755	2357.1	112205	24448.7	28067	32440.5
2007	44624	2590.3	113466	17819.3	28708	33330.4
2008	43529	2805.1	112314	27814	29389	34186.3
2009	44266	2930.2	110509	25894.7	30308	34037
2010	45649	3151.9	110509	23133	31817	33258.2
2011	44510	3317.9	109544	31383	33802	32690.3
2012	46662	3593.7	110060	22267	36118	32334.5
2013	50454	3827.9	112548	21234	38546	32260.4

Data source: China Statistical Yearbook

China is a large agricultural country and has the second largest population in the world, so it is of great practical importance to investigate the factors influencing grain yield [1]. Grain production is used as the dependent variable, and fertiliser application, sown area, disaster area, farm machinery power and agricultural labour are used as independent variables. A multiple linear regression analysis is conducted on the data related to grain production from 2000 to 2013 in China to explore the influencing factors of grain yield, and some of the data are show as table 1.

2. Model Assumptions and Establishment

2.1 Regression analysis

Regression analysis is a mathematical and statistical method that deals with the statistical correlation of variables [2]. The basic idea of regression analysis is that although there is no strict, deterministic functional relationship between the independent and dependent variables. It is possible to find a mathematical expression that best represents the relationship between them. Regression analysis can be classified according to the number of dependent variables and independent variables, and can be divided into linear regression analysis and non-linear regression analysis according to the functional expressions of the dependent and independent variables. The linear regression model of random y and $x_1,x_2,...x_k$ the variables is as follows Equation.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon \tag{1}$$

Where $\beta_0, \beta_1, \ldots, \beta_k$ are $k+1$ unknown parameters, β_0, which is called regression constant, and $\beta_1, \ldots, \beta_k$ are called regression coefficient.

The following basic assumptions are satisfied:

(1) Explanatory variable $x_1, x_2, \ldots x_k$ are non-random variables, and the observed value $x_{i1},x_{i2},...x_{ik}$ are constants.

(2) Equivariances and irrelevant assumptions (Equation 2).

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

2.2 Correlation analysis

Correlation analysis is a statistical method to study the correlation between random variables by studying whether there is a certain dependency relationship between phenomena and exploring the correlation direction and degree of the specific dependency phenomenon. Correlation is a kind of non-deterministic relation, which is described by correlation coefficient r.

$|r| > 0.95$ There was a significant correlation.

$|r| \geq 0.8$ High correlation.

$0.5 \leq |r| < 0.8$ Moderate Correlation.

$0.3 \leq |r| < 0.5$ Low correlation.

$|r| < 0.3$ The relationship is so weak that it is considered irrelevant.

Spearman rank correlation analysis was used in this paper, which is suitable for small sample size. The calculation formula is $p = 1 - \frac{6\sum d_i^2}{n^3 - n}$.

2.3 Stepwise regression

When there are many independent variables, some of these factors may not have a significant

effect on the corresponding variable and the x's may not be completely independent of each other and may have various interactions. In such cases, stepwise regression analysis can be used to screen the x-factors[3].

The multiple regression model thus established will work better. Stepwise regression analysis begins by establishing the total regression equation between the dependent variable y and the independent variable x. The total equation and each independent variable are then hypothesis tested. When the total equation is not significant, it indicates that the linear relationship of that multiple regression equation does not hold. And when one of the independent variables does not have a significant effect on y, it should be removed and rebuilt[4]. The multiple regression equation that does not contain the factor is screened out to identify the factor that has a significant effect as the independent variable, and the optimal regression equation is created. The more independent variables the regression equation contains, the larger the regression sum of squares, the smaller the residual sum of squares and consequently the smaller the residual square. The error in the predicted values is also smaller and the better the model fit [5].

2.4 Ridge regression

Ridge regression is a biased estimation regression method dedicated to the analysis of collinear data. It is actually an improved least squares estimation method. By giving up the unbias of the least squares method, the regression coefficient is more in line with the reality and more reliable regression method at the cost of losing part of the information and reducing the accuracy.

Adding a penalty term to the objective function of a linear regression model.

$$J(\beta) = \sum(y - X\beta)^2 + \sum \lambda\beta^2 \tag{3}$$

To solve for the minimum of the objective function, you need to find the derivative of it and make the derivative function 0.

$$\frac{\partial J(\beta)}{\partial \beta} = 2(X'X + \lambda I)\beta - 2X'y = 0 \tag{4}$$

The problem of minimizing the objective function value J(B) of the ridge regression model is equivalent to

$$\arg min \sum(y - X\beta)^2, \quad \sum \beta^2 \leq t \tag{5}$$

3. Analysing the data

3.1 Plotting matrix scatter plots

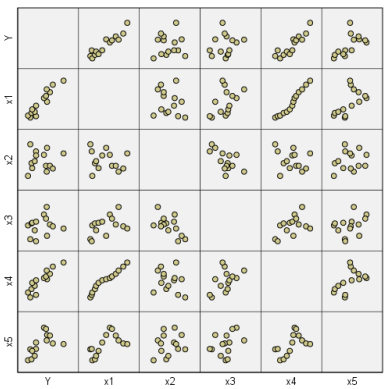

Figure 1: Matrix scatter plot

Analysis: Figure 1 is a matrix scatter plot of the dependent variable and all independent variables. The first row of the graph shows that there is a more obvious linear relationship between fertiliser application, farm machinery power and the dependent variable, while the remaining variables do not have an obvious linear relationship with the dependent variable, and their effects are not significant, as shown in the following specific analysis.

3.2 Linear regression analysis

We first set up a multiple linear regression equation between the respective variables and the dependent variable to initially analyse this problem, and the following are the outputs (Table 2).

Table 2: Descriptive statistics

	N	Minimal Value	Maximum Value	Mean Value	Standard Deviation	Variance
x1	14	1659.8	3827.9	2558.636	724.5038	524905.802
x2	14	108845.0	114047.0	111375.357	1556.4360	2422493.016
x3	14	15264.0	31383.0	22583.336	4351.5860	18936300.604
x4	14	18022.0	38546.0	27824.857	6153.9000	37870484.901
x5	14	30351.5	34186.3	32215.136	1220.0434	1488505.872
Number of valid cases (column)	14					

The table above briefly describes the maximum and minimum values, mean values, variances and standard deviations of each variable.

Table 3: Model summary

Model	R	R^2	Adjusted R^2	Errors in standard estimates	Durbin-Watson
1	0.987[a]	0.975	0.959	731.8312	2.132

The fitting effect of this model can be obtained from the table 3. The adjusted R^2 of the fourth column is 0.959. It can be seen that the five variables in this model can explain 95.9% of the model changes, so the fitting effect of this model is good. The Durbin-Watson test value in the table is 2.132. Generally, the test values are distributed between 0 and 4, and the closer they are to 2, the more likely they are to be independent of each other. That is, the observed values of simple linear regression in this study are mutually independent.

Table 4: Variance analysis table

Model		Quadratic Sum	Degree of freedom	Mean square	F	Significance
1	Regression	165834737.540	5	33166947.508	61.928	0.000[b]
	Residual error	4284615.317	8	535576.915		
	Aggregate	170119352.857	13			

As can be seen from Table 4, the significance of F test (P value) =0.000<0.01. Therefore, it can be considered that the linear relationship established by each variable and dependent variable has significant statistical significance at the significance level of 0.01.

It can be seen from the significance column of the above table 5 that only the significance of variable x_1, x_3 are less than 0.05, indicating that there is a significant linear relationship between these two variables and y. From the VIF value, a value of x_1, x_4 greater than 10 indicates that there should be multicollinearity. The regression model can be obtained from the table as follows.

$$y = 5.137x_1 + 0.226x_2 - 0.2x_3 + 0.008x_4 + 0.147x_5 + 3923.229$$

In conclusion, the model has good fitting effect and statistical significance. However, not all variables are significant and there may be multicollinearity between variables, so the model needs to be tested.

Table 5: Regression coefficient table

Model		Unstandardized coefficient		Standardization coefficient	t	significance	Collinear statistics	
		B	Standard error	Beta			tolerance	VIF
1	constant quantity	3923.229	20607.655		0.190	0.854		
	x1	5.137	1.500	1.029	3.425	0.009	0.035	28.658
	x2	0.226	0.209	0.097	1.077	0.313	0.388	2.579
	x3	-0.200	0.079	-0.241	-2.541	0.035	0.351	2.852
	x4	0.008	0.172	0.014	0.047	0.964	0.037	27.168
	x5	0.147	0.270	0.050	0.543	0.602	0.379	2.640
a. dependent variable:Y								

4. Model testing

4.1 Cointegration test

Table 6: Collinear diagnosis table

Model	Dimension	Eigenvalue	Conditional index	Variance Proportion					
				(constant)	x1	x2	x3	x4	x5
1	1	5.911	1.000	0.00	0.00	0.00	0.00	0.00	0.00
	2	0.066	9.485	0.00	0.02	0.00	0.00	0.00	0.00
	3	0.022	16.393	0.00	0.00	0.00	0.44	0.00	0.00
	4	0.001	68.940	0.00	0.58	0.00	0.02	0.72	0.05
	5	0.000	118.712	0.05	0.39	0.01	0.04	0.27	0.69
	6	3.971E-5	385.802	0.95	0.02	0.99	0.50	0.01	0.26
a. Dependent variable:Y									

The condition index can be used to judge whether multicollinearity exists and the severity of multicollinearity. Eg. It is generally thought that 0<k<10, there is no multicollinearity. When 10≤k＜100, there is strong multicollinearity. Severe multicollinearity exists when k≥100. It can be seen from table 6, $k_5 = 118.712$, $k_6 = 385.802$, that serious multicollinearity exists in this problem. It can be roughly concluded from the variance ratio $x_2 = 0.99$that there should be multicollinearity between this variable and other variables.

4.2 Test autocorrelation

As can be seen from Table 7, usually the correlation coefficient between variables is below 0.5. As can be seen from the above table, x_1 is highly correlated with x_4,x_5, and x_2 is highly correlated with x_3,x_4. Therefore, multicollinearity exists among these variables, which will cause model distortion.

Table 7: Spearman correlation test

			x1	x2	x3	x4	x5
Spearman Rho	x1	Correlation index	1.000	-0.286	0.420	1.000**	0.604*
		Sig.(two-tailed)	0.00	0.322	0.135	0.00	0.022
		N	14	14	14	14	14
	x2	Correlation index	-0.286	1.000	-0.557*	-0.286	0.174
		Sig.(two-tailed)	0.322	0.00	0.039	0.322	0.552
		N	14	14	14	14	14
	x3	Correlation index	0.420	-0.557*	1.000	0.420	0.393
		Sig.(two-tailed)	0.135	0.039	0.00	0.135	0.164
		N	14	14	14	14	14
	x4	Correlation index	1.000**	-0.286	0.420	1.000	0.604*
		Sig.(two-tailed)	0.00	0.322	0.135	0.00	0.022
		N	14	14	14	14	14
	x5	Correlation index	0.604*	0.174	0.393	0.604*	1.000
		Sig.(two-tailed)	0.022	0.552	0.164	0.022	0.00
		N	14	14	14	14	14
**. At Level 0.01 (two-tailed), Significant correlation.							
*. At Level 0.05 (two-tailed), Significant correlation.							

4.3 Heteroscedasticity test

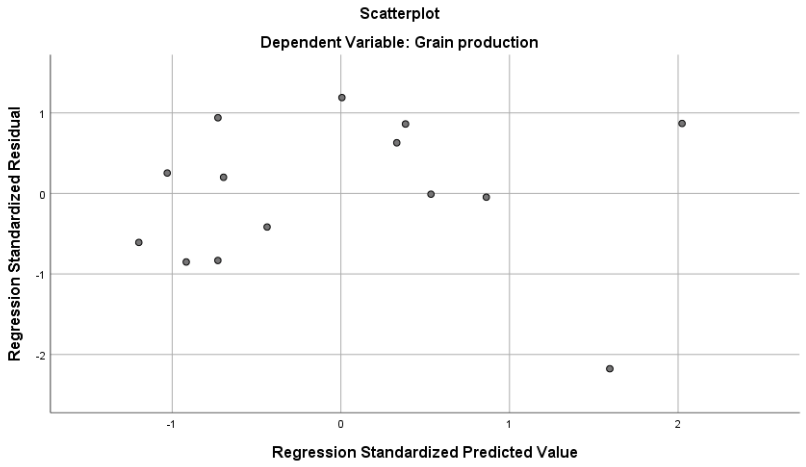

Figure 2: Residual graph

Table 8: Residual statistics table

	Minimal Value	Maximum Value	Mean Value	Standard Deviation	number of cases
Predicted value	37673.660	49767.949	42619.714	3571.6268	14
residual error	-1126.1721	686.0491	0.0000	574.0956	14
Normal expected value	-1.385	2.001	0.000	1.000	14
Standardized residual	-1.539	0.937	0.000	0.784	14
a. Dependent variable: Y					

The Figure 2 is the residual graph, from which it can be seen that the standardized residual randomly distributed around the 0 horizontal line, without an obvious trend, and its fluctuation range basically remained stable, and did not change with the change of the standardized predicted value, so

it can be considered that the problem has homogeneity of variance. And table 8 shows the maximum and minimum residuals and other statistics.

It can be concluded from the above model test that the model meets the homogeneity of variance, but there is a serious collinearity among variables, which affects the establishment of the model and causes the distortion of the model. So next we use some processing methods to eliminate multicollinearity between variables to optimize the model.

5. Elimination of covariance

5.1 Stepwise regression

Table 9: Variables entered/removed

Model	Input variable	Removed variable	Method
1	x1	0.00	Step (condition: probability of F to be input <=.050, probability of F to be removed >=.100)
2	x3	0.00	Step (condition: probability of F to be input <=.050, probability of F to be removed >=.100)
a. Dependent variable: Y			

As can be seen from Table 9, the final variables selected by stepwise regression method are fertilizer application amount x_1 and disaster area x_3.

Table 10: Model summary

Model	R	R^2	Adjusted R^2	Errors in standard estimates	Durbin-Watson
1	0.945[a]	0.892	0.883	1236.13747	
2	0.982[b]	0.964	0.958	745.31311	2.189
a. Prediction variable: (constant), x1					
b. Prediction variable: (constant), x1, x3					
c. Dependent variable:Y					

As can be seen from Table 10, the third column in the table is the square of the correlation coefficient, which measures the overall fit degree of the regression equation and expresses the overall correlation degree between the dependent variable and the independent variable. It shows that Model 1 and Model 2 can explain 88.3% and 95.8% of the total profit of high-tech industry respectively. Because the second model has a better fitting degree, Model 2 is chosen to establish the regression equation. In the fifth column, the Durbin-Watson statistic is 2.189, close to 2, indicating that there is no obvious correlation between residuals and independence is satisfied.

Table 11: ANOVA analysis of variance

Model		Quadratic Sum	Degree of freedom	Mean square	F	Significance
1	Regression	151782922.852	1	151782922.852	99.332	0.000[b]
	Residual error	18336430.005	12	1528035.834		
	Aggregate	170119352.857	13			
2	Regression	164008944.853	2	82004472.427	147.625	0.000[c]
	Residual error	6110408.004	11	555491.637		
	Aggregate	170119352.857	13			
a. Dependent Variable:Y						
b. Prediction variable: (constant), x1						
c. Prediction variable: (constant), x1, x3						

Table 11 is the variance analysis table, and the F test of variance analysis is the significance test of regression equation. The original assumption is that all independent variables have no influence

on dependent variables. Based on this assumption, it is observed that the F statistic of Model 2 is 147.625, and the corresponding significance level is 0.000 < 0.05, that is, whether all variables in the model are statistically significant. Further testing of the respective variables is needed.

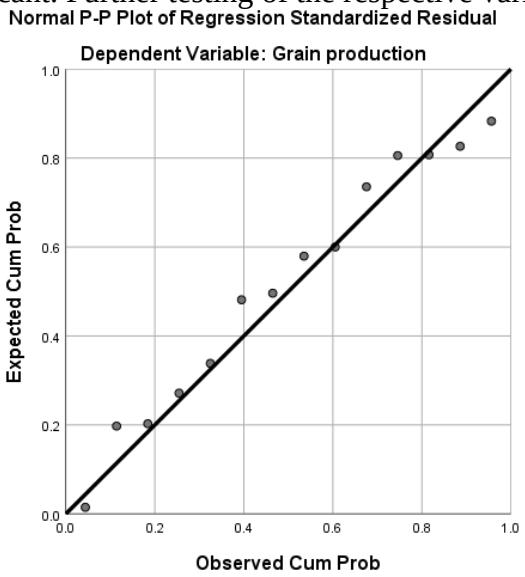

Figure 3: P-P diagram

From Figure 3, we can see that the standard residuals are all around a straight line, indicating that the residuals basically obey the normal distribution are evenly distributed and the fit is good.

Table 12: Regression coefficient significance test table

Model		Unstandardized coefficient		Standardization coefficient	t	Significance	Collinear statistics	
		B	Standard error	Beta			tolerance	VIF
1	Constant quantity	30552.485	1255.037		24.344	0.000		
	x1	4.716	0.473	00.945	9.967	0.000	1.000	1.000
2	Constant quantity	34457.363	1124.903		30.631	0.000		
	x1	5.397	0.320	1.081	16.862	0.000	0.795	1.258
	x3	-0.250	0.053	-0.301	-4.691	0.001	0.795	1.258
a. Dependent variable:Y								

Table 12 shows regression coefficient and significance test, while t test is significance test for a single independent variable. In Model 2, the significance values of the independent variables of fertilizer application amount and disaster area are all lower than the significant level of 0.05, so the coefficient of each variable is very significant and has statistical significance, which cannot be removed from the regression equation. The stepwise regression equation is as follows.

$$y = 3953.463 + 0.226x_2 + 0.021x_5$$

The value of the last column is VIF, and it can be seen that the VIF values corresponding to x_1 and x_3 are all less than 10, so there is no multicollinearity, indicating that this method has well solved the collinearity problem between independent variables.

5.2 Ridge Regression

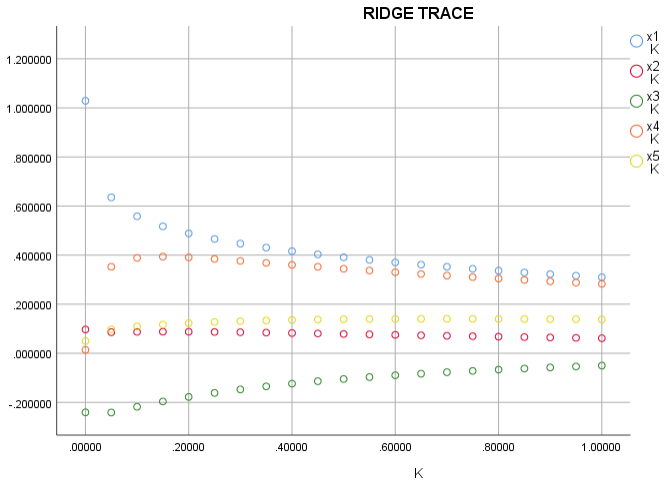

Figure 4: Ridge trace map

As shown in Figure 4, the abscissa is the value of ridge parameter k, and the ordinate is the coefficient of the independent variable. The change of coefficient is relatively stable, but its absolute value is always small. It can be removed from the independent variable according to the principle of ridge regression variable selection. And because there is a strong collinearity between and, so the two variables can be retained one. Compared with, the coefficient is small, so the variable is eliminated. Ridge regression analysis was performed again.

Table 13: Table of Ridge parameters after elimination of variables

K	RSQ	x1	x3	x5
0	0.97106	1.022065	-0.30969	0.10484
0.05	0.96681	0.93903	-0.26851	0.133923
0.1	0.95693	0.871331	-0.2347	0.154281
0.15	0.94415	0.814823	-0.20649	0.168689
0.2	0.92992	0.766758	-0.18263	0.178912
…	…	…	…	…
0.7	0.78987	0.503396	-0.06194	0.195022
0.75	0.77784	0.487944	-0.05589	0.193567
0.8	0.76618	0.473549	-0.0504	0.191948
0.85	0.75488	0.460096	-0.04543	0.190205
0.9	0.74393	0.447488	-0.0409	0.188366
0.95	0.73329	0.43564	-0.03676	0.186458
1	0.72297	0.42448	-0.03298	0.184499

The first column in Table 13 is ridge parameter k. The software default k value ranges from 0-1 and is 0.05. There are 21 k values in total. The second column is the decision coefficient R^2, and the third to fifth column is the standardized ridge regression coefficient, in which the value corresponding to the first row k=0 is the standardized regression coefficient estimated by the ordinary least squares.

It can be seen from Figure 5 that when the ridge parameter k=0.2, the ridge trace map has been basically stable. Let's look at the complex coefficient of determination. When k=0.2, it is still large, so the ridge parameter k=0.2 can be selected, and then the ridge regression can be re-performed with the given k=0.2.

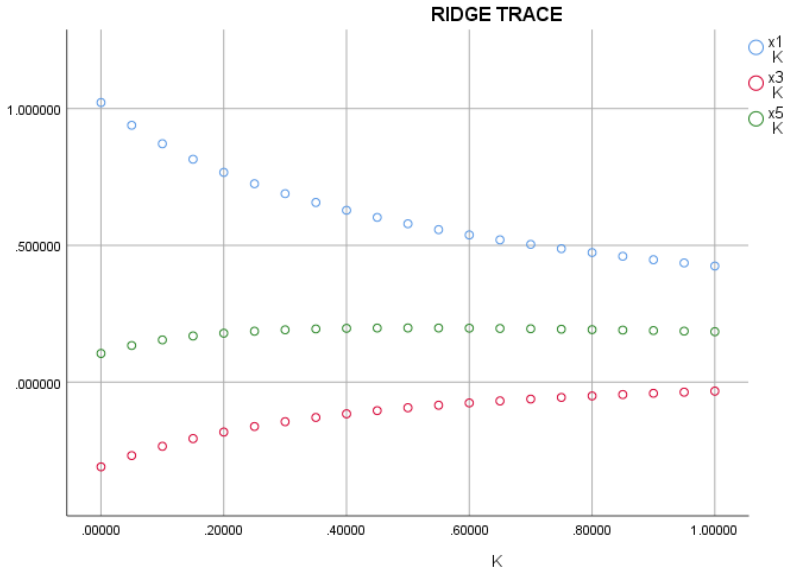

Figure 5: Ridge trace map after removing variables

****** Ridge Regression with k = 0.2 ******

Mult R .995197
RSquare .990417
Adj RSqu .984668
SE 1746.650653

ANOVA table

	df	SS	MS
Regress	3.000	1.58E+009	525516156
Residual	5.000	15253943	3050788.5

F value	Sig F
172.2558462	.0000183

----------Variables in the Equation----------

	B	SE(B)	Beta	B/SE(B)
x1	4.4615809	.4146000	.2707270	10.7611706
x3	.5767566	.0271849	.3060572	21.2160662
x5	.4769212	.0379747	.3615795	12.5589318
Constant	3.2562242	625.8435756	.0000000	.0052029

Figure 6: Ridge regression analysis results

As can be seen from Figure 6, the adjusted R square of the model is 0.9847, indicating that several variables of the model explain 98.47% of the total population. The standardized Ridge regression equation of y pairs, which proves that the model has a good fit, is shown below.

$$\hat{y} = 0.270727x_1 + 0.3060572x_3 + 0.3615795x_5$$

The unstandardized ridge regression equation is shown below.

$$\hat{y} = 3.2562242 + 4.4615809x_1 + 0.5767566x_3 + 0.4769212x_5$$

5.3 Model Comparison

Table 14: Comparison table of regression models

	Stepwise regression model	Ridge regression model
The fitted regression equation p value	<0.001	<0.001
The number of arguments it contains	2	3
Adjusted R^2	0.958	0.985

Both regression models were good at addressing the high level of multicollinearity between the independent variables in the original model. Both fitted regression equations have p-values < 0.001 and are statistically significant overall (table 14). The number of independent variables in the stepwise regression model is two and the number of independent variables in the ridge regression model is three, and the stepwise regression model is better than the ridge regression model based on the principle that fewer independent variables is a better model. The adjusted R^2 for the stepwise regression model is 0.958 and the adjusted R^2 for the ridge regression model is 0.985. The ridge regression model is larger than the principal component regression model, and the adjusted R^2 reflects the goodness of fit of the model.

In summary, if this is intended to explain the information of the original variables through fewer variables, a stepwise regression model can be used, with the regression equation as follows.

$$\hat{y} = 3.2562242 + 4.4615809x_1 + 0.5767566x_3 + 0.4769212x_5$$

6. Conclusion

In this paper, in order to study the factors affecting grain production in China, we took annual grain production as the dependent variable y, and fertilizer application, sown area, disaster area, agricultural machinery power and agricultural labour as the main factors affecting grain production. Using SPSS, a multiple regression model was established and the model fit was good, with the observations being independent of each other. However, there was a serious problem of covariance among the variables, so we tested the model for heteroskedasticity, autocorrelation and covariance, and concluded that the model satisfied variance chi-square, but there was more serious covariance among the variables, which affected the establishment of the model and could cause model distortion. So we built stepwise regression and ridge regression models respectively to eliminate the effect of multicollinearity between variables on the model. It was concluded that if one wanted to explain the information of the original variables through fewer variables, one could use a stepwise regression model in which the fertiliser application and the area affected by the disaster were the main factors affecting grain yield. If one wants to use more information about the variables to obtain a better-fitting model, one can choose a ridge regression model, in which fertiliser application, disaster area and agricultural labour are the main factors influencing yield.

References

[1] Cameron A C, Trivedi P K. Regression Analysis of Count Data: Contents. 1998.
[2] Rawlings J O, Dickey D A, Pantula S G, et al. Applied regression analysis: a research tool. Wiley, 1998.
[3] Wang S J, Li J Y, et al. Analysis of the Main Factors Influencing Food Production in China Based on Time Series Trend Chart [J]. Agricultural Research in Asia: English edition, 2014(6):P. 37-42.
[4] Zheng D, An Z, Yan C, et al. Spatial-temporal characteristics and influencing factors of food production efficiency based on WEF nexus in China [J]. Journal of Cleaner Production, 2022, 330:129921.
[5] Zhang M X, Zhao H Y. Analysis on Factors Influencing the Food Production in Tonghua City [J]. Journal of Jilin Agricultural Sciences, 2014.