Linear Regression Analysis of Gross Output Value of Farming, Forestry, Animal Husbandry and Fishery Industries

Liu Haime, Chen Yun
School of Management, Wuhan University of Technology, Wuhan, P.R.China, 430070
(Email: liuhaimei319@gmail.com, chenyun315@126.com)

Abstract: This paper applies software SPSS to analyze 1999-2009 relevant statistics of China agriculture industry promulgated by National Bureau of Statistics of China, trying to establish a Dualistic Linearity Regression Model to calculate the effect of the total agricultural machinery dynamic and total corps planting area in the total output of farming, forestry, animal husbandry and fishery industries, but the results indicate that the independence "total corps planting area" doesn't pass the significance test. So the paper chooses stepwise method to do the regression again, and the software eliminates independence "total corps planting area", and works out a one-dimensional linear regression model. The article aims at finding out the factors that significantly influence the total output of farming, forestry, animal husbandry and fishery industries, and what degree they impact, hence, to put forward advices for practices of farming, forestry, animal husbandry and fishery industries.
Key words: SPSS; Linear regression analysis; Total output of farming; Forestry; Animal husbandry and fishery industries

1 Introduction

Our motherland is an agricultural country, even though these years primary industry weights less in GDP than before, but it's still the main part of GDP, as table 1 shows[1]. Moreover, the products of farming; forestry; animal husbandry and fishery are necessary for people's life. Therefore, research on China farming; forestry; animal husbandry and fishery industries output's influencing factors and the degree they impacts can give suggestions for the improving of its total output, and also is significant in practices.

Table 1 The Proportion of Primary Industry in GDP in 1999-2009

Time	The proportion of primary industry in GDP（%）
1999	16.5
2000	15.1
2001	14.4
2002	13.7
2003	12.8
2004	13.4
2005	12.1
2006	11.1
2007	10.8
2008	10.7
2009	10.3

Data source: 2010 China statistical yearbook

It is easy to get a regression model between dependent variable and independent variable under the help of statistic analysis software. So, this paper try to take SPSS as tool, on the base of 1999-2009 agriculture industry and correlated industry data provided by State Statistic Bureau, to establish regression model of national farming; forestry; animal husbandry and fishery industries output and tests it.

2 Date Collection

Farming; forestry; animal husbandry and fishery output is national macroeconomic indicator, so the data this paper uses is mainly collected from State Statistic Bureau's official website. In the 2010 China statistical yearbook, it provided main results of farming; forestry; animal husbandry and fishery production status, such as farming; forestry; animal husbandry and fishery output and indicator, main agriculture machine owned per unit, effective irrigated area, Agricultural chemical fertilizer used, rural hydropower station and electricity used, rural family main productive capital assets owned per 100 families, total corps planting area, main corps planting structure etc. indicator value, and their rang from 1978 to 2009. Concerned the integrity and coherence of data, the paper chooses arming; forestry; animal

husbandry and fishery output as dependent variable; and combined with analysis of arming; forestry; animal husbandry and fishery output in the angle of economic theory, the paper chooses total corps planting area and total agricultural machinery dynamic as independent variable.

There into, farming; forestry; animal husbandry and fishery output refers to the total value of total products of farming; forestry; animal husbandry and fishery and supportive service activities for them which expressed by money, which reflected total farming; forestry; animal husbandry and fishery production scale and achievement[2]. Total agricultural machinery dynamic means area actually planted or transplanted corps, corps includes food, cotton, oil plants, sugar plants, bastfibreplants, tobacco, vegetable and melon, medical material and others nine kinds[2].

Table 2 China farming, Forestry, Animal Husbandry and Fishery Output, Total Agricultural Machinery Dynamic and Total Corps Planting Area in 1999-2009

Time	farming; forestry; animal husbandry and fishery output(billion) Y	total agricultural machinery dynamic（kw） X1	total corps planting area （thousand hectare） X2
1999	24519.1	14106.2	156373
2000	24915.8	13873.6	156300
2001	26179.6	14462.8	155708
2002	27390.8	14931.5	154636
2003	29691.8	14870.1	152415
2004	36239.0	18138.4	153553
2005	39450.9	19613.4	155488
2006	40810.8	21522.3	152149
2007	48893.0	24658.1	153464
2008	58002.2	28044.2	156266
2009	60361.0	30611.1	158639

Data source: 2010 China Statistics Yearbook

We can make theoretical analysis as followed. Total agricultural machinery dynamic and farming; forestry; animal husbandry and fishery output are statistic aggregate indicators, total agricultural machinery dynamic is investment of farming; forestry; animal husbandry and fishery industry. Normally, grow in investment leads grow in output. So, we suppose total agricultural machinery dynamic is positive correlated with farming; forestry, animal husbandry and fishery output to a large extent. Total corps planting area is the main condition of farming; forestry; animal husbandry and fishery production, but corps planting area only comes down to agriculture which is a part of farming; forestry; animal husbandry and fishery industry, so we suppose it limitedly positive correlated with farming; forestry; animal husbandry and fishery output.

Depend on these analysis, and because of timeliness of data, the article chooses data between 1999-2009 of the dependent variable and independent variable talked above, as the base of regression analysis, showed as table 2[3].

3 Linear Regression Analysis

3.1 Dualistic linearity regression analysis

Depending on theoretical analysis, firstly chooses total agricultural machinery dynamic and total corps planting area as independent variable, do Dualistic Linearity Regression analysis of these data, to explain the influence total agricultural machinery dynamic and total corps planting area impact on farming; forestry; animal husbandry and fishery output. As, table 2 shows, defined variable "Y" as farming; forestry; animal husbandry and fishery output, variable "X1" as total agricultural machinery dynamic, and variable "X2" as total corps planting area. Choose "Linear regression analyze" in SPSS software, and choose "enter" while screen independent variable, choose regression, and choose regression coefficient, 95% coefficient interval, PP plot, unstandardized individual prediction value, unstandardized residual and so on, and get the results as table 3.

Table 3 Model Summary(b)

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.995(a)	.991	.988	1412.91489

a Predictors: (Constant), X2, X1 b Dependent Variable: Y

Table 3 indicates, goodness of fit R^2=0.991, adjusted R square is 0.988, both are near to 1, which

indicates that the model fits well. And the standard error of the estimate is 1412.915.

Table 4 Analysis of Variance

ANOVA(b)

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	1689912814.578	2	844956407.289	423.255	.000(a)
	Residual	15970627.929	8	1996328.491		
	Total	1705883442.507	10			

a Predictors: (Constant), X2, X1 b Dependent Variable: Y

The F test of Multiple Linear Regression equations aims to test whether the whole regression equation is linear significant or not, and puts forward null hypothesis and alternative hypothesis as followed: $\boldsymbol{H_0}$: $\beta_1=\beta_2=0$ $\boldsymbol{H_1}$: $\beta_j$not all are 0（j=1,2）

Relying on analysis of variance, we can find that the value of F is 423.255, and the relevant significance probability is 0.000, is less than 0.01. So it's apparent to turn down $\boldsymbol{H_0}$, and makes the conclusion that regression equation is significant in linear relation.

Table 5 Coefficients

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Correlations			Collinearity Statistics	
		B	Std. Error	Beta			Zero-order	Partial	Part	Tolerance	VIF
1	(Constant)	35702.975	36519.902		.978	.357					
	X1	2.198	.078	1.006	28.085	.000	.995	.995	.961	.911	1.097
	X2	-.263	.238	-.040	-1.104	.302	.260	-.364	-.038	.911	1.097

a. Dependent Variable: Y

The t test of coefficients significance, is to test whether independent variable X_j impacts dependent variable Y significantly. Put forward null hypothesis and alternative hypothesis as followed:

$\boldsymbol{H_0}$: $\beta_j=0$ （j=1,2） $\boldsymbol{H_1}$: $\beta_j\neq0$（j=1,2）

To independent variable X1, the estimation of its regression coefficient is 2.198, standard error is 0.078, t test value is 28.085, t test significance is 0.000,lower than 0.01. Therefore we refuses $\boldsymbol{H_0}$, and think independent variable X1 "total agricultural machinery dynamic" is highly significant.

Table 6 Casewise Diagnostics(a)

Case Number	Std. Residual	Y	Predicted Value	Residual
1	-.746	24519.10	25573.4620	-1054.36196
2	-.117	24915.80	25081.2983	-165.49831
3	-.250	26179.60	26532.4078	-352.80776
4	-.321	27390.80	27844.8871	-454.08709
5	.989	29691.80	28294.2473	1397.55270
6	.749	36239.00	35180.1807	1058.81926
7	1.088	39450.90	37913.8606	1537.03937
8	-1.542	40810.80	42989.0636	-2178.26359
9	-.456	48893.00	49537.1275	-644.12746
10	1.244	58002.20	56244.2432	1757.95678
11	-.639	60361.00	61263.2219	-902.22193

a Dependent Variable: Y

To independent variable X2,the estimation of its regression coefficient is -2.63,standard error is 0.238, t test value is -1.104, t test significance is 0.302, higher than 0.05. So we accepts $\boldsymbol{H_0}$, and think that the coefficient of independent variable "total corps planting area" is not significant.

From the column "correlations" of table 5, we can find that no matter zero-order, partial, or part correlation, "total agricultural machinery dynamic" is stronger correlated with "farming; forestry; animal husbandry and fishery output" than "total corps planting area".

From the column "collinearity statistics" of table 5, we can find that the two collinearity statistics "tolerance" and "VIF" are valued 0.911 and 1.097, "tolerance" close to 1, "VIF" near to 1, too. Hence, there is no collinerarity exists between the two independent variables "total agricultural machinery dynamic" and "total corps planting area".

Table 6 is used to observe the exceptional value of "farming; forestry; animal husbandry and

fishery output", it outputs all values of "farming; forestry; animal husbandry and fishery output". If chooses to output exceptional values outside three times standard level, the software will not output this table. Because of table 7, the minimum and maximum of standardized residual is respectively -1.542 and 1.244, both lower than 3. So, there is no exceptional value in table 6.

Table 7 Residuals Statistics(a)

	Minimum	Maximum	Mean	Std. Deviation	N
Predicted Value	25081.2988	61263.2227	37859.4545	12999.66467	11
Std. Predicted Value	-.983	1.800	.000	1.000	11
Standard Error of Predicted Value	441.190	1112.946	717.554	180.350	11
Adjusted Predicted Value	25145.0859	62738.1641	38039.5580	13232.30439	11
Residual	-2178.26367	1757.95679	.00000	1263.74950	11
Std. Residual	-1.542	1.244	.000	.894	11
Stud. Residual	-1.936	1.486	-.050	1.077	11
Deleted Residual	-3434.86206	2508.09375	-180.10343	1872.30345	11
Stud. Deleted Residual	-2.484	1.634	-.075	1.204	11
Mahal. Distance	.066	5.296	1.818	1.404	11
Cook's Distance	.002	.721	.185	.251	11
Centered Leverage Value	.007	.530	.182	.140	11

a Dependent Variable: Y

As Figure 1 indicates, most splashes are close to diagonal, which indicates that standardized residual are obeyed to the Normal distribution.

Besides, in data edit window, we can also see standardized individual prediction value PRE_1, unstandardized residual value RES_1, prediction interval left and right endpoint LICI_1and UICI_1, as figure 2.

To make a conclusion, from all analysis, we can find that the established Dualistic Linearity Regression model fits well, and pass F test, independent colineraity test, standardized residual normality test and other statistical tests, it's reasonable totally. But it has a very serious flaw is that independent variable X2 "total corps planting area" doesn't pass t test. So, the chosen of independent variable need to improve.

3.2 Regression analysis rejected an independent variable

Choose stepwise to screen independent variable, gets results as followed:

From table 8, the standard of choose independent variable is: while entering the model, its F probable value is 0.05, while reject it from the model, the F probable value is 0.10. Combined with table 9, we can find that, SPSS choose independent variable X1 to enter and rejects independent variable X2. Tabel 9 also indicates clearly that independent variable X2's F probable value is 0.302, far bigger than 0.1, so rejects it.

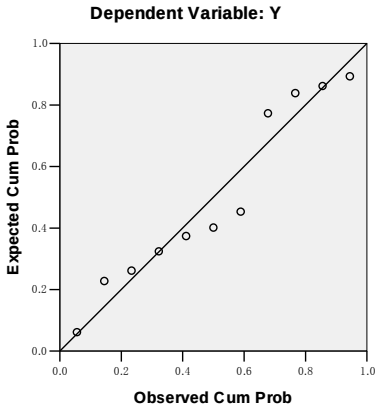

Figure 1 Normal P-P Plot of Regression Standardize Residual

Y	X1	X2	PRE_1	RES_1	LICI_1	UICI_1
24519	14106.20	156373.00	25573.46196	-1054.36196	21892.48452	29254.43941
24916	13873.60	156300.00	25081.29831	-165.49831	21397.67501	28764.92162
26180	14462.80	155708.00	26532.40776	-352.80776	22957.29656	30107.51895
27391	14931.50	154636.00	27844.88709	-454.08709	24350.43806	31339.33612
29692	14870.10	152415.00	28294.24730	1397.55270	24607.14225	31981.35236
36239	18138.40	153553.00	35180.18074	1058.81926	31693.50251	38666.85897
39451	19613.40	155488.00	37913.86063	1537.03937	34500.52470	41327.19656
40811	21522.30	152149.00	42989.06359	-2178.26359	39181.25184	46796.87533
48893	24658.10	153464.00	49537.12746	-644.12746	45847.34893	53226.90600
58002	28044.20	156266.00	56244.24322	1757.95678	52530.64327	59957.84316
60361	30611.10	158639.00	61263.22193	-902.22193	57115.63378	65410.81009

Figure 2 Unstandardized Individual Prediction and Predicted Interval

From table 10, we can find that, the regression model with only one independent variable X1, its goodness of fit is 0.989, a little lower than the dualistic linearity regression model; And adjusted R square is 0.988, equal to the dualistic linearity regression model. These indicate that this one-dimensional linear regression model fits well.

Table 8 Variables Entered/Removed(a)

Model	Variables Entered	Variables Removed	Method
1	X1	.	Stepwise (Criteria: Probability-of-F-to-enter <= .050, Probability-of-F-to-remove >= .100).

a Dependent Variable: Y

Table 9 Excluded Variables(b)

Excluded Variables[b]

Model		Beta In	t	Sig.	Partial Correlation	Collinearity Statistics		
						Tolerance	VIF	Minimum Tolerance
1	X2	-.040[a]	-1.104	.302	-.364	.911	1.097	.911

a. Predictors in the Model: (Constant), X1

b. Dependent Variable: Y

Table 10 Model Summary(b)

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.995(a)	.989	.988	1429.96566

a Predictors: (Constant), X1 b Dependent Variable: Y

Table 11 Analysis of Variance

ANOVA(b)

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	1687480226.536	1	1687480226.536	825.254	.000(a)
	Residual	18403215.971	9	2044801.775		
	Total	1705883442.507	10			

a Predictors: (Constant), X1 b Dependent Variable: Y

From table 11, we can test whether the whole regression equation is linear significant or not, just as table 4. We can see F value is 825.254, probable value is 0.000, and they indicate that regression equation is significant in linear relation. What's more, the F value of one-dimensional linear regression model is bigger than the dualistic linearity regression, which indicates that the former is an improvement of the latter.

Table 12 Coefficients

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Correlations			Collinearity Statistics	
		B	Std. Error	Beta			Zero-order	Partial	Part	Tolerance	VIF
1	(Constant)	-4575.329	1538.798		-2.973	.016					
	X1	2.173	.076	.995	28.727	.000	.995	.995	.995	1.000	1.000

a. Dependent Variable: Y

And from table 12, we can test whether independent variable X1 impacts dependent variable Y significantly. In the table, the t statistic value is 28.727; probable value is 0.000, passes the t test.

Compositing analysis of table 8 to table 12, we conclude that, the one-dimensional linear regression model rejected variable X2 passes statistical tests, and fits well, is a good improvement of the dualistic linearity regression model.

4 Conclusion

This paper applies software SPSS to analyze 1999-2009 relevant statistics of China agriculture industry promulgated by National Bureau of Statistics of China, trying to establish a Dualistic Linearity Regression Model to calculate the effect of the total agricultural machinery dynamic and total corps planting area in the total output of farming, forestry, animal husbandry and fishery industries, but the results indicate that the independence "total corps planting area" doesn't pass the significance test. So the paper chooses stepwise method to do the regression again, and the software eliminates independence "total corps planting area", and works out a one-dimensional linear regression model. the one-dimensional linear regression model rejected variable X2 passes statistical tests, and fits well, is a good improvement of the dualistic linearity regression model.

References

[1] State Bureau of Statistics of PRC. State Statistic Yearbook of 2010: 2-2 the Constitution of GNP[EB/OL]. http://www.stats.gov.cn/tjsj/ndsj/2010/indexch.htm

[2] State Bureau of Statistics of PRC. State Statistic Yearbook of 2010: Explanation of Main Statistical Indicators[EB/OL]. http://www.stats.gov.cn/tjsj/ndsj/2010/indexch.htm

[3] State Bureau of Statistics of PRC. State Statistic Yearbook of 2010: 13-4,13-5,13-13[EB/OL]. http://www.stats.gov.cn/tjsj/ndsj/2010/indexch.htm