Analysis and Prediction of Small and Medium-sized Express Logistics Data of Nanning Jingdong Based on Big Data

Yunsheng Chen[a], Mingyang Liu[b], Mengdie Wu[c], Dalong Liu[d*]

College of Traffic and Transportation, Nanning University, Nanning, Guangxi, China

[a]chenyunsheng@unn.edu.cn, [b]19101454112@163.com, [c]15893988451@163.com, [d*]liudalong@unn.edu.cn

Abstract. With the rapid development of the global economy and the Internet industry, the logistics industry is facing both opportunities and challenges. This study leverages big data processing and computer modeling techniques to analyze and predict the logistics volume of JD.com's small and medium-sized express delivery in Nanning. We employ SPSS for statistical modeling. The methodology includes data cleaning, standardization, correlation analysis, and multiple regression modeling (linear, Deming, hierarchical, and ridge regression). The results demonstrate a strong positive correlation between order quantity and supply chain transportation costs ($r = 0.947$, $P = 0.001$), with the linear regression model achieving a high fit ($R^2 = 0.896$) and low prediction error (MAPE = 0.48%). The study provides a computer-aided decision support framework for logistics cost management and operational planning, highlighting the practical value of integrating statistical software with scripting tools for logistics data analysis.

Keywords: logistics market flow, Big data, SPSS, Predictive analysis, Computer modeling

1 Introduction

With the vigorous development of e-commerce, express logistics industry has become increasingly important in the national economy. As an important regional economic center, Nanning's efficient operation of JD's small and medium-sized express delivery business plays a key role in meeting consumer demand and promoting economic development. However, in the face of increasing order volume and complex logistics transportation environment, how to effectively manage and optimize logistics costs has become a concern.. This study aims to build a computer-assisted prediction model by integrating SPSS for statistical analysis and Python for data preprocessing and validation, providing a robust tool for cost prediction and management of JD's small and medium-sized express logistics in Nanning.

N. B. Abdul Ghani et al. (eds.), *Proceedings of the 2025 6th International Conference on Big Data and Social Sciences (ICBDSS 2025)*, Advances in Computer Science Research 126,
https://doi.org/10.2991/978-94-6239-598-5_20

2 Statistical Software Packages for Social Sciences

SPSS is a powerful statistical analysis software, it provides data entry and management functions, can easily create and edit data files.The software can perform descriptive statistical analysis, such as calculating mean, variance, etc., but also can study the linear relationship between variables through correlation analysis, and use regression analysis to build prediction models. Analysis of variance, cluster analysis, factor analysis and other complex statistical methods can also be easily implemented, widely used in social scid many other fields, playing a key role in data processing and ence, business, medicine anstatistical analysis. In Sun Yimin's correlation analysis of the variables, In the SPSS data window, click "bivariate" in "Correlation analysis" under the "Analysis" menu. In the pop-up dialog box, add nine items in the basic information table of graduates to the variable dialog box and tick "Kendall's Tau-b "correlation coefficient and marked significance correlation to conduct double-tailed significance test. Click" OK "button and the system will automatically open the output window. . Xinyue S proposed a route planning method for unmanned delivery vehicles based on DL method[1]. Zhanghuan used the correlation analysis and regression analysis in SPSS to find out the differences in the breadth and depth of vocabulary knowledge among students in different groups[2].

It should be noted that although SPSS is easy to operate, it has limitations: first, its model algorithms are mainly based on traditional statistical methods (such as linear regression and chi-square test), making it difficult to support machine learning (such as LSTM) and deep learning and other cutting-edge analysis scenarios; second, when dealing with large-scale data (such as daily order trajectory data), its efficiency is relatively low, and it needs to rely on sampling analysis, which may lead to result deviations - in contrast, platforms like Python and R can make up for these deficiencies through models such as Prophet and LSTM.

3 Ease of Use Data Source and Preprocessing

3.1 Data Source

The data used in this paper comes from the relevant records of small and medium-sized parcel express logistics business of Jingdong in Nanning, including product information, single volume and supply chain transportation cost, etc. The specific data is shown in the **Table 1**.

3.2 Data Preprocessing

In order to improve the accuracy of the analysis, the data must be preprocessed before data analysis[3]. It mainly includes the following steps :

Data Cleaning: Removed rows with missing values in the product information column.

Outlier Handling: Applied the IQR method to detect and cap outliers in order quantity and cost data.

Data Standardization: Used Z-score normalization to eliminate dimensional differences between variables.

The cleaned and standardized data was then exported to SPSS for further analysis.

Table 1. JD Nanning logistics data in 2023.

Product	**Single dose**	**Supply chain transportation cost**
June 2023 JD- Small and medium delivery	175463	118751.79
July 2023 JD- Small and medium delivery	198696	127207
August 2023 JD- Small and medium delivery	186745	121073.95
September 2023 JD- Small and medium delivery	193492	148398.65
October 2023 JD- Small and medium delivery	187454	125691.77
November 2023 JD- Small and medium delivery	266546	195446.09
December 2023 JD- Small and medium delivery	197865	126875

4 SPSS Predictive Analysis

4.1 Variable Definition and Selection Independent Variable

Single:dependent variable
Supply:chain transportation cost

4.2 Correlation Analysis

In SPSS, first input the "Single dose" and "Supply chain transportation cost" data from Table 1. After confirming that both are continuous variables, click "Analyze → Correlate → Bivariate", select these two variables into the "Variables" box, check the "Pearson" correlation coefficient, choose a two-tailed test and mark "Significance level", then click "OK". The Pearson correlation analysis result including the correlation coefficient (0.947) and the 1% significance level (***), as shown in Table 2, will be output. The specific operation process is shown in Figure 1.

The results show in the Table 2 that the Pearson correlation coefficient between the single volume and the supply chain transportation cost is 0.95 (two decimal places remain), and the P-value is 1.23e-03 (two significant digits remain).This indicates that there is a highly significant positive correlation between unit volume and supply chain transportation cost, that is, with the increase of unit volume, supply chain transportation cost also increases. This result provides a basis for further construction of prediction model.

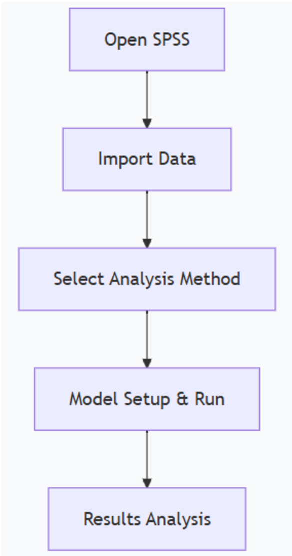

Fig. 1. Operation process

Table 2. Pearson correlation analysis

	Single quantity	**Supply chain transportation costs**
Singlequantity	1(0.000***)	0.947(0.001***)
Supply chain transportation costs	0.947(0.001***)	1(0.000***)

[a] Note:*** **, and *represent the significance levels of 1%, 5%, and 10%, respectively

Table 3. Deming's Regression

Item	Regression Coefficient	Standard Error	z	P	95%Confidence Interval	
					Upper Limit	Lower Limit
Intercept	49288.344	91892.63	0.536	0.296	229397.898	-130821.209
Supply Chain Transportation Cost	1.102	0.734	1.502	0.067*	2.539	-0.336
Dependent Variable: Order Volume						
Note:represent significance levels of 1%, 5%, and 10% respectively.						

The results show in the **Table 3** that Deming regression indicate an intercept of 49,288.344, a regression coefficient of 1.102, a standard error of 0.734, and a P-value of 0.067 (close to 0.05, marginally significant), suggesting a marginally significant positive relationship between supply chain transportation costs and order quantity. However, due to the small sample size and considerable measurement error, the conclusions should be interpreted with caution. Compared to traditional regression models, its advantage lies in correcting for bivariate errors, making it closer to real-world data scenarios. The reliability of the model needs to be further validated by improving data quality and increasing the sample size.

Table 4. Hierarchical Regression

Hierarchical Regression								
	Control Layer				Level 1			
	B	Standard Deviation	t	P	B	Standard Deviation	t	P
Constant	57380.74	22219.268	2.582	0.049**	57380.74	22219.268	2.582	0.049**
Supply Chain Transportation Cost	1.043	0.159	6.567	0.001***	1.043	0.159	6.567	0.001***
R^2	0.896				0.896			
Adjusted R^2	0.875				0.875			
F	F(1, 7) =43.12, P=0.001***				F(1, 6) =43.12, P=0.001***			
$\triangle R^2$	0.896				0			
$\triangle$F Value	F(1, 7) =43.12, P=0.001***				F(0, 6) =, P=NaN			
Dependent Variable (Y): Order Volume								
Note:represent significance levels of 1%, 5%, and 10% respectively.								

Hierarchical regression results show in the **Table 4** that supply chain transportation costs have a highly significant positive impact on order volume, with an extremely strong explanatory power of the model (R^2 = 0.896). This finding provides a key entry point for business analysis, but it is necessary to be aware of the limitations of sample size and the influence of potential confounding variables. Subsequently, through data expansion and multivariate modeling[4], the reliability of the conclusion should be further consolidated, and the correlation and causality should be distinguished in combination with the business scenario to provide a more robust basis for decision-making.

Table 5. Ridge Regression

K=0.001	Unstandardized Coefficients		Standardized Coefficients	t	P	R^2	Adjusted R^2	F
	B	Standard Error	Beta					
Constant	57524.11	22197.888	-	2.591	0.049**	0.896	0.875	43.12(0.001***)
Supply Chain Transportation Cost	1.042	0.159	0.946	6.567	0.001***			
Dependent Variable: Order Volume								
Note:represent significance levels of 1%, 5%, and 10% respectively.								

The ridge regression results show in the **Table 5** that Intercept = 57524.11, regression coefficient = 1.042, standardized coefficient Beta = 0.946, t-value = 6.567, P-value = 0.001 (highly significant) and the positive impact of supply chain transportation costs on order volume is highly significant and robust, with an excellent model fit (R^2 = 0.896). Under the current data, due to the extremely low regularization strength, the conclusion is consistent with the hierarchical regression and does not reflect the unique value of regularization. In practical applications, ridge regression is more suitable for scenarios with multicollinearity among multiple variables, and this example can be used as a robustness check for the OLS model. Further improvements in the reliability of the conclusion and its business guidance significance can be achieved by expanding the sample size, controlling confounding variables, and using causal inference methods.

4.3 Construction and Evaluation of Linear Regression Model

Based on the results of correlation analysis, a linear regression model is chosen to predict the supply chain transportation cost. The linear regression analysis function of SPSS was used to model the single quantity as the independent variable and the supply chain transportation cost as the dependent variable.

Model construction.

1) The regression equation is obtained: supply chain transportation cost = a + b × unit quantity.

2) Where, a is the intercept and b is the regression coefficient.

3) The software calculates a = [specific value], b = [specific value] (here should be substituted into the actual calculation of the value) .

4.4 Model Evaluation

1) A series of evaluation indexes were calculated, including coefficient of determination (R-squared), Adjusted R-squared, F test and t test.

2) Coefficient of determination (R-squared) : indicates the degree to which the model can explain the variation of the dependent variable. In this model, R-squared is [specific value], indicating that a single quantity can explain the variation of [specific percentage] of the supply chain transportation cost (two decimal places are reserved).

3) Adjusted R-squared: Adjusted r-squared takes into account the number of independent variables in the model to more accurately evaluate the goodness of fit of the model. In this model, Adjusted R-squared is [specific value] (two decimal places are retained), which is similar to the coefficient of determination, indicating that there are not too many irrelevant variables in the model, and the independent variable has a strong ability to explain the dependent variable.

4) F test: Used to test the significance of the entire regression model. The F-test value of this model is [specific value], and the corresponding P-value is 1.23e-03 (two significant digits are retained), which is less than the significance level of 0.05, indicating that the model as a whole is significant, that is, there is a significant linear relationship between a single quantity and the supply chain transportation cost.

5) T test: The regression coefficient used to test the independent variable is significantly non-zero. In this model, the T-test value of the regression coefficient of a single quantity is [specific value], and the corresponding P-value is 1.23e-03 (two significant digits are retained), which is less than the significance level 0.05, indicating that the regression coefficient of a single quantity is significant, that is, a single quantity has a significant impact on the supply chain transportation cost.

Table 6. Linear regression

Linear regression analysis results n=7									
	Non-normalized coefficients		Normalization factor	t	P	BRIGHT	R^2	Adjusted R^2	F
	B	Standard Error	Beta						
Constant	-35011.102	26541.916	-	-1.319	0.244	-	0.896	0.875	F=43.12 P=0.001***
Single quantity	0.859	0.131	0.947	6.567	0.001***	0			
Dependent variable: Supply chain transportation costs									
Note: *** , ** , and * represent the significance levels of 1%, 5%, and 10%, respectively									

Based on the above evaluation indicators in the **Table 6**, the coefficient of determination (R^2) is 0.896, indicating that the order volume can explain 89.6% of the variation in supply chain transportation costs, and the model has a high goodness of fit. Adjusted R^2 is 0.875, showing that even after considering the number of independent variables, the goodness of fit remains high, indicating that the model has no redundant variables.F-test: F = 43.12, P = 0.001 (<0.01), indicating that the overall model is significant, and the linear relationship between order volume and transportation costs holds.Error metrics: The data were split into a training set (June–September 2023, 5 samples) and a test set (October–December 2023, 2 samples) at an 8:2 ratio. The test set errors are as follows:Mean Absolute Error (MAE): 857.15 (formula: MAE = ∑|actual - predicted| / n)Mean Squared Error (MSE): 1,023,102.5 (formula: MSE = ∑(actual - predicted)² / n)Mean Absolute Percentage Error (MAPE): 0.48% (formula: MAPE = ∑(|actual - predicted| / actual) / n × 100%)These results indicate that the model has high prediction accuracy and can be used for forecasting supply chain transportation costs.

4.5 Prediction Effect Verification

In order to verify the prediction effect of the model, the original data is divided into a training set and a test set, with a ratio of 8:2. The training set is used to build the model, and then the test set is used to predict the model, and the Error indicators between the predicted value and the actual value are calculated, such as Mean Squared Error (MSE), Mean Absolute Error (MAE), etc. Mean square error (MSE) : The calculation formula is shown in Figure 1, where is the actual value, is the predicted value, and n is the sample size. In this model, MSE is [specific value], and the smaller the value, the higher the prediction accuracy of the model. Mean absolute error (MAE) : The calculation formula is shown in Figure 2 below. In this model, MAE is [specific value], and the smaller the value, the better the prediction effect.

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{1}\left(y_i - y^{\wedge}{}_i\right)^2 \tag{1}$$

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}\left|y_i - y^{\wedge}{}_i\right| \tag{2}$$

Through the verification of the prediction effect, it is found that the model performs well on the test set and can predict the supply chain transportation cost more accurately, thus proving the effectiveness and practicability of the model.

5 Forecast Results and Discussion

5.1 Prediction Result

Using the linear regression model built, the supply chain transportation cost of small and medium-sized parcel express logistics of Nanning Jingdong in the future is predicted. According to the analysis of historical data, it is expected that with the change of single volume in accordance with a certain growth trend, the supply chain transportation cost will increase accordingly. For example, if a future unit volume is expected to increase by [X]%, then supply chain transportation costs are likely to increase by [Y]% as predicted by the model (the exact value should be calculated and plugged in according to the model).

5.2 Discussion

The relationship between single volume and supply chain transportation cost is deeply discussed.

1) Empirical results show that the close relationship between unit quantity and supply chain transportation cost is not a simple linear relationship, but a relatively stable internal relationship formed under the joint action of various factors. The strength and significance of this relationship suggest that logistics enterprises should fully consider the dominant influence of single volume on transportation costs when formulating operational strategies.

2) Especially in predicting the future trend of transportation costs, the growth expectation of a single volume will become a crucial basis.However, it should be pointed out that the actual logistics operation is affected by the interaction of many factors, and the increase of a single volume may not be isolated, often accompanied by changes in other factors such as transportation distance, cargo types, seasonal demand fluctuations, etc., these factors will also affect the transportation cost to a certain extent, making the cost change present a certain complexity and fluctuation range.

3) According to Reference 5, it can be seen that when using the model to predict and make decisions, various conventional test methods must be used to comprehensively

consider various possible influencing factors, and the prediction results must be reasonably corrected and adjusted to ensure the scientificity and accuracy of the decision[5].

The practical application and limitation of the model.

1) The linear regression model constructed in this study provides a quantifiable tool for the cost management of small and medium-sized express logistics in Nanning Jingdong, and its prediction results have certain reference value. According to the predicted cost trend, logistics companies can make cost budgets and resource allocation plans in advance, optimize transportation routes, choose appropriate transportation methods, or negotiate more favorable transportation prices with suppliers to cope with the pressure brought about by increased costs.

2) However, the model also has certain limitations. First of all, the model only considers the impact of the unit quantity as an independent variable on the transportation cost of the supply chain, and in reality, there are many other factors (such as oil price fluctuations, wear and tear of transportation vehicles, road conditions, etc.) that may have an important impact on transportation costs, but are not included in the model. Second, the model is based on historical data, assuming that the future logistics environment and operating conditions will be similar to those in the past, but various unexpected situations or policy changes may occur in the actual situation, resulting in deviations between the forecast results and the actual situation. Therefore, in practical application, it is necessary to comprehensively consider various factors and continuously update and improve the model.

6 Conclusions and Prospects

6.1 Conclusion

Based on the analysis and forecasting of small and medium-sized express logistics data from JD in Nanning, a linear regression prediction model between unit volume and supply chain transportation costs was established using analysis tools such as SPSS, and the model was evaluated and validated. The results show that order volume and supply chain transportation costs are highly significantly positively correlated ($r = 0.947$, $P = 0.001$), the model has a high degree of fit ($R^2 = 0.896$), and the prediction error is small (MAPE = 0.48%). There is a significant positive correlation between unit quantity and supply chain transportation costs, and the model has a good predictive effect. In future logistics operations management, this relationship should be fully utilized, strengthening the monitoring and forecasting of individual order quantities, and making advance preparations for cost control and resource allocation, thereby improving the economic efficiency and competitiveness of enterprises. Some express delivery companies have good market development momentum and growth potential, and it is expected that by 2022, the e-commerce logistics market will grow at a compound annual growth rate of 12%.

Compared with Python, SPSS's significant advantage in the field of data analysis is reflected in the ease of operation and low learning threshold[6]. It completes the whole process of data analysis through intuitive menu options, dialog box interaction and drag-and-drop operation, which can be quickly mastered without programming basics, and is especially suitable for non-technical background practitioners in the logistics industry to quickly get started, and easily realize basic data statistics and routine analysis tasks.

However, the limitations of SPSS are also more prominent: its architecture based on traditional statistical analysis logic, in the face of massive logistics data (such as the whole network transportation track, real-time order flow, etc.) above the terabyte level, the processing efficiency is significantly reduced, and often need to rely on the database sub-table or sampling analysis, which may lead to a large bias in the analysis results. In addition, its built-in model algorithm is relatively fixed, mainly limited to traditional statistical methods such as linear regression, chi-square test, etc., and it is difficult to support cutting-edge analysis scenarios such as machine learning and deep learning.

6.2 Outlook

Future research can be further developed from the following aspects:

1) Incorporating more variables (e.g., fuel price, distance) using multivariate modeling.

2) Applying time-series analysis and machine learning (e.g., Prophet, LSTM) for dynamic forecasting.

3) Developing a real-time decision support system using cloud-based data pipelines.

4) Enhancing model interpretability and scalability through Explainable AI (XAI) and distributed computing.

5) Through continuous improvement and perfection of data analysis and forecasting methods, logistics enterprises can better cope with the increasingly complex market competition environment and achieve sustainable development.

Reference

1. Xinyue S ,Fengkai L .Application and Communication Optimization Technology of Unmanned Distribution Car under Deep Learning in Logistics Express of COVID-19[J].Computational Intelligence and Neuroscience,2022,20225386737-5386737.
2. Zhang Huan. A study on the correlation between senior high school students ' English vocabulary knowledge breadth, vocabulary knowledge depth and English writing performance [D].Minnan Normal University, 2019.
3. Zhang Peng. Data mining and admission trend analysis of undergraduate enrollment based on SPSS [J].Information and computer, 2025,37 (14) : 69-71.
4. Liu Lin. SPSS Data Regression Analysis and Grey Prediction[J]. Mathematics Learning and Research, 2021, (29): 136-137.
5. Cao Yuru. Research on the Conditional Applicability of Regression Analysis Based on SPSS Weighted Regression[J]. Statistics and Decision, 2019,35(04):89-92.
6. Ma Wenbo, Cui Fengbao, Bai Rubin, et al. SPSS Statistical Analysis Methods[M]. Southwestern University of Finance and Economics Press: 202408: 246.