Research on Optimization of Rural Financial Policies Based on Econometrics

Ziyi Chen\(^1, a\)

\(^1\) School of Economics and Business Administration, Central China Normal University.
\(^a\) ziyichen@mails.ccnu.edu.cn

**Abstract.** Rural finance is vital for rural economic development. This study uses extensive empirical data and econometric models to analyze factors influencing rural finance development in China. It employs multiple linear regression, panel data, and time series models, optimizing them through feature selection, non-linear processing, and cross-validation. Scenarios such as baseline, fiscal support, interest rate preferences, and inclusive finance are simulated to assess impacts on rural loan balances and resident income levels. The findings highlight that comprehensive measures yield better policy synergies, offering policy optimization suggestions for decision-makers.

**Keywords:** rural finance; econometric models; policy simulation.

1. Introduction

Rural finance is crucial for modern agriculture and rural economic development. Despite progress, challenges like insufficient service coverage and inadequate credit supply hinder growth. This study aims to identify key factors affecting rural finance development in China, using econometric models and big data analysis[1]. By evaluating policy impacts and simulating various scenarios, we will provide theoretical foundations and decision-making references to enhance rural finance policies and promote economic revitalization in rural areas.

2. Data Collection and Preprocessing

2.1 Data Sources

The data for this study are sourced from the China Rural Finance Yearbook, China Statistical Yearbook, provincial and municipal statistical yearbooks, and third-party platforms like Alibaba Cloud. The Rural Finance Yearbook offers historical data on loan and deposit balances, and the number of rural financial institutions. The Statistical Yearbooks provide macroeconomic data such as per capita disposable income, agricultural output, and rural consumption levels. Third-party platforms contribute alternative data, such as mobile user traffic and geographic information, reflecting local economic activity levels (see Table 1).

<table>
<thead>
<tr>
<th>Year</th>
<th>Rural Credit Cooperatives</th>
<th>Rural Commercial Banks</th>
<th>Rural Mutual Aid Funds</th>
</tr>
</thead>
<tbody>
<tr>
<td>2015</td>
<td>1235.6</td>
<td>328.4</td>
<td>75.2</td>
</tr>
<tr>
<td>2016</td>
<td>1424.7</td>
<td>405.3</td>
<td>88.6</td>
</tr>
<tr>
<td>2017</td>
<td>1689.5</td>
<td>519.8</td>
<td>102.7</td>
</tr>
<tr>
<td>2018</td>
<td>1876.2</td>
<td>672.5</td>
<td>121.4</td>
</tr>
</tbody>
</table>

By collecting data through multiple channels, we can obtain comprehensive data support, providing a data foundation for subsequent model construction and policy simulations[2].

2.2 Data Cleaning

Due to differences in sources and definitions of indicators, raw data often contain issues such as missing values, outliers, and duplicates. Therefore, data cleaning is necessary. Initially, the data is explored using Python libraries such as statsmodels and pandas. Missing value distribution charts, outlier distribution charts, etc., are plotted, as shown in Figure 1.
To handle missing values, continuous variables are filled using interpolation or mean imputation, while discrete variables are addressed by deletion or mode imputation. Outliers are managed by combining statistical criteria and expert judgment, deleting erroneous values and retaining reasonable outliers. Duplicate values are considered based on specific circumstances[3]. Data cleaning ensures accuracy and completeness, providing a solid foundation for feature engineering and model construction. This process results in a high-quality dataset suitable for model input.

2.3 Feature Engineering

Feature engineering is crucial for model performance. For this study, we constructed: 1) Static features: region, year, rural population, per capita net income, consumption level, and number of rural employees; 2) Lag features: previous periods' loan and deposit balances; 3) Combination features: cross features of region and year, income-to-consumption ratios; 4) Trend features: annual growth rates of rural income; 5) Text features: keywords from economic policy documents using TF-IDF. These features are based on insights from rural finance business[4]. Table 2 provides a statistical summary of selected features for 2017.

<table>
<thead>
<tr>
<th>Feature</th>
<th>Mean</th>
<th>Standard Deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Region</td>
<td>11.5</td>
<td>6.78</td>
</tr>
<tr>
<td>Per Capita Net Income (RMB)</td>
<td>13471</td>
<td>2836</td>
</tr>
<tr>
<td>Number of Employees (ten thousand)</td>
<td>3258</td>
<td>1027</td>
</tr>
</tbody>
</table>

Through feature engineering, we can extract more potential information from limited raw data, providing high-quality input for the model.

2.4 Data Visualization Analysis

Data visualization is an important and indispensable part of the data analysis process. It helps us intuitively understand the distribution characteristics of the data, discover potential patterns, and identify anomalies. In this study, we used several Python visualization libraries to conduct multidimensional visualization analysis of the data. Firstly, we plotted the distribution histogram of rural loan balances for each province (Figure 2). We found that the loan balances of most provinces were concentrated within a certain range, but there were also outliers in some provinces, which may require separate analysis.

Secondly, we studied the relationship between rural resident income, consumption level, and financial institution loan balances. Through scatter plots (Figure 3), we observed a positive correlation between resident income, consumption level, and loan balances, with a positive slope of the fitted line.[5]
Figure 2: Distribution Histogram of Rural Loan Balances for Each Province

Figure 3: Relationship between Resident Income, Consumption, and Loan Balances

Finally, we visualized the temporal trends of rural financial indicators, as shown in Figure 4. It can be seen that both loan balances and deposit balances demonstrate a continuous growth trend over time.

Figure 4: Time Series Trends of Rural Financial Indicators

Through visual analysis, we gained a deeper understanding of the data, which is helpful for feature construction, model optimization, and provides insights for subsequent policy analysis. In summary, data visualization is an indispensable part of the analysis process.

3. Model Construction and Optimization

3.1 Selection of Econometric Models

In the field of econometrics, various model methods can be chosen depending on the research question. For this study, we primarily considered the following types of models:

1. Multiple Linear Regression Model: This is the most basic and commonly used regression model, capable of capturing the linear relationship between independent and dependent variables. Its basic form is as follows:

\[ y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n + \epsilon \]

Where \( y \) is the dependent variable, \( x_1, x_2, \ldots, x_n \) are the independent variables, \( \beta_0 \) is the constant term, \( \beta_1, \beta_2, \ldots, \beta_n \) are the regression coefficients, and \( \epsilon \) is the random error term.

2. Panel Data Models: Since our research data are panel data across provinces, cities, districts, and years, panel data models are a natural choice. They mainly include fixed effects models and random effects models:

   Fixed effects: \( y_{it} = \alpha_i + \beta x_{it} + u_{it} \),
   Random effects: \( y_{it} = \alpha + \beta x_{it} + u_i + \epsilon_{it} \)
Where \( i \) represents the individual, \( t \) represents time, \( \alpha \) and \( \alpha_i \) are intercept terms for the overall population and individual, \( u_{it} \) and \( \epsilon_t \) are composite error terms.

3. Time Series Models: With the passage of time, financial and economic data often exhibit certain trends and periodicity. Therefore, time series models such as the ARIMA model are useful. Its basic form is:

\[
y_t = c + \phi_1 y_{t-1} + \ldots + \phi_p y_{t-p} + \theta_1 \epsilon_{t-1} + \ldots + \theta_q \epsilon_{t-q} + \epsilon_t
\]

Where \( c \) is the constant term, \( \phi_1, \ldots, \phi_p \) are autoregressive parameters, \( \theta_1, \ldots, \theta_q \) are moving average parameters, \( \epsilon_t \) is the white noise. In the actual modeling process, we will choose one or several combinations of the above models based on the statistical characteristics of the data and practical business needs. Table 3 shows an example of coefficient estimation of a multiple linear regression model in practical application:

<table>
<thead>
<tr>
<th>Feature</th>
<th>Coefficient Estimate</th>
<th>Standard Error</th>
<th>t-Statistic</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant Term</td>
<td>182.56</td>
<td>24.77</td>
<td>7.38</td>
<td>0</td>
</tr>
<tr>
<td>Per Capita Income</td>
<td>0.032</td>
<td>0.005</td>
<td>6.21</td>
<td>0</td>
</tr>
<tr>
<td>Consumption Level</td>
<td>0.028</td>
<td>0.007</td>
<td>4.15</td>
<td>0.002</td>
</tr>
</tbody>
</table>

By comparing the explanatory power and predictive accuracy of different models, we will select the optimal model for subsequent policy simulations and decision analysis[6].

3.2 Model Training

After selecting the model, we need to train it to fit the research data. Taking the multiple linear regression model as an example, we used the statsmodels library in Python to estimate the regression coefficients through the method of least squares. The specific code is as follows:

```python
import statsmodels.api as sm

X = data[['income', 'consumption', ...]]  # 独立变量
y = data['loan_balance']  # 因变量

X = sm.add_constant(X)  # 增加常数项
model = sm.OLS(y, X).fit()  # 最小二乘法拟合
print(model.summary())  # 输出模型汇总信息
```

Part of the output during the model training process is shown below:

It can be observed that the variables of per capita income and consumption level have a significant positive impact on loan balances, consistent with the results of the previous data visualization analysis[7].

For the time series model, we used the ARIMA model and achieved data stationarity through differencing. Part of the Python code is as follows:

```python
from statsmodels.tsa.arima.model import ARIMA

# 干预性检验和差分
...

# ARIMA模型训练
model = ARIMA(data['loan_balance'], order=(2,1,1)).fit()
print(model.summary())
```
Through the above training, we obtained models that can fit the research data well, laying the foundation for subsequent model evaluation, optimization, and policy simulations.

3.3 Model Evaluation

After constructing the initial model, we comprehensively evaluated its performance on the training and testing datasets to guide subsequent optimization. 1) Model Fit Assessment: Using R-squared and adjusted R-squared metrics, the multiple linear regression model showed R-squared of 0.782 and adjusted R-squared of 0.776 on the training set, indicating acceptable fit but room for improvement. 2) Residual Analysis: This included tests for normality, heteroscedasticity, and autocorrelation of residuals, revealing areas needing enhancement:

![Figure 5: Normality Test Results of Residuals for the Multiple Linear Model](image)

From Figure 5, it can be seen that the residuals passed the normality test. From Figure 6, it can be observed that most of the autocorrelation and partial autocorrelation of residuals are within the confidence interval, indicating that the residuals are close to white noise. Therefore, the residual analysis results are satisfactory.

3.4 Model Optimization

To enhance model performance, we optimized through: 1) Feature Selection: Identifying and removing redundant features improved generalization. 2) Nonlinear Processing: Constructing polynomial features and using nonlinear models like decision trees and SVM helped capture nonlinear relationships. 3) Hyperparameter Optimization: Grid search and random search were used to fine-tune parameters for optimal model configurations. 4) Cross-Validation: K-fold cross-validation reduced overfitting and improved generalization. These optimizations significantly improved the models[9]. For instance, the multiple linear regression model’s R-squared increased from 0.782 to 0.856, and RMSE decreased from 128.95 to 95.31, enhancing both accuracy and generalization.

4. Policy Simulation

4.1 Setting Policy Scenarios

After constructing a satisfactory model, we conduct policy simulation analysis by setting different policy scenarios to explore changes in rural finance development under various interventions. The scenarios considered are: 1) Baseline Scenario: No new policies; rural finance develops on its own inertia. 2) Fiscal Support Policy: Government increases fiscal subsidies to rural
financial institutions by 10% annually to expand loan sizes. 3) Interest Rate Preferential Policy: Interest rates for rural loans are reduced by 20% to stimulate demand. 4) Financial Inclusion Policy: Establishes at least one financial service outlet in each county, increasing the number of outlets by 30%. 5) Comprehensive Policy: Combines fiscal support, interest rate preferences, and financial inclusion. We input these policy parameters into the model to dynamically simulate and forecast rural finance development for the coming years[10].

4.2 Policy Simulation Result Analysis

After conducting simulations under various policy scenarios, we proceed to interpret and analyze the simulation results. Visualization is a useful tool for discovering underlying patterns and insights behind the data. Figure 7 presents a comparative graph of policy simulation results, illustrating the development trends of rural loan balances under different policy scenarios.

![Figure 7: Comparison of Rural Loan Balance Development Trends under Different Policy Scenarios](image)

The graph shows that policy interventions stimulate rural loan growth, with the comprehensive policy having the most significant impact, raising the loan balance to 4529.8 billion yuan by 2027. Fiscal support and interest rate discounts are also effective, while financial inclusion has a smaller impact. Multiple policies combined yield better benefits.

5. Conclusion

This study examines the current state and challenges of rural finance in China, constructing econometric models to evaluate different policy scenarios. Findings indicate that coordinated policy measures, such as increasing fiscal support, implementing interest rate discounts, and improving financial infrastructure, significantly enhance rural finance development. These measures inject vitality into rural economic revitalization. Continuity and implementation of policies are crucial for effectiveness. The proposed models and policy suggestions offer theoretical support and decision-making references for developing effective rural finance strategies.

Reference


