

Prediction of the Household Consumption in China by Linear Regression

Junyu Qian *

Nanjing Foreign Language School, Nanjing, China

* Corresponding Author Email: junyuqian9@gmail.com

Abstract. Household consumption plays a crucial role in economic growth and accounts for a large proportion of GDP. So, making accurate predictions of household consumption trends is highly important to predict macroeconomic activity. This paper estimates the impact of disposable income per head and the Gini coefficient on household consumption in China by a multiple linear regression equation. It is constructed based on annual data coming from the CEI website during the recent 20 years. The multiple linear regression equation is calculated by the Excel software to get the coefficients of each determinant. The paper also predicts the disposable income per head using Log-linear Trend Model, which is used to input into the multiple linear regression equation and get the final prediction of household consumption. Moreover, a VIF method is used to prove the effectiveness of the multiple linear regression equation. The results show that both variables have significant positive effects and the household consumption will have a slight increase in 2025.

Keywords: Household consumption; Multiple linear regression; Log-linear Trend Model.

1. Introduction

Household consumption is a fundamental economic knowledge, representing the total expenditure by households on both essential and non-essential goods and services in order to satisfy their daily needs, such as food, healthcare, housing, education, and transportation. These expenditures reflect not only the living standards of residents but also the economic activity in a country. In macroeconomy, household consumption is widely regarded as a important driver of economic growth. It is also a significant component of Gross Domestic Product (GDP). For instance, household consumption accounted for approximately 66.87% of China's GDP in 2024 on average [1]. This large proportion indicates that fluctuations in household consumption can have significant impacts on overall economic performance. So accurate prediction of household consumption can make a significant contribution to forecasting future GDP and understanding economic trends.

Over time, experts have developed different methods to explain consumption behavior with its determinants. In past research on this subject, a classical approach is the Engel Curve [2], which emphasizes the relationship between income and household expenditure. Engel's Law suggests that the proportion of income spent on necessities such as food decreases as income increases but the share allocated to other goods and services increases. This relationship provides valuable suggestions about how consumption patterns change as the rise of income levels and has been widely applied in studies of household expenditure.

After that, the Permanent Income Hypothesis appeared and provided better model [3]. According to this model, household consumption is influenced more strongly by permanent income rather than by temporary fluctuations in current income. In other words, individuals adjust their consumption based on their expectations of long-term income. While both the Engel Curve and the Permanent Income Hypothesis offer important models, they are limited because they mainly focus on the effect of income, but do not account for other potential determinants of consumption.

To overcome these limitations, there are better analytical models. One example is the Demand System [4], which allows researchers to use multiple determinants of consumption simultaneously. This approach can cover the effects of prices, income, and substitution of a good. It means a more complete model of consumer behavior and expenditure allocation

However, these methods are generally more suitable for dissolving microeconomic problems. To analyze macroeconomic data, the paper chose to use Multiple Linear Regression [5]. This approach accounts for multiple determinants at the same time. This paper selects disposable income per head and the Gini coefficient as the explanatory variables. The prediction results can promote government policy adjustments and help businesses decide whether it is an opportune time to expand production to meet consumer demand.

2. Method

2.1. Dataset Perparation

There are only 5 years' data shown above. However, the paper will include 19 years' data items, which are between 2005 to 2024. The data are latest and suitable for the prediction.

Table 1. The data in China on the CEI from 2005 to 2024(sample) [6].

Year	Disposable income per resident (Income)	Gini coefficient (Gini)	Household consumption
Unit	ten thousand RMB	\	ten thousand RMB
2020	3.22	0.47	3949934122
2021	3.51	0.47	4504424935
2022	3.69	0.47	4661812203
2023	3.92	0.47	5121205504
2024	4.13	0.47	5388282000

2.2. Data Preprocessing

At this processing, the paper still uses the Table 1 data above. But they are processed by natural logarithms. The t value is the year gap shown in Table 2, which starts from 2004.

Table 2. The data in China on the CEI from 2005 to 2024(sample) [6].

t	Gini coefficient (Gini)	Disposable income per resident (Income)	$\ln(Gini)$	$\ln(Income)$
16	3.22	0.47	-0.7550226	1.16938136
17	3.51	0.47	-0.7550226	1.25561604
18	3.69	0.47	-0.7550226	1.30562646
19	3.92	0.47	-0.7550226	1.36609165
20	4.13	0.47	-0.7550226	1.41827741

2.3. Multiple Linear Regression

Multiple linear regression is a model to estimate multiple determinants' effect on only one variable. The general form of the model is $Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p + \epsilon$, where the coefficients represent the effect of each predictor on the dependent variable.

MLR is widely used for prediction. It can estimate outcomes, such as predicting house prices based on size, location, and age. It can also help to understand which variables have the greatest impact. The coefficients of the model are typically estimated using the ordinary least squares (OLS) method [7].

However, MLR relies on several assumptions [7], including linearity, independence of errors, constant variance, and low multicollinearity among predictors. When these assumptions are held, MLR provides a powerful and interpretable tool for data analysis across fields such as economics, finance, and social sciences [7].

So, the equation needs to make the two determinants as two variables and set the household consumption as the dependent variable and get the Multiple Linear Regression equation.

$$\text{Household Consumption} = \beta_1 \text{Income} + \beta_2 \text{Gini coefficient} + \epsilon \quad (1)$$

In this equation the β_1 and β_2 represent the coefficient of both Disposable income per resident (Income) and Gini coefficient (Gini). The ϵ is the intercept term.

2.4. The Feasibility of the Equation

The equation can use the VIF to check the Multicollinearity [7].

$$VIF = \frac{1}{1-R^2} \quad (2)$$

In this equation VIF is a constant to represent the Multicollinearity of the function. The degree of multicollinearity decreases as the data approaches zero [7].

2.5. The Prediction of Determinants

If household consumption wants to be predicted accurately, it needs to predict the two determinants first. In this regard, the paper chooses Log-linear Trend Model to solve it [8].

$$\ln(Y_t) = \alpha + \beta t \quad (3)$$

In this equation the Y_t is the Gini coefficient in t year. t is the time gap between the standard year (2004) and another year.

3. The Results and Discussion

3.1. The Result of Multiple Linear Regression Equation

The result of multiple linear regression equations is provided in Table 3, Table 4 and Table 5. Firstly, the equation can use Excel to calculate the coefficients.

Table 3. The summary output of the Multiple linear regression equation.

SUMMARY OUTPUT		
Regression Statistics		
Multiple R		0.99921047
R Square		0.99842157
Adjusted R Square		0.99823587
Standard error		62856155.8
Observations		20

Table 4. The coefficient of the Multiple linear regression equation

	Coefficients	Standard error	t Stat	P-value	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%
Intercept	-2.697E+09	1148852028	-2.34750270	0.03126642	-5.121E+09	-273067368	-5.121E+09	-273067368
Disposable income per resident	1360145083	18624677.1	73.02919021	1.184E-22	1320850450	1399439717	1320850450	1399439717
Gini coefficient	5153856765	23551125682	1.88369610	0.04289405	185003581	1.0123E+10	185003581	1.0123E+10

Table 5. The residual output of the Multiple linear regression equation.

RESIDUAL OUTPUT			
Observations	predict Household consumption	Residuals	Standardized Residuals
1	698949396	57128849	0.96085974
2	807761002	38054486.4	0.64004482
3	946642746	57356379.4	0.96468661
4	1188601626	-35905556	-0.6039016
5	1324616134	-63268112	-1.064117

The result shows that the two coefficients are both significantly relevant in household consumption. Because the P-value of two variable are both less than 0.05. So, these two determinants can affect

household consumption significantly. The value of R Square is 0.99842157, which means the equation can estimate 99% data. So, the equation is valid

3.2. The Result of VIF

The Gini coefficient is conducted a regression on per resident disposable income and get the R² value.

Table 6. Summary output.

SUMMARY OUTPUT	
Regression Statistics	
Multiple R	0.72465379
R Square	0.52512311
Adjusted R Square	0.49874106
Standard error	0.00629071
Observations	20

The result shows the R Square is 0.52512311 shown in Table 6. Put this result into the equation and find the VIF is 2.08, which means these two determinants have little multicollinearity.

3.3. The Result of Gini Coefficient

Table 7. The coefficient of Log-linear Trend Model equation.

	Coefficients	Standard error	t Stat	P-value	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%
Intercept	-0.7195433	0.00577275	-124.64487	6.907E-28	-0.7316714	-0.7074152	-0.7316714	-0.7074152
t	-0.0023869	0.0004819	-4.95302320	0.0010279	-0.0033993	-0.0013744	-0.0033993	-0.0013744

So the coefficient of Gini coefficient is got (Table 7), which means the figure for next year will be 0.463.

3.4. The Result of The Household Consumption In 2025

According to the data from CEI, the Disposable income per resident is 43.4 thousand RMB [6]. Then put this data into the equation of household consumption

$$\text{Household Consumption} = 1360145083 \times \text{Income} + 5153856765 \times \text{Gini-coefficient} - 2.697\text{E}+09 \quad (4)$$

The result is 55922653424.15 thousand RMB.

3.5. Discussion

This study utilizes a multiple linear regression model to explore the relationship between household consumption, disposable income, and income inequality in China. The empirical results indicate that the model achieves a high level of explanatory power, with an R² value close to 1, suggesting that most of the variation in consumption can be explained by the selected variables. By incorporating both income level and income distribution, the analysis provides a broader perspective compared with traditional single-variable approaches.

Another strength of the model lies in its interpretability. The estimated coefficients clearly indicate the direction and magnitude of the influence of each variable, making the results easier to understand and apply in policy analysis. In addition, multicollinearity is examined using the VIF test, and the relatively low value confirms that the regression results are statistically reliable. However, despite these advantages, the method also has several limitations.

One important limitation is that the model relies on the assumptions of linear regression, that is the data must be linear. In reality, the relationship between income inequality and consumption may be nonlinear or involve some special time period, especially during periods such as economic crises.

Compared with more flexible models (such as nonlinear regression [9]), MLR may oversimplify these relationships.

Second, the study uses a relatively small sample size covering annual data from 2005 to 2024. While the model shows high goodness-of-fit, such a small dataset may lead to overfitting and limit the generalizability of the results. Previous studies using lots of data or cross-regional data can provide a more general conclusion.

Finally, the prediction part of the study relies on a log-linear trend model [8], which assumes a stable time trend. This may not adequately reflect future uncertainties or structural economic changes. In contrast, more advanced forecasting methods (like ARIMA [10]) could potentially provide more accurate predictions under changing economic conditions.

4. Conclusion

This paper analyzes household consumption in China by estimating the effects of disposable income per head and the Gini coefficient using a multiple linear regression model. The results indicate that both variables have significant positive impacts on household consumption, and the model has strong interpretability and reliability. The findings suggest that both income growth and income distribution play important roles in influencing household consumption, which provide useful advice for policymakers and businesses in forecasting future demand.

However, the model relies on linear assumptions and a relatively small sample size, which may limit its generalizability. Future research may use more comprehensive datasets and advanced models to improve prediction accuracy and decrease effect of economic uncertainties.

References

- [1] The Global Economy. Household consumption as percent of GDP [Internet]. 2024 [cited 2026 Apr 14]. Available from: The World Bank.
- [2] Chai A, Moneta A. Retrospectives: Engel curves. *J Econ Perspect*. 2010;24(1):225–240.
- [3] Friedman M. *Theory of the consumption function*. 2018.
- [4] Deaton A, Muellbauer J. An almost ideal demand system. *Am Econ Rev*. 1980;70(3):312–326.
- [5] Campbell JY, Mankiw NG. Consumption, income, and interest rates: Reinterpreting the time series evidence. *NBER Macroecon Annu*. 1989;4:185–216.
- [6] China Economic Net. Gini coefficient (2005–2024) and disposable income per resident (2005–2025) [Internet]. [cited 2026 Apr 14]. Available from: National Bureau of Statistics of China.
- [7] James G, Witten D, Hastie T, Tibshirani R, Taylor J. Linear regression. In: *An Introduction to Statistical Learning: With Applications in Python*. Cham: Springer; 2023. p. 69–134.
- [8] Wooldridge JM. *Introductory econometrics: A modern approach*. 2016.
- [9] Gallant AR. Nonlinear regression. *Am Stat*. 1975;29(2):73–81.
- [10] Ariyo AA, Adewumi AO, Ayo CK. Stock price prediction using the ARIMA model. In: *Proc 16th UKSim-AMSS Int Conf Comput Model Simul*. IEEE; 2014. p. 106–112.