Dean&Francis

ISSN 2959-6157

Application of Linear Regression Models in Multidisciplinary Data Analysis: From Economics to Health

Zihan Xu

Qingdao Number 9 High School, Qingdao, China

Corresponding author: xuzhiqiang@hisense.com

Abstract:

This paper introduces the application of regression analysis in multiple fields, with a focus on conducting linear regression analysis on five case studies. First, the study on the Consumer Price Index (CPI) and Retail Price Index (RPI) reveals a positive correlation between the two, indicating that fluctuations in retail prices significantly affect consumer prices. Second, an analysis was conducted on the impact of academic pressure on students' mental health, finding that academic pressure significantly affects emotional and psychological states. Third, the paper compares the animation markets in Japan and China, showing differences in market development models and growth trends between the two countries. Next, the paper explores the relationship between socioeconomic status and family educational expectations, indicating that family income and parents' education levels significantly influence educational expectations. Finally, the paper examines the relationship between age and sleep quality and duration, finding that both sleep quality and duration decrease as age increases. This paper uses linear regression models and related statistical methods, providing data support for further research in related fields.

Keywords: Linear Regression Model, Consumer Price Index, Data Analysis, Economic Analysis.

1.1 Introduction

With the advent of the big data era, data analysis has become increasingly widely used in various fields. The linear regression model, one of the most fundamental and commonly used tools in statistics, has been applied across many disciplines, such as economics, education, and social sciences. It helps researchers analyze the relationship between independent and dependent variables, providing strong support for policy formulation and decision-making [1].

The importance of the research topic is reflected in its broad application across multiple fields and its profound impact on real-world social issues. First, the relationship between the Consumer Price Index (CPI) and the Retail Price Index (RPI) directly influences inflation forecasting and control. Particularly in times of economic fluctuations, this data can provide

ISSN 2959-6157

strong support for governments to formulate reasonable price policies. Second, as students' mental health issues become more prominent, the impact of academic pressure on psychological and physical health has become an important research topic in the fields of education and psychology. Understanding these impacts can help schools and policymakers alleviate student stress and promote their overall development [2]. Additionally, the booming animation market in the cultural industry has brought great growth potential to the economy. A comparative study of the animation markets in Japan and China helps to understand the development models of cultural industries in the context of globalization. At the same time, the relationship between socioeconomic status and family educational expectations reflects the unequal distribution of social resources. Studying this topic not only has academic value but also provides reference for the formulation of educational equity policies. Finally, research on the relationship between age, sleep quality, and sleep duration is closely related to the health issues of an aging society, providing scientific evidence for the formulation of public health policies [3]. Therefore, these research topics not only address important real-world problems but also have significant social implications in areas such as policy, education, and health management.

This paper is divided into four main parts. First, it introduces the basic theory and methods of the linear regression model. Second, the author conducts linear regression analysis on five case studies, discussing the relationship between the Consumer Price Index and the Retail Price Index, and revealing its importance in macroeconomic forecasting through regression analysis. Then, the paper studies the impact of academic pressure on students' mental health, analyzing the effects of pressure on emotions and willpower. Next, it compares the development of the animation markets in Japan and China, revealing differences in market scale and growth models. Subsequently, it explores the influence of socioeconomic status on family educational expectations, particularly the roles of family income and parents' education levels in shaping these expectations. At the same time, the paper analyzes the impact of age on sleep quality and duration, summarizing the challenges aging presents to health. Finally, the paper draws conclusions from the various studies and looks forward to future research directions.

2.Method and model

2.1 Linear Regression Model

Linear regression is a statistical method used to study the linear relationship between one or more independent variables and a dependent variable. The simplest form of linear regression is called simple linear regression, which describes the relationship between one independent variable and one dependent variable. The regression equation can be written as [4]

$$Y = \beta_0 + \beta_1 X + \epsilon \tag{1}$$

where, y is the dependent variable, which author wants to predict or explain. x is the independent variable, which affects the dependent variable. β_0 is the intercept, representing the value of y when x equals 0. β_1 is the slope, indicating how much y changes for each unit change in x. ε is the error term, which represents the part of y that cannot be explained by the model.

When there are multiple independent variables, the model can be extended to multiple linear regression, which can be written as:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n + \epsilon \tag{2}$$

The primary goal of a linear regression model is to find the best-fitting coefficients for β_0 and β_1 (or $\beta_1, \beta_2, ..., \beta_n$ in the case of multiple regression) to maximize the explanatory power of the independent variables over the dependent variable.

2.2 Least Squares Method

The goal of the least squares method is to minimize the sum of squared errors (SSE) [5]

$$SSE = \sum \left(y_i - \hat{y}_i \right)^2 = \sum \left(y_i - \left(\beta_0 + \beta_1 x_i \right) \right)^2 \tag{3}$$

By solving for the intercept (β_0) and slope (β_1), author can obtain the parameter estimates that minimize SSE. The optimal values are derived by differentiating SSE and setting the derivatives equal to zero. The formula for calculating the slope (β_1) is:

$$\beta_i = \frac{\sum \left(\left(x_i - \bar{x} \right) \left(y_i - \bar{y}_i \right) \right)}{\sum \left(x_i - \bar{x} \right)^2}, \tag{4}$$

while the formula for calculating the intercept (β_0) is:

$$\beta_0 = \bar{y} - \bar{\beta}_1 x. \tag{5}$$

Here, $\bar{x}$ and $\bar{y}$ are the means of the independent and dependent variables, respectively.

For multiple linear regression, the solution of the least squares method can be expressed in matrix form. Assume X is the design matrix containing all independent variables, and y is the target vector. The parameter estimates for the regression model can be written as:

$$\hat{\beta} = \left(X ? X \right)^{-1} X^T y \tag{6}$$

This formula allows people to compute regression coefficients and construct regression model.

2.3 Model Evaluation of Linear Regression

About the R^2 Coefficient of Determination, R^2 is used to measure how well the model explains the variation in the dependent variable. It ranges from 0 to 1, with values closer to 1 indicating a better fit of the model. The formula for R^2 is:

$$R^2 = 1 - \left(\frac{SSE}{SST}\right) \tag{7}$$

where SSE is the sum of squared errors, and SST is the total sum of squares, representing the total variation in the dependent variable [6]. Residual Analysis is also used. Residual analysis is a critical step in evaluating the model. By analyzing the residuals, author can determine whether the assumptions of linear regression are met, such as normality and homoscedasticity of the error terms.
3.Applications

3.1 Analysis of Consumer Price Index and the Retail Price Index

Here is an analysis of the Consumer Price Index and Retail Price Index - Based on Simple-Linear Regression Analysis. The author uses a simple linear regression model:

$$Y = \alpha + \beta x \tag{8}$$

where Y represents the consumer price index, X represents the retail price index, α is the constant term, and β is the regression coefficient, indicating the average change in the consumer price index for each unit increase in the retail price index. In this paper, author uses the annual CPI and RPI data from 2005 to 2017 to conduct regression analysis. After that, author did the Model Test. The goodness-of-fit test was performed on the model, and through calculations using Eviews software, the coefficient of determination $R^2 = 0.960$ indicates that the regression model can explain 96% of the data variance, showing that the model fits excellently.
The significance test of the model showed that the t-test and F-test results were significant. The F-value was 266.4 , and the P-value was 0.000, indicating that the overall significance of the model passed, and there is a significant linear relationship between the variables. A multicollinearity and autocorrelation test were also conducted on the model, and the Durbin-Watson statistic was 1.77, which is close to 2, indicating that the model does not have significant autocorrelation.
About the results, the final regression equation obtained is:

$$Y = 11.604 + 0.893X \tag{9}$$

This indicates that for each unit increase in the retail price index, the consumer price index increases by 0.893 units. There is a significant positive correlation between the consumer price index and the retail price index, indicating that changes in the retail price index significantly impact the consumer price index. Based on this conclusion, this paper suggests that the government should optimize the consumption environment and encourage consumer spending when formulating macroeconomic policies to promote economic development.

3.2 The impact of academic stress on students‘ mental and physical health

In this paper, author sets learning pressure as the independent variable, and various factors of physical and mental health (such as sensory thinking, emotional response, willpower, physical symptoms, etc.) as the dependent variables, constructing a simple linear regression model:

$$Y = b_0 + b_1 X \tag{10}$$

where Y represents the health factor score, X represents the learning pressure score, b_0 is the constant term, and b_1 is the regression coefficient, indicating the average change in the health factor score for each unit increase in learning pressure [7].
Reliability and validity tests were conducted using Cronbach's alpha coefficient method, and the α coefficient was 0.9755, indicating that the questionnaire has high reliability and validity.
About the results, the linear equation obtained is:

$$Y = 1.7689 + 0.2828X \tag{11}$$

The regression analysis results show that learning pressure has a significant impact on students' perceptions, emotions, and willpower, especially on psychological issues such as weakened willpower, memory loss, and emotional anxiety. Increased learning pressure also leads to fatigue and emotional depression among students, although its impact on physical symptoms like palpitations and dizziness is relatively minor. It is recommended that schools and parents pay more attention to students' mental health by reducing academic burdens and providing psychological counseling to help students cope with learning pressure.

3.3 Comparison of the development of the Chinese and Japanese animation industries.

A simple linear regression model was used to predict the market size of the animation industry in China and Japan.

The model is of the form:

$$Y = \beta_0 + \beta_1 X \qquad (12)$$

where Y represents the market size, X represents the year, β_0 is the intercept, and β_1 is the regression coefficient. A Grey Dynamic Model GM(1,1) was also used for predicting time-series data [8]. The model is of the form:

$$\frac{dX^{(1)}}{dt} + aX^{(1)} = u \qquad (13)$$

Where a is the grey parameter, and u is the constant term. The model was built and predictions were made based on cumulative generation of time series data.

About the Model Test, author using the GM(1,1) model to predict the market size of China's animation industry, the posterior error ratio c = 0.0818 and the probability of small errors p = 1.0000 indicate that the model has high prediction accuracy.

About the results, the simple linear regression equation for China's animation market is obtained as:

$$Y = 11.43 + 34.61X \qquad (14)$$

The coefficient of determination $R^2 = 0.94$ indicates a significant linear relationship between market size and the year. Based on the regression equation, the market size of China's animation industry is predicted to be 350.5 billion yuan in 2013, and it will increase by 55.37 billion yuan annually. The simple linear regression equation for Japan's animation market is:

$$Y = 1147.93 + 107.07X \qquad (15)$$

According to the simple linear regression model, the market size of Japan's animation industry is predicted to increase by 3107.07 billion yen annually. Using this regression model, the market size of Japan's animation industry in 2013 is predicted to be 2488.63 billion yen.

The Grey Dynamic Model for China's animation market is:

$$x(t+1) = 191.44e^{0.32386t} - 134.14 \qquad (16)$$

Based on the Grey Dynamic Model, the predicted market sizes of China's animation industry for 2013-2015 are 434.65 billion, 599.44 billion, and 826.92 billion yuan, respectively. The Grey Dynamic Model for Japan's animation market is:

$$x(t+1) = 10851.18e^{0.02078t} - 106618.18 \qquad (17)$$

According to the Grey Dynamic Model, the predicted market sizes of Japan's animation industry for 2013-2015 are 2476.42 billion, 2528.42 billion, and 2581.51 billion yen, respectively. The positive exponent of e (0.02078) indicates that Japan's animation industry is in a stable growth period, which aligns with the actual development over the past several years.

3.4 Socioeconomic status and family educational expectations.

This study uses a linear regression model:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n + \epsilon \qquad (18)$$

Where Y is the number of years of education expected by the family for their children as the highest level of education expected by the family, converted into the number of years (e.g., a bachelor's degree corresponds to 16 years). $X_1, X_2, \cdots, X_n$ are the independent variables, including parents' educational level, family income, gender of the child, household registration, ethnicity, number of siblings. β_0 is the constant term, and $\beta_1, \beta_2, \cdots, \beta_n$ are the regression coefficients, indicating the influence of each independent variable on the educational expectation. ϵ is the error term.

A binary Logit model is also used:

$$P(Y=1) = \frac{1}{1+e^{-(\beta_0+\beta_1 X_1+\beta_2 X_2+\cdots+\beta_n X_n)}} \qquad (19)$$

where $P(Y=1)$ indicates the probability that the family expects their child to receive higher education. $X_1, X_2, \cdots, X_n$ are the independent variables, including parents' educational level, family income, gender, household registration, and ethnicity.

About the results, the linear regression model is used to analyze the years of education expected by the family for their children, as the expected length of time their children would be educated (in years). The model mainly examines the influence of different factors, such as family socioeconomic status and cultural beliefs, on educational expectations:

In terms of parents' educational level, the results show a significant impact on family educational expectations. Compared to families where parents' education level is at elementary school or below, families where the parents have a junior high school education expect their children to receive about 3.2 more years of education; families where the parents have a high school education expect their children to receive 4.7 more years of education. This shows that the higher the parents' education level, the higher their expectations for their children's education. In terms of family income, there is a positive relationship between family income and educational expectations. For every unit increase in family income (in the natural logarithm), the family's expected years of education for their children increase by 0.1 years. This means that families with higher incomes have higher educational expectations for their children, possibly because they can provide more educational resources and opportunities. Regarding gender differences, although families slightly favor boys in terms of expected years of education (regression coefficient of 0.2), the difference is not significant. This suggests that, considering other variables, gender has a weak influence on educational expectations, especially in one-child families, where gender differences are less apparent. In terms of the number of siblings, an increase in the number of children in a family significantly lowers the educational expectations for each child. For every additional sibling, the expected years of education decrease by 1.1 years.

The binary Logit model analyzes whether the family expects their child to pursue higher education (whether they expect their child to attend college or above). The dependent variable is binary, where 0 means no expectation and 1 means the expectation exists. The conclusions are as follows:

In terms of parents' educational level, the effect of parental education on higher education expectations is very significant. Compared to parents with an education level of elementary school or below, parents with a junior high school education are 2.1 times more likely to expect their children to receive higher education; parents with a high school education are 3.5 times more likely to have this expectation. This indicates that the more educated the parents, the more they tend to expect their children to receive higher education. Family income also significantly influences expectations for higher education. For every unit increase in family income, the probability of expecting the child to receive higher education increases by 7%. This implies that wealthier families are more capable of supporting their children to attend college, thus having higher expectations for them. In terms of gender differences, while the model results show that boys are more likely to be expected to receive higher education, the gender difference is not significant. This suggests that although there is a "son preference" in families with multiple children when it comes to higher education expectations, it is not a universally strong phenomenon. The number of siblings significantly reduces the expectation for higher education. For every additional sibling, the probability of the child receiving higher education expectations decreases by 15%. This demonstrates that in larger families, the "resource dilution effect" causes parents' expectations for their children's higher education to decline, especially for girls.

3.5 The impact of age on sleep duration and quality.

The author randomly selected 20 data points from the referenced data to conduct a detailed analysis of the relationships between age, sleep quality, and sleep duration, see Table 1.

Table 1. The age and Sleep Quality and Duration

Age	Sleep Quality (score)	Sleep Duration (hours)	Age	Sleep Quality (score)	Sleep Duration (hours)
45.0	6.0	6.5	31.0	8.0	7.0
44.0	5.0	5.75	50.0	5.0	6.0
32.0	9.0	9.0	25.0	8.0	7.5
47.0	5.0	4.5	33.0	9.0	8.5
33.0	8.0	7.5	24.0	9.0	8.0
41.0	6.0	6.75	50.0	4.0	5.5
23.0	9.0	8.5	47.0	5.0	5.75
22.0	8.0	7.5	36.0	8.0	8.0
30.0	8.0	8.0	45.0	6.0	7.0
40.0	9.0	7.5	46.0	4.0	5.5

First, the author conducted a simple linear regression analysis between age and sleep duration, using the following simple linear regression model:

$$Y = \beta_0 + \beta_1 X + \epsilon \tag{20}$$

Where Y represents sleep duration (hours), X represents age (years), β? represents the intercept (the predicted sleep duration when age is 0), β? represents the slope (the change in sleep duration with each additional year of age), and ϵ represents the error term (the difference between actual and predicted values).

Using the least squares method to fit the data, the following regression equation was obtained:

$$Y = -0.1631X + 13.0170 \tag{21}$$

The fitted plot based on the data is shown in Fig. 1. Fig.1 indicates that for each additional year of age, sleep quality decreases by approximately 0.1631 points. The intercept suggests that, theoretically, when the age is 0, the predicted sleep quality score would be 13.0170 points. The correlation coefficient between age and sleep quality is $R = -0.847$, indicating a strong negative correlation between the two variables. As age increases, sleep quality significantly declines.

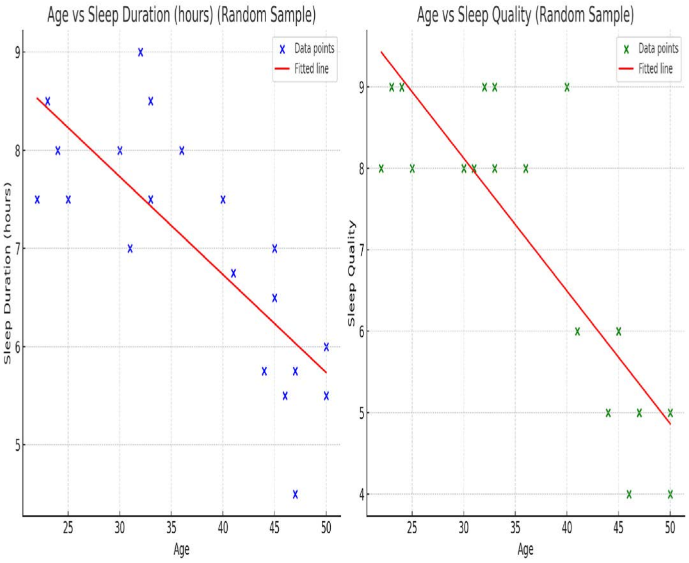

Fig. 1 The fitted plot based on the age and sleep duration

Next, the author conducted a simple linear regression analysis between age and sleep quality, using the following simple linear regression model:

$$Y = \beta^0 + \beta^1 X + \epsilon \tag{22}$$

Where Y represents sleep quality (score), X represents age (years), β_0 represents the intercept (the predicted sleep quality score when age is 0), and β_1 represents the slope (the change in sleep quality score with each additional year of age).

Using the least squares method to fit the data, the following regression equation was obtained:

$$Y = -0.1631X + 13.0170 \tag{23}$$

The fitted plot based on the data is shown in Fig. 2. Fig.2 suggests that with each additional year of age, sleep quality decreases by approximately 0.1631 points. The intercept indicates that, theoretically, when age is 0, the predicted sleep quality score would be 13.0170 points. The correlation coefficient between age and sleep quality is $R = -0.847$, demonstrating a strong negative correlation between the two variables. As age increases, sleep quality significantly declines.

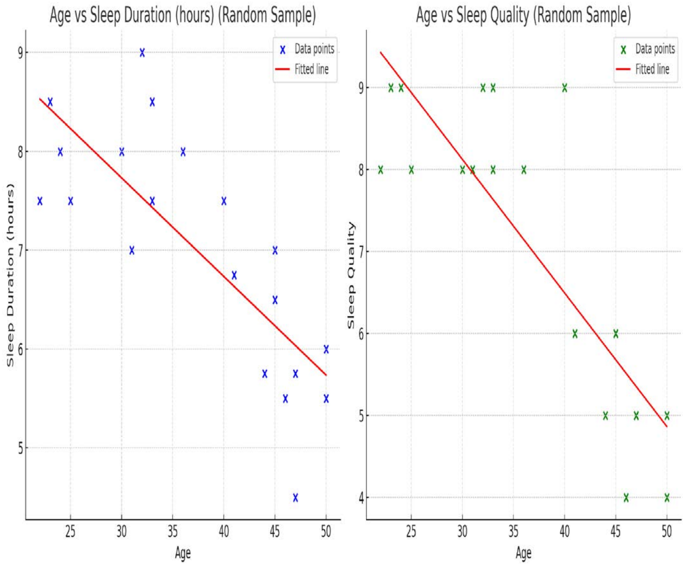

Fig. 2 The fitted plot based on the age and sleep quality

4. Conclusion

This paper conducts regression analysis on the relationship between the Consumer Price Index and Retail Price Index, the impact of academic pressure on students' mental health, the comparison between the animation markets in Japan and China, the relationship between socioeconomic status and family educational expectations, and the correlation between age, sleep quality, and duration. The research results indicate that CPI and RPI have a significant positive correlation, where changes in retail prices affect consumer prices. Academic pressure negatively impacts students' emotions and mental health, particularly leading to emotional anxiety and weakened willpower. There are significant differences in the development models of the animation markets between Japan and China. Socioeconomic status significantly influences family educational expectations, and there is a negative correlation between age, sleep quality, and duration. These findings provide references for theoretical research and policy formulation in related fields. However, the study has some limitations, such as only using a simple linear regression model and not fully considering more complex variable relationships. Future research can introduce nonlinear models or multiple regression models to improve analysis accuracy. Additionally, the sample size should be expanded to enhance the applicability of the research results. The linear regression model will continue to play an important role in various fields, especially as a fundamental tool for analyzing the relationships between variables. However, with the increasing complexity of data, researchers will need to incorporate more advanced regression techniques and methods to address more complex data structures and provide more accurate predictions and decision support.

References

[1] Han, M. Analysis of the Consumer Price Index and Retail Price Index—Based on Simple Linear Regression Analysis. Modern Business, 2020 , (17), 12-13.

[2] Liu, B., Zhang, Y., & Li, J. Socioeconomic Status, Cultural Concepts, and Family Educational Expectations. Youth Studies, 2019 , (3), 22-35.

[3] Zheng, L., Wan, L., & Li, Z. Linear Regression Analysis of the Impact of Learning Pressure on Students' Mental and Physical Health. Chinese School Health, 2001, 22(3), 224-226.

[4] Zhu, Y. Forecast and Comparative Study of the Development of the Animation Industry in China and Japan. Statistics and Decision, 2014 , (9), 117-120.

[5] Aksoy, H. (2023). Health and Sleep Statistics [Data set]. Kaggle [https://www.kaggle.com/datasets/hanaksoy/health-and-sleep-statistics].

[6] Suzuki, Frank S., et al. Multivariate linear regression analysis to evaluate multiple-set performance in active and inactive individuals. Motriz: Revista de Educação Física, 2019, 25(2): e101919.

[7] Altman, Naomi and Martin Krzywinski. Simple Linear Regression. Nature Methods, 2015, 12(11): 999-1000.

[8] Nikolaos Pandis. Multiple linear regression analysis. American journal of orthodontics and dentofacial orthopedics 2016, 149 (3): 431-434.