**MODELING FARMERS’ INCOME : KEY ECONOMIC DETERMINANTS**

**DATA ANALYSIS USING STATISTICAL PACKAGES**

**SUBMITTED BY,**

**MOHAMED DILROSE PM**

**23375040**

**MSC STATISTICS**

**INTRODUCTION**

Agriculture plays a crucial role in the economic stability of many countries, particularly in regions where farming is the primary source of livelihood. In West Bengal, India, rice farming is a major agricultural activity, and farmers' annual income is influenced by various economic, environmental, and policy-driven factors. Understanding these influences can help policymakers and stakeholders develop effective strategies to improve farmers' financial stability. This study aims to predict farmers' annual income using a linear regression model based on key determinants such as land size, farming experience, irrigation type, crop loss, weather impact, fertilizer usage, government subsidies, loan amount taken, interest rate of loans, and market price per kilogram of produce. By identifying the most significant predictors of income, this study seeks to provide insights into financial planning and agricultural development policies.

This study was conducted using SPSS software, which facilitated data processing, statistical analysis and model building to derive meaningful insights from the dataset.

**OBJECTIVES**

1. To analyze the relationship between various agricultural and economic factors with farmers' annual income.
2. To develop a linear regression model that accurately predicts farmers' earnings based on selected economic and agricultural variables.
3. To assess the impact of financial support mechanisms, such as subsidies and loans, on farmers' financial well-being.
4. To provide data-driven recommendations for policymakers to improve farmers' income stability.

**METHODOLOGY AND DATA**

1. **DATA COLLECTION**

This study utilizes secondary data obtained from **Kaggle**, containing records of **rice** **farmers** from **West Bengal** who took loan and received subsidies from the government for the year **2022**. The dataset comprises a combination of categorical and numerical variables that capture various aspects of farmers' financial and agricultural conditions. The variables include **“Name”, “Age”, “Gender”, “Annual earnings”, “Farming experience”, “Crop loss percentage”, “Land size in acres”, “Loan amount”, “Interest rate”, “Government subsidy” and “Market price per kg”, “Fertilizer usage”, “Irrigation type”, “Weather impact” .**

1. **RESEARCH DESIGN**

This study employs a **quantitative research approach** using a **predictive modeling framework** to analyze the impact of financial support mechanisms, such as government subsidies and loans, on farmers' annual earnings.

**3. DATA PREPROCESSING**

* **Handling Missing Values:** Missing values are identified and treated using mean computation for numerical variables.
* **Converting Categorical Variables:** categorical variables are converted into numerical representations to be included as predictors
* **Identifying and Correcting Inconsistent Data:** The logical inconsistencies in the data were identified and corrected if necessary.
* **Transforming Skewed Variables:** Transforming skewed data improves the accuracy of regression modeling by meeting normality assumptions.
* **Outlier Detection & Treatment:** Outliers are identified using **Z-score analysis** and **Cook’s Distance method** and either transformed or removed if necessary.

**4. EXPLORATORY DATA ANALYSIS (EDA)**

* **Summary statistics** (mean,standard deviation) are computed.
* **Correlation analysis** is performed to check multicollinearity between predictors.
* **Histograms** are used to visualize variable distributions and detect anomalies.
* **Scatter plots** are created to explore the relationships between variables.

**5. MODEL DEVELOPMENT**

* A **multiple linear regression model** was constructed to assess the influence of independent variables, including land size, farming experience, loan amount, interest rate, government subsidy, and market price per kilogram, on farmers' earnings.
* The dataset was split into **training (90%) and test (10%)** sets to validate model performance.
* The regression model was built using the training dataset, and the estimated coefficients were used to **predict farmers' earnings in the test dataset**.

**6. EVALUATION OF MODEL PERFORMANCE**

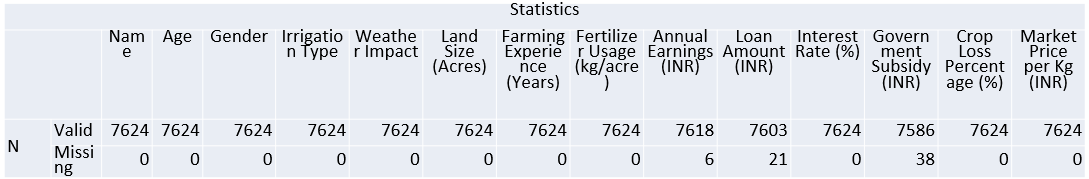
* **Residual analysis** was conducted to check the normality, homoscedasticity, and independence of residuals.
* The **histogram and Q-Q plot** confirmed that the residuals were approximately normally distributed.
* The model was applied to the test dataset, and the predicted values were compared with actual values to assess its **predictive accuracy**.

**7. FINDINGS AND POLICY RECOMMENDATIONS**

* **Data-driven recommendations** will be provided for policymakers to improve farmers' income stability by optimizing subsidy allocations, refining loan policies, and identifying alternative measures to enhance economic resilience.

**DATA PREPROCESSING**

* **Handling missing values**



* **Annual Earnings** has 7618 valid values and 6 values were missing.
* **Loan Amount** has 7603 valid values and 21 values were missing.
* **GovernmentSubsidy** has 7586 values and 38 values were missing

**Since these are numerical variables, the missing values can be replaced using the mean**

* **Converting Categorical Variables**

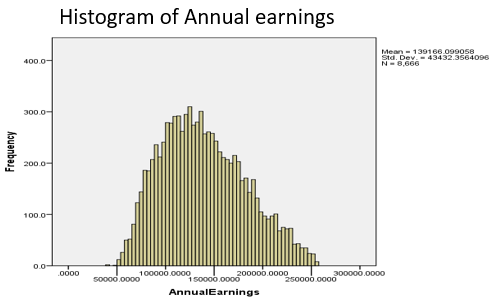
In this study, two categorical variables—**Weather Impact** and **Irrigation Type**—are included in the dataset. Since regression analysis requires numerical inputs, these categorical variables were converted into numerical form using **dummy coding** in SPSS.

* **Identifying and Correcting Inconsistent Data**

A logical inconsistency was identified in **Farming Experience**, where in some cases, the reported farming experience exceeded the farmer’s age, which is not possible.

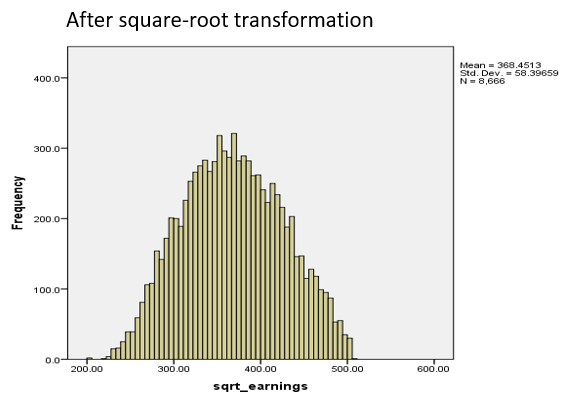
The cases where **Farming Experience** is greater than **Age** were filtered out using “Select cases” function and keep only valid records.

* **Checking for Normality**



This histogram of **AnnualEarnings** looks roughly normal,

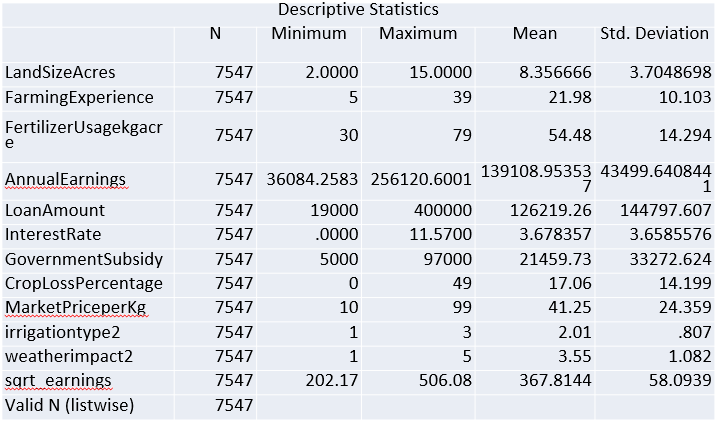
it appears slightly **right-skewed** (longer tail on the right)



Thehistogram of **square-root** transformed **AnnualEarnings** appears to be normal.

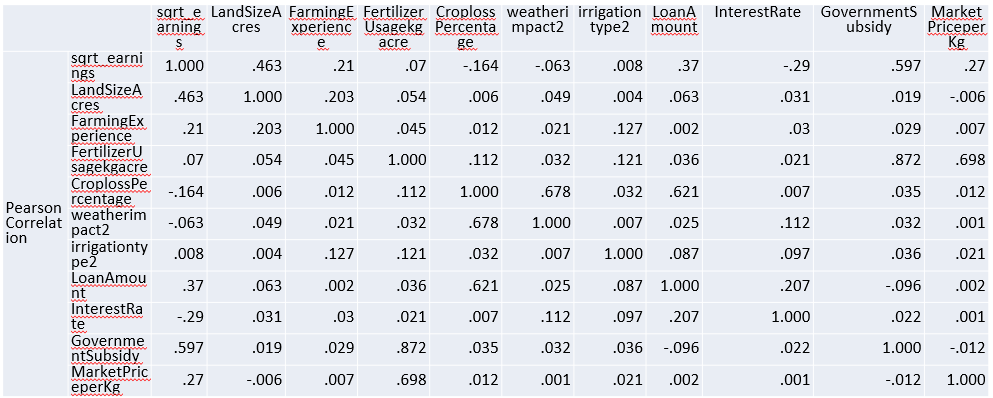
The regression model will be developed using the square-root transformed **Annual Earnings** variable to improve normality, reduce skewness, and enhance the reliability of statistical inferences.

* **Descriptive statistics**

****

**DATA ANALYSIS AND RESULTS**

* **Correlation Table**



**Correlation with sqrt\_earnings (Target Variable)**

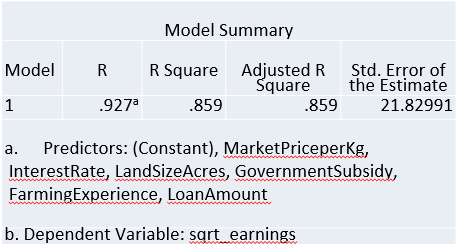
* **GovernmentSubsidy (0.597):** The highest positive correlation, indicating that government subsidies strongly influence earnings.
* **LandSizeAcres (0.463):** A moderate positive correlation, meaning that larger land sizes are associated with higher earnings.
* **LoanAmount (0.37):** A moderate positive correlation, suggesting that higher loan amounts contribute to increased earnings.
* **MarketPriceperKg (0.27):** A low positive correlation, implying that higher market prices for crops slightly increase earnings.
* **FarmingExperience (0.21):** A low positive correlation, meaning experience has an impact on earnings.
* **InterestRate (-0.29):** A negative correlation, indicating that higher interest rates are linked to lower earnings, possibly due to increased loan repayment burdens.
* **Irrigation Type (0.008)** : Almost no correlation with earnings.
* **Weather Impact (-0.063)** : Very weak correlation.
* **Crop Loss Percentage (-0.164)** : Weak correlation.
* **Fertilizer Usage (0.07)** : Very Weak correlation.

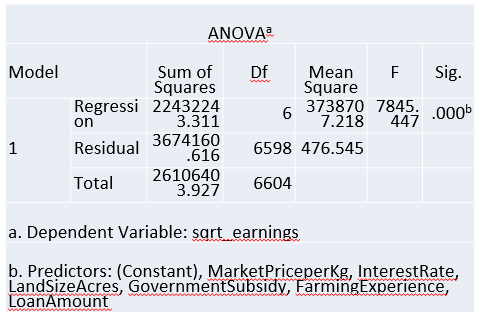
These variables that exhibit a very weak correlation with the dependent variable (**sqrt\_earnings**) are not considered for model building.

* **Data splitting for model development**

To ensure the robustness and generalizability of the regression model, the dataset was **split into training and test sets**. A **90:10 split ratio** was chosen, where **90% of the data was used for training** the model, and the remaining **10% was reserved for testing**.

The training dataset was used to build and optimize the regression model by identifying relationships between the independent variables and the **square-root** **transformed** **Annual Earnings** variable. The test dataset was kept separate to evaluate the model’s performance on unseen data, preventing overfitting and ensuring the model’s predictive power in real-world scenarios.

**Model development**



* **R = 0.927** → Indicates a high correlation between the predictors and the dependent variable (sqrt\_earnings).
* **R² = 0.859** → Suggests that **85.9% of the variance** in sqrt\_earnings is explained by the independent variables, which is a **strong model fit**.
* **Adjusted R² = 0.859** → Since it's nearly the same as R², it means that adding predictors hasn't introduced unnecessary complexity.
* **Standard Error of the Estimate (21.82991)** → Measures the typical deviation of observed values from the predicted values.
* **F-Statistic = 7845.44**
* **Significance Level (Sig.) = 0.000**
* The **p-value (< 0.05)** confirms that the regression model is **highly significant**, meaning that the predictor variables collectively have a strong relationship with **sqrt\_earnings.**

**Outlier detection and removal**

Outliers are data points that deviate significantly from the pattern of the rest of the data. They can distort statistical analyses, affect regression coefficients, and reduce the accuracy of predictive models. In regression analysis, two effective methods for detecting and removing outliers are **Standardized Residuals** and **Cook’s Distance**.

A **residual** is the difference between the observed and predicted values of the dependent variable. Standardized residuals (ZRESID) are residuals that have been transformed into a standard normal distribution with a mean of 0 and a standard deviation of 1.

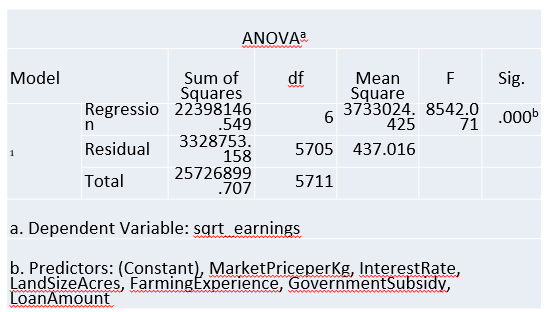
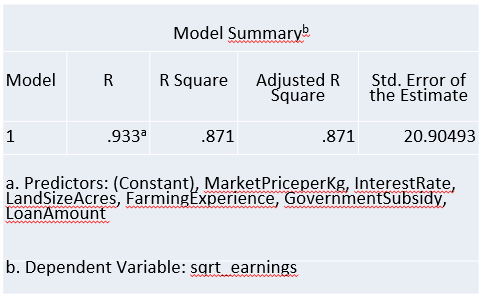
In general:

* If **|Standardized Residual| > 3**, the observation is considered an outlier.

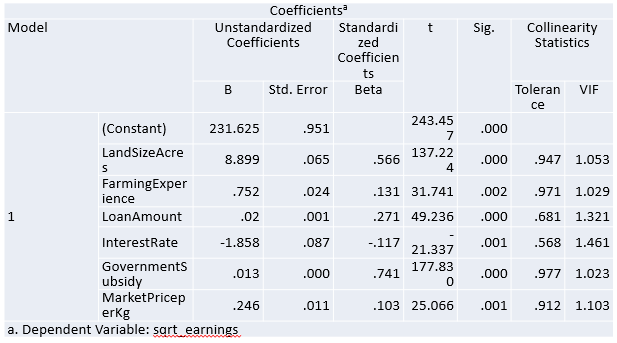
Cook’s Distance measures the influence of an individual data point on the overall regression model. It assesses how much the fitted values would change if that data point were removed.

* A high Cook’s Distance value suggests that the point is **highly influential** and could be distorting the model.
* The common rule is:
  + If **Cook’s Distance > 1**, the observation is influential and should be examined.

**After outlier removal,**



The new updated model summary shows an **R² of 0.871**, which means that **87.1% of the variance** in the square-root-transformed earnings is explained by the predictor variables. This is an improvement from the previous model (R² = 0.859). The **standard error of the estimate (20.90)** has also slightly decreased, indicating better model precision.



**Significant Predictors (p-value < 0.05):**

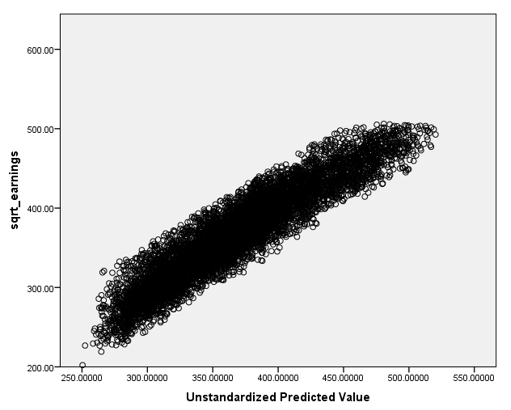
* **LandSizeAcres (p = .000)**
* **FarmingExperience (p = .002)**
* **LoanAmount (p = .000)**
* **InterestRate (p = .001)**
* **GovernmentSubsidy p = .000)**
* **MarketPriceperKg (p = .001)**.
* **Multicollinearity (VIF Analysis):**

**VIF < 5** for all variables, indicating **no serious multicollinearity issues**.

* **Regression Equation:**

**sqrt\_earnings^=231.625+8.899(LandSizeAcres)+0.752(FarmingExperience)+0.02(LoanAmount)−1.858(InterestRate)+0.013(GovernmentSubsidy)+0.246(MarketPriceperKg)**

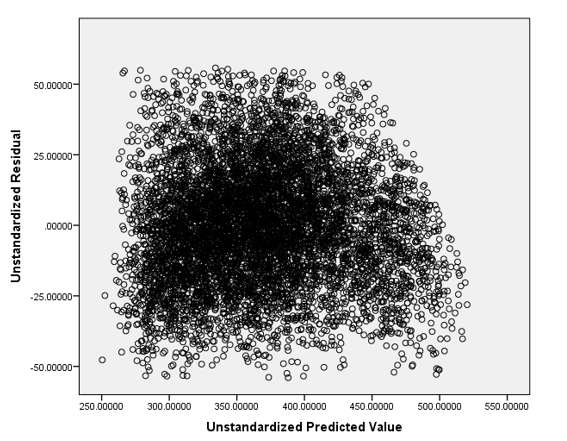
**Scatter Plot (Predicted vs. Transformed Earnings) after outlier removal**



The points remain closely clustered around a linear trend, further validating the model’s predictive strength.The refined model shows **less deviation at lower and upper earnings values**, suggesting a more **robust** prediction across different earnings levels.

The spread of points appears more controlled and compact, with fewer extreme deviations, indicating improved homoscedasticity and better model fit.

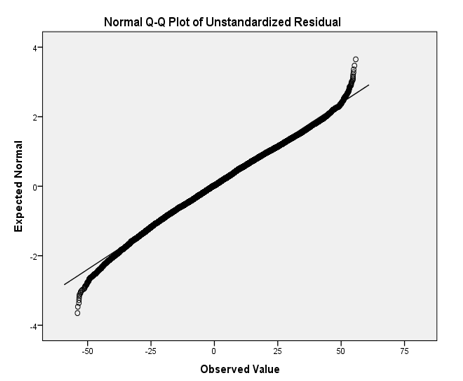
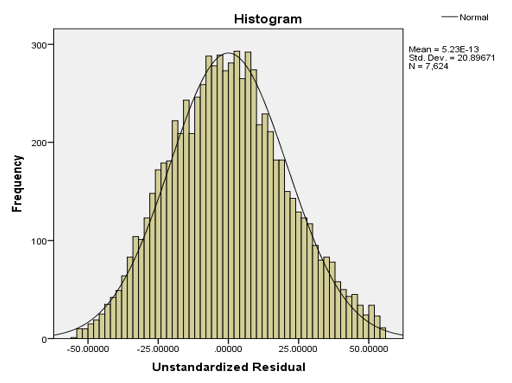
**Residual Plot**



**Random Scatter (Good Sign)**

* The points appear randomly distributed around zero without any clear pattern.
* This suggests that the model meets the **assumption of homoscedasticity** (constant variance of residuals).

**Residual normality checks**

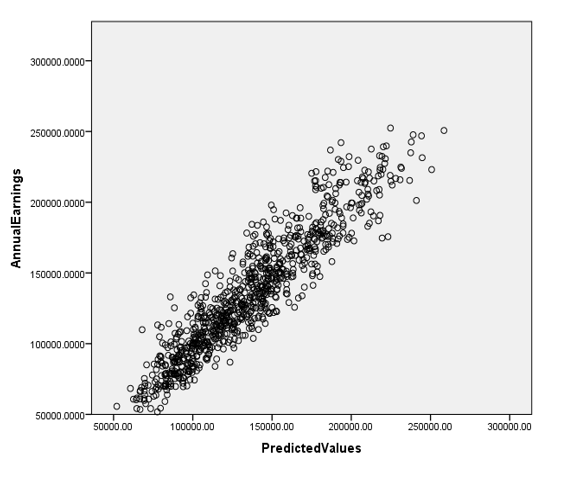


* The histogram of residuals follows a symmetric normal shape.
* The Q-Q plot shows an alignment with the normal distribution line, except for slight deviations at the tails.

**Applying the Model to the Test Data**

Once the regression model was built using the training set, the estimated regression equation was applied to the testing dataset to generate predicted values for farmers' annual earnings.

**Scatter Plot (Predicted values and Actual values) of the test data**



**Strong Positive Correlation**

* The data points form a tight cluster along a diagonal trend, suggesting that the model’s predicted values closely match the actual earnings.
* This indicates that the model performs well on unseen data, confirming its predictive capability.

**FINDINGS**

The linear regression model was developed to predict farmers' annual earnings using key independent variables such as **Land Size, Farming Experience, Loan Amount, Interest Rate, Government Subsidy, and Market Price per Kg**. The major findings from the model are:

1. **Strongest Predictors of Earnings**

* **Government Subsidy (β = 0.741, p < 0.001)**
  + This is the **most influential** variable in the model.
  + Higher government subsidies are strongly associated with increased earnings.
* **Land Size (β = 0.566, p < 0.001)**
  + Larger farm sizes **increase earnings** significantly.
  + This suggests that farmers with more land have **greater production capacity and income potential.**

**B. Moderate Predictors**

* **Loan Amount (β = 0.271, p < 0.001)**
  + Farmers who take larger loans tend to **earn more**, suggesting loans are being used effectively to boost productivity.
* **Farming Experience (β = 0.131, p = 0.002)**
  + More experienced farmers have higher earnings, though the impact is relatively small.
  + Experience helps in better decision-making, but **alone is not enough** to significantly boost earnings.
* **Market Price per Kg (β = 0.103, p < 0.001)**
  + While higher crop prices slightly increase earnings, this impact is relatively small.
  + Suggests that **production volume and input costs** might be more important than price fluctuations.

**C. Negative Impact on Earnings**

* **Interest Rate (β = -0.117, p = 0.001)**
  + Higher interest rates **reduce earnings**, as loan repayments cut into farmers' profits.
  + This suggests that **affordable credit options** are necessary to improve farmer incomes.

**SUGGESTIONS FOR POLICY AND PRACTICAL IMPLICATIONS**

Based on the findings, the following recommendations are proposed:

**1. Enhancing Government Subsidy Programs**

* + - Since **government subsidies have the highest positive impact on earnings**, authorities should **increase the accessibility and efficiency of financial aid programs**.
    - Implement **direct benefit transfers (DBT)** to ensure farmers receive subsidies **on time** and **without intermediaries**.

**2. Expanding Access to Farmland**

* + - Policies should be designed to **facilitate land consolidation** and **ease restrictions on leasing farmland**.
    - Encourage **cooperative farming models** where small farmers pool resources and share profits.

**3. Providing Affordable Credit Facilities**

* + - Since **high interest rates negatively impact earnings**, **lower interest loan programs should be promoted** for farmers.
    - **Subsidized loan schemes** or **interest waivers for small farmers** can help **reduce financial burdens**.

**4. Improving Market Access and Pricing Policies**

* + - The **relatively low impact of market prices on earnings** suggests that farmers **may not be getting fair prices**.
    - Strengthen **minimum support price (MSP) policies** to protect farmers from price volatility.
    - Develop **better storage and transportation infrastructure** to help farmers **reduce post-harvest losses** and **sell their produce at better prices**.

**5. Skill Development and Technology Adoption**

* + - Since **farming experience plays a role in earnings**, **providing technical training and workshops** on modern farming techniques can **further improve productivity**.
    - Promote the use of **technology, precision farming, and efficient irrigation** methods to **optimize land use**.

**CONCLUSION**

The regression analysis provides valuable insights into the **key determinants of farmers' earnings**. The major findings indicate that **government subsidies, land size, and access to credit have the highest impact on increasing earnings**, while **high-interest rates act as a financial burden on farmers**.

To enhance farmers' earnings, policymakers should **increase financial support, improve access to affordable credit, strengthen pricing mechanisms, and promote modern farming practices**. These steps will ensure **sustainable agricultural growth and better financial stability for farmers**.

**REFERENCES**

* **Montgomery, D. C., Peck, E. A., & Vining, G. G. (2021).** *Introduction to Linear Regression Analysis* (6th ed.). Wiley.
* **Kutner, M. H., Nachtsheim, C. J., Neter, J., & Li, W. (2005).** *Applied Linear Statistical Models* (5th Edition). McGraw-Hill.
* **Draper, N.R , & Smith H. (1998).** *Applied Regression Analysis* (3rd Edition). Wiley.
* **Food and Agriculture Organization (FAO**). (2022). *The State of Food and Agriculture 2022: Leveraging agricultural finance for sustainable development.*
* <https://www.kaggle.com/datasets/renuprasad/farmers-earnings-westbengal>
* [www.geeksforgeeks.org](http://www.geeksforgeeks.org/)