VIRGINIA COMMONWEALTH UNIVERSITY

Statistical analysis and modelling (SCMA 632)

A2: Regression - Predictive Analytics

RAKSHITHA VIGNESH SARGURUNATHAN

V01109007

Date of Submission: 23-06-2024

CONTENTS

PERFORMING MULTIPLE REGRESSION ANALYSIS ON NSSO68 DATA AND IPL BALL_BY_BALL DATA

INTRODUCTION:

This Assignment using **IPL_ball_by_ball dataset** aims to analyze the relationship between player performance, specifically runs scored, and their corresponding salaries in cricket. By leveraging data matching techniques and regression analysis, we seek to uncover the extent to which runs scored influence player salaries. The analysis employs **Ordinary Least Squares (OLS) regression** to quantify this relationship, providing insights into the significance and strength of the correlation. Understanding these dynamics is essential for team management, enabling informed decision-making in player acquisitions, contract negotiations, and performance-based incentives. The study not only highlights the impact of performance metrics on financial outcomes but also underscores the need for incorporating additional factors to capture the complexity of salary determination in professional sports.

By using **NSSO68 dataset** we explore the socio-economic factors influencing household welfare metrics, using a comprehensive dataset that includes variables such as rice quantity, pulse quantity, household size, land possession, cultivated and irrigated land, and monthly per capita expenditure (MPCE). The study aims to uncover how these factors affect key economic outcomes, providing insights into household behavior and economic status. Through descriptive statistics, exploratory data analysis (EDA), and **linear regression modeling**, we examine relationships and identify significant predictors.

ABOUT THE DATASET:

- IPL_ball_by_ball_updated till 2024.csv
- IPL SALARIES 2024.xlsx

```
Columns in Ball_by_ball Dataset:
 Index(['Match id', 'Date', 'Season', 'Batting team', 'Bowling team',
       'Innings No', 'Ball No', 'Bowler', 'Striker', 'Non Striker',
       'runs_scored', 'extras', 'type of extras', 'score', 'score/wicket',
       'wicket_confirmation', 'wicket_type', 'fielders_involved',
       'Player Out'],
      dtype='object')

Columns in Salary Dataset:
 Index(['Player', 'Salary', 'Rs', 'international', 'iconic'], dtype='object')
```

```
Shape of Ball_by_ball dataset:  (255759, 19)

Shape of salary dataset:  (166, 5)
```

- NSSO68.csv

```
[5] nss.shape

(101662, 384)

nss.columns

Index(['slno', 'grp', 'Round_Centre', 'FSU_number', 'Round', 'Schedule_Number',
       'Sample', 'Sector', 'state', 'State_Region',
       ...
       'pickle_v', 'sauce_jam_v', 'Othrprocessed_v', 'Beveragestotal_v',
       'foodtotal_v', 'foodtotal_q', 'state_1', 'Region', 'fruits_df_tt_v',
       'fv_tot'],
      dtype='object', length=384)
```

OBJECTIVES:

Objective A: Multiple Regression Analysis on NSSO68 Data

1. **Perform Multiple Regression Analysis**
   - Identify the dependent and independent variables.
   - Fit a multiple regression model using the identified variables.
   - Present the regression equation and interpret the coefficients.
2. **Carry Out Regression Diagnostics**
   - Assess multicollinearity using Variance Inflation Factor (VIF).
   - Check for homoscedasticity using residual plots and evaluate the normality of residuals using Q-Q plots. Identify any influential points or outliers using Cook's distance.
3. **Correct and Revisit the Results**
   - Address any issues found during diagnostics, such as transforming variables or removing outliers.
   - Refit the regression model after corrections.Compare the revised model's performance with the original model and Explain significant differences observed after corrections.

Objective B: Relationship Between Player's Performance and Payment in IPL Data

1. **Establish the Relationship Between Performance and Payment**
   - Define key performance metrics for players (e.g., runs scored, wickets taken, strike rate).
   - Fit a regression model to establish the relationship between performance metrics and player payments.
   - Interpret the regression coefficients and discuss the significance of the relationships.
2. **Analyze the Relationship Between Salary and Performance Over the Last Three Years**
   - Filter the data for the last three years and Perform regression analysis to understand the trend and strength of the relationship between salary and performance metrics.
   - Compare results year-over-year to identify any trends or changes in the relationship.
3. **Discuss Findings**
   - Summarize key findings from the regression analysis.Discuss any patterns or insights related to player performance and payments.
   - Provide recommendations based on the analysis (e.g., suggestions for team management, potential areas for further research).

BUSINESS SCOPE:

IPL DATA:

The business scope outlines the potential applications and benefits of the analysis for business purposes.

1. **Player Valuation**: Teams can use the insights to make informed decisions on player acquisitions and salary negotiations based on performance metrics.
2. **Performance Incentives**: Implement performance-based incentives to motivate players, aligning their personal goals with team success.
3. **Strategic Planning**: Use data-driven insights for strategic planning, such as optimizing team composition and budgeting for player salaries.

NSSO68 DATA :

The analysis provides valuable insights for businesses and policymakers:

- **Agricultural Sector**: Investment in irrigation and land management can be optimized.
- **Consumer Goods**: Understanding household size and expenditure patterns can help in targeting marketing efforts.
- **Financial Services**: Insights on household economics can aid in designing better financial products.

INTERPRETATION:

NSS068 DATA ANALYSIS:

From the regression model output:

- **Rice quantity (ricepds_q)**: Negative coefficient, indicating that as rice quantity increases, the dependent variable decreases, assuming the dependent variable is expenditure or income.
- **Pulse quantity (pulsep_q)**: Positive coefficient, suggesting a direct relationship with the dependent variable.
- **Household size (hhdsz)**: Negative coefficient, showing larger households are associated with lower per capita values of the dependent variable.
- **Land Total Possessed**: Positive impact, indicating more land is associated with higher values of the dependent variable.
- **Irrigated Land**: Positive coefficient, showing a beneficial effect.
- **MPCE_URP**: Strong positive relationship, suggesting higher expenditures are associated with higher values of the dependent variable.

The analysis indicates several significant factors impacting the dependent variable. Notably:

- Larger household size reduces per capita values.
- Land possession and irrigation have positive impacts.
- MPCE is a strong positive predictor.
- The model explains a moderate portion of the variance (R-squared on test data: 0.152).

IPL DATA ANALYSIS:

Dependent Variable : Salary

Independent Variable : Runs Scored

R-squared (R^2) Score : 0.074

The R-squared value is a measure of the proportion of variance in the dependent variable that is predictable from the independent variable. Here, an R^2 score of 0.074 indicates that approximately 7.4% of the variance in the dependent variable (salary) can be explained by the independent variable (runs scored).

1. **Positive Relationship**:
   - The positive coefficient for runs scored (0.6895) suggests that as the number of runs scored by a player increases, their salary also tends to increase. This aligns with the expectation that better performance (more runs) would lead to higher salaries.
2. **Statistical Significance**:
   - The p-value associated with runs scored is very low (0.000), indicating that the relationship between runs scored and salary is statistically significant. This means we can confidently say that there is a real association between these variables in the population from which the sample is drawn.
3. **Model Fit**:
   - The R^2 score of 0.074 is relatively low, indicating that runs scored alone do not explain much of the variability in salaries. This suggests that other factors (e.g., player experience, position, team performance, endorsements, etc.) also play significant roles in determining a player's salary.

RESULTS:

PYTHON

MODEL FITTING

```python
# Split the data into training and testing sets
from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
```

```python
[15] # Standardize the features
from sklearn.preprocessing import StandardScaler
import statsmodels.api as sm
scaler = StandardScaler()
X_train_scaled = scaler.fit_transform(X_train)
X_test_scaled = scaler.transform(X_test)
```

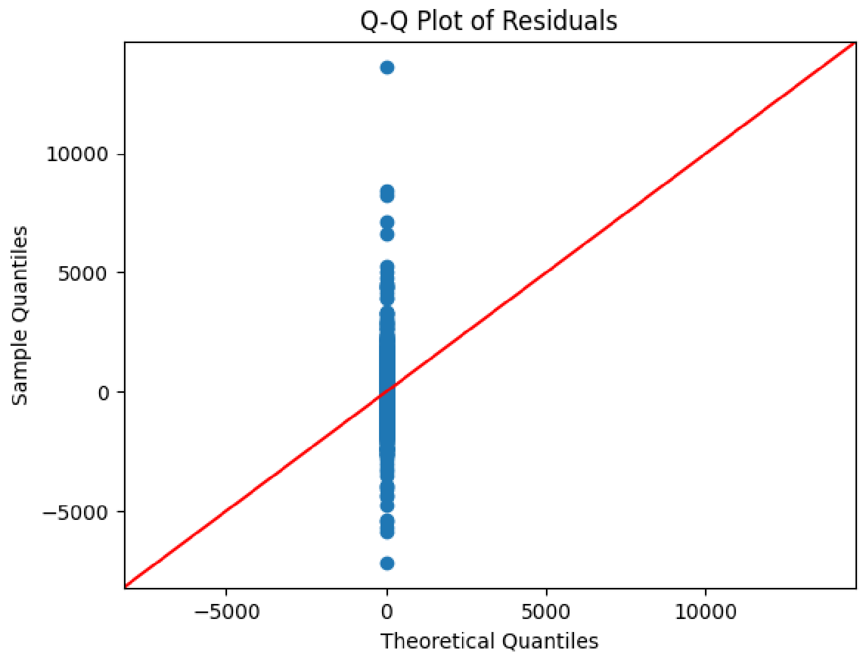
```python
[16] # Convert the scaled features back to a DataFrame and add the constant term
X_train_scaled = pd.DataFrame(X_train_scaled, columns=X_train.columns)
X_train_scaled = sm.add_constant(X_train_scaled)
X_test_scaled = pd.DataFrame(X_test_scaled, columns=X_test.columns)
X_test_scaled = sm.add_constant(X_test_scaled)
```

```python
[17] # Ensure the indices are aligned by resetting them
X_train = X_train.reset_index(drop=True)
y_train = y_train.reset_index(drop=True)
X_test = X_test.reset_index(drop=True)
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:            foodtotal_v   R-squared:                       0.358
Model:                            OLS   Adj. R-squared:                  0.358
Method:                 Least Squares   F-statistic:                     5658.
Date:                Sun, 23 Jun 2024   Prob (F-statistic):               0.00
Time:                        15:34:33   Log-Likelihood:            -5.7550e+05
No. Observations:               81329   AIC:                         1.151e+06
Df Residuals:                   81320   BIC:                         1.151e+06
Df Model:                           8
Covariance Type:            nonrobust
=================================================================================================
                                   coef    std err          t      P>|t|      [0.025      0.975]
-------------------------------------------------------------------------------------------------
const                          632.0312      1.004    629.370      0.000     630.063     633.999
ricepds_q                      -27.1985      1.027    -26.489      0.000     -29.211     -25.186
pulsep_q                        20.1771      1.008     20.017      0.000      18.201      22.153
hhdsz                          -27.9504      1.069    -26.135      0.000     -30.046     -25.854
Land_Total_possessed            15.7663      1.355     11.637      0.000      13.111      18.422
During_July_June_Cultivated     -0.2092      1.008     -0.208      0.835      -2.184       1.766
During_July_June_Irrigated       5.8450      1.339      4.365      0.000       3.221       8.469
MPCE_URP                        -8.0372      1.209     -6.648      0.000     -10.407      -5.668
MPCE_MRP                       201.2678      1.259    159.908      0.000     198.801     203.735
==============================================================================
Omnibus:                    60880.411   Durbin-Watson:                   2.003
Prob(Omnibus):                  0.000   Jarque-Bera (JB):         51737019.529
Skew:                           2.294   Prob(JB):                         0.00
Kurtosis:                     126.476   Cond. No.                         2.29
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

```python
# P-values
p_values = model.pvalues
print("P-values:\n", p_values)
```

```
P-values:
 const                             0.000000e+00
ricepds_q                        5.809005e-154
pulsep_q                          6.435995e-89
hhdsz                            6.094092e-150
Land_Total_possessed              2.836760e-31
During_July_June_Cultivated       8.354853e-01
During_July_June_Irrigated        1.269943e-05
MPCE_URP                          2.987472e-11
MPCE_MRP                          0.000000e+00
dtype: float64
```



	feature	VIF
0	const	1.000000
1	ricepds_q	1.045395
2	pulsep_q	1.007532
3	hhdsz	1.134131
4	Land_Total_possessed	1.820230
5	During_July_June_Cultivated	1.006648
6	During_July_June_Irrigated	1.777600
7	MPCE_URP	1.449267
8	MPCE_MRP	1.570890

INFERENCE:

Based on the VIF (Variance Inflation Factor) values, multicollinearity does not appear to be a significant issue since all VIF values are below 10. However, it's still important to check for the normality of the residuals. If the residuals are not normally distributed, we might need to transform some features.

```
                            OLS Regression Results
==============================================================================
Dep. Variable:            foodtotal_v   R-squared:                       0.156
Model:                            OLS   Adj. R-squared:                  0.156
Method:                 Least Squares   F-statistic:                     2140.
Date:                Sun, 23 Jun 2024   Prob (F-statistic):               0.00
Time:                        15:34:54   Log-Likelihood:            -5.8662e+05
No. Observations:               81329   AIC:                         1.173e+06
Df Residuals:                   81321   BIC:                         1.173e+06
Df Model:                           7
Covariance Type:            nonrobust
=================================================================================================
                                  coef    std err          t      P>|t|      [0.025      0.975]
-------------------------------------------------------------------------------------------------
const                         632.0312      1.151    548.957      0.000     629.775     634.288
ricepds_q                     -50.3737      1.165    -43.225      0.000     -52.658     -48.090
pulsep_q                       27.4328      1.154     23.762      0.000      25.170      29.696
hhdsz                         -70.0767      1.188    -58.970      0.000     -72.406     -67.748
Land_Total_possessed           14.2670      1.553      9.185      0.000      11.223      17.311
During_July_June_Cultivated    -0.0371      1.155     -0.032      0.974      -2.301       2.227
During_July_June_Irrigated     18.5628      1.532     12.114      0.000      15.559      21.566
MPCE_URP                       97.0717      1.163     83.448      0.000      94.792      99.352
==============================================================================
Omnibus:                    71685.972   Durbin-Watson:                   2.002
Prob(Omnibus):                  0.000   Jarque-Bera (JB):         65397534.791
Skew:                           3.118   Prob(JB):                         0.00
Kurtosis:                     141.780   Cond. No.                         2.29
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
R-squared on test data: 0.15242530917597852
```

```python
from sklearn.metrics import mean_squared_error
from sklearn.linear_model import LinearRegression
from sklearn.metrics import r2_score, mean_absolute_percentage_error

X = df_merged[['runs_scored']] # Independent variable(s)
y = df_merged['Rs'] # Dependent variable
# Split the data into training and test sets (80% for training, 20% for testing)
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
# Create a LinearRegression model
model = LinearRegression()
```

[53]
```python
# Fit the model on the training data
model.fit(X_train, y_train)
```

```
▾ LinearRegression
LinearRegression()
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                     Rs   R-squared:                       0.080
Model:                            OLS   Adj. R-squared:                  0.075
Method:                 Least Squares   F-statistic:                     15.83
Date:                Sun, 23 Jun 2024   Prob (F-statistic):           0.000100
Time:                        15:42:43   Log-Likelihood:                -1379.8
No. Observations:                 183   AIC:                             2764.
Df Residuals:                     181   BIC:                             2770.
Df Model:                           1
Covariance Type:            nonrobust
===============================================================================
                  coef    std err          t      P>|t|      [0.025      0.975]
-------------------------------------------------------------------------------
const         430.8473     46.111      9.344      0.000     339.864     521.831
runs_scored     0.6895      0.173      3.979      0.000       0.348       1.031
==============================================================================
Omnibus:                       15.690   Durbin-Watson:                   2.100
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               18.057
Skew:                           0.764   Prob(JB):                     0.000120
Kurtosis:                       2.823   Cond. No.                         363.
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

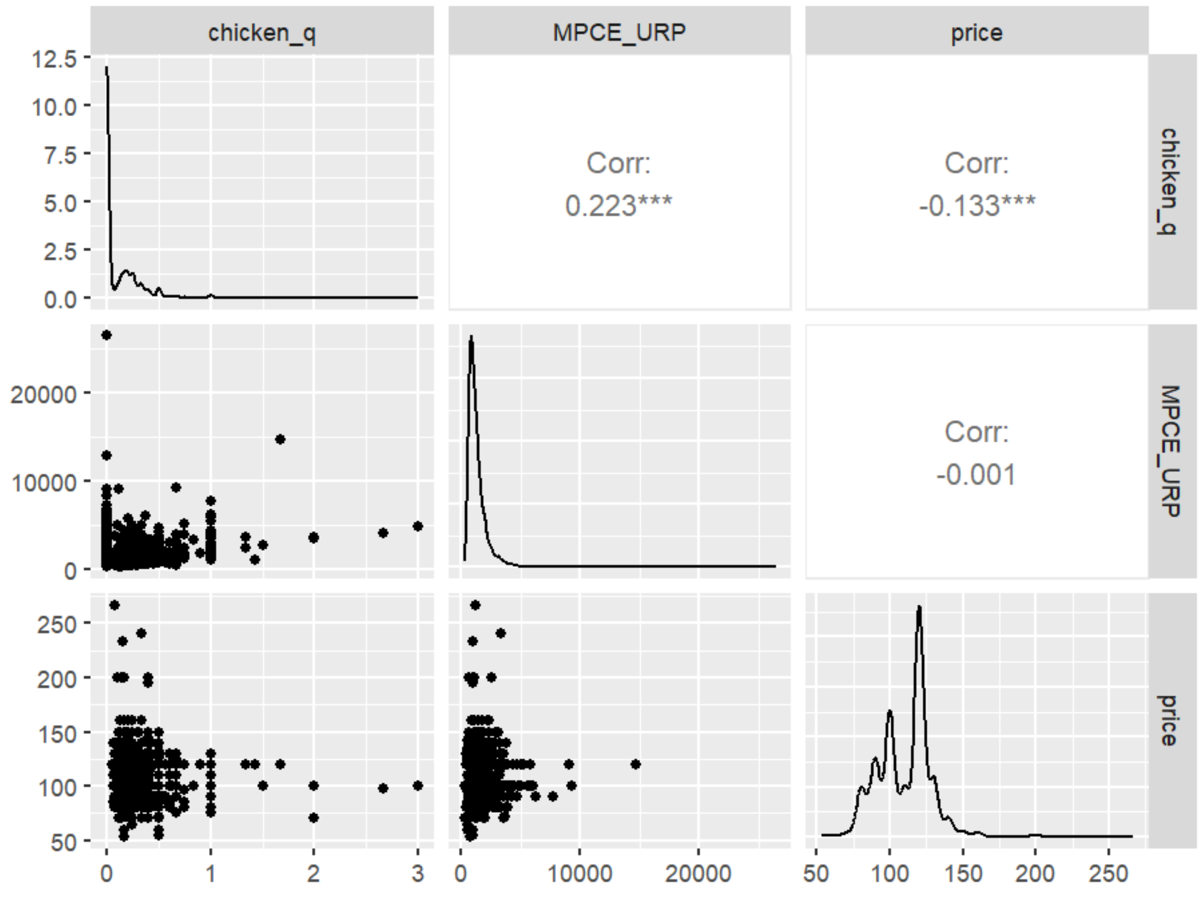
```
                            OLS Regression Results
==============================================================================
Dep. Variable:                     Rs   R-squared:                       0.074
Model:                            OLS   Adj. R-squared:                  0.054
Method:                 Least Squares   F-statistic:                     3.688
Date:                Sun, 23 Jun 2024   Prob (F-statistic):             0.0610
Time:                        15:43:00   Log-Likelihood:                -360.96
No. Observations:                  48   AIC:                             725.9
Df Residuals:                      46   BIC:                             729.7
Df Model:                           1
Covariance Type:            nonrobust
=======================================================================================
                          coef    std err          t      P>|t|      [0.025      0.975]
---------------------------------------------------------------------------------------
const                 396.6881     91.270      4.346      0.000     212.971     580.405
wicket_confirmation    17.6635      9.198      1.920      0.061      -0.851      36.179
==============================================================================
Omnibus:                        6.984   Durbin-Watson:                   2.451
Prob(Omnibus):                  0.030   Jarque-Bera (JB):                6.309
Skew:                           0.877   Prob(JB):                       0.0427
Kurtosis:                       3.274   Cond. No.                         13.8
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

```
> # Print the regression results
> print(summary(model))

Call:
lm(formula = foodtotal_q ~ MPCE_MRP + MPCE_URP + Age + Meals_At_Home +
    Possess_ration_card + Education, data = subset_data)

Residuals:
    Min      1Q  Median      3Q     Max
-68.609  -3.971  -0.654   3.291 239.668

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)          1.138e+01  8.243e-01  13.811  < 2e-16 ***
MPCE_MRP             1.140e-03  5.659e-05  20.152  < 2e-16 ***
MPCE_URP             9.934e-05  3.422e-05   2.903  0.00372 **
Age                  9.884e-02  9.613e-03  10.282  < 2e-16 ***
Meals_At_Home        5.079e-02  6.420e-03   7.911 3.27e-15 ***
Possess_ration_card -2.187e+00  3.025e-01  -7.229 5.79e-13 ***
Education            2.458e-01  3.564e-02   6.898 6.11e-12 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7.667 on 4028 degrees of freedom
  (59 observations deleted due to missingness)
Multiple R-squared:  0.202,	Adjusted R-squared:  0.2008
F-statistic: 169.9 on 6 and 4028 DF,  p-value: < 2.2e-16
```

```
> vif(model) # VIF Value more than 8 its problematic
           MPCE_MRP            MPCE_URP                 Age        Meals_At_Ho
me
           1.636493            1.478309            1.106082            1.1182
80
Possess_ration_card           Education
           1.147250            1.208647
```

```
> print(equation)
[1] "y = 11.38 + 0.00114*x1 + 9.9e-05*x2 + 0.09884*x3 + 0.050789*x4 + -2.1869
64*x5 + 0.245842*x6"
```

```
> head(subset_data$MPCE_MRP,1)
[1] 1124.92
> head(subset_data$MPCE_URP,1)
[1] 982
> head(subset_data$Age,1)
[1] 38
> head(subset_data$Meals_At_Home,1)
[1] 54
> head(subset_data$Possess_ration_card,1)
[1] 1
> head(subset_data$Education,1)
[1] 6
> head(subset_data$foodtotal_q,1)
[1] 17.92535
>
```



RECOMMENDATIONS:

1. **Strategic Player Selection and Retention**:
   - Focus on Consistent Performers
   - Invest in Emerging Talents

2. **Performance-Based Compensation**:
   - Enhance Salary Structures
   - Incentive Programs

3. **Player Development and Training**:
   - Tailored Training Programs
   - Use of Analytics in Training

4. **In-Game Strategy and Planning**:
   - Data-Driven Decision Making:
   - Match-Up Analysis:

5. **Fan Engagement and Marketing**:
   - Highlight Top Performers
   - Transparency with Fans

6. **Future Research and Continuous Improvement**:
   - Regular Data Analysis
   - Explore Advanced Metrics

By implementing these recommendations, IPL teams can improve their performance, make informed financial decisions, and engage fans more effectively. Continuous use of data analysis and statistical modeling will ensure that strategies evolve with the game, maintaining competitive advantage.

CONCLUSION:

In IPL data analysis, it indicates a statistically significant positive relationship between runs scored and player salary, but the low R^2 score suggests that many other factors are also at play. Runs scored is a significant but not a comprehensive predictor of salary.

NSSO68 data analysis underscores the importance of targeted interventions in land management and household support to improve economic conditions. By leveraging these findings, stakeholders can develop strategies to address socio-economic disparities and promote sustainable development, ensuring better economic resilience and growth for households.