# **Data Variables Summary with Time Series Plots: (Sarfaraz)**

* + Nitrogen Consumption
  + Nitrogen Price
  + Water
  + Cropped Acreage
  + Construction of New Tech Proxy Explanation (Mathematical Vs Intuitive Logic)
  + Agric GDP PKR and USD (explain why it’s better to estimate results with USD then PKR)
* Extreme fluctuation in specific time years (5 times) 1998,99 2008, 09 2012, 13 2016-17 using Agric GDP PKR.
* Leading to Skewing of macro numbers in PKR. Impact of foreign exchange value resetting will be impacting 4/6 variables in regression model other than water and cropped acreage.

**Other Data Issues:**

* Explain why we choose 29-year database instead of 30 years. (Large Variability in last years)

# **Statistical Test Parameters Threshold (Uzair)**

The study used statistical tests to validate the robustness of the regression models. These tests have helped us assess the variability of the coefficients, mitigate problems of multicollinearity, heteroscedasticity and helped in identification of observational values in the database. A careful consideration of these thresholds is necessary to ensure reliable statistical inferences for the time series database.

1. Durbin Watson Test:

The Durbin-Watson test is used to detect the presence of autocorrelation in the residuals of a regression analysis. Autocorrelation occurs when the residuals (errors) of the regression model are not independent of each other. The test statistic ranges from 0 to 4, where:

* A value close to 2 suggests no autocorrelation.
* 0 to 1.5 suggests strong positive autocorrelation.
* 1.5 to 2.5 suggests no autocorrelation
* 2.5 to 4 suggests strong negative autocorrelation
* Values outside the range of 1.5 to 2.5 suggest that the model has issues with autocorrelation, warranting further investigation.

1. White Test:

The White test is used to detect heteroscedasticity in a regression model, which occurs when the variance of the residuals is not constant across observations.

* The white test checks for the null hypothesis of homoscedasticity (constant variance). If the test is rejected typically at a 5% significance level (p-value < 0.05), it suggests that the regression model’s assumption of homoscedasticity will be violated.

1. Shapiro-Wilk Test:

The Shapiro-Wilk test is used to assess the normality of a sample. It tests that a sample comes from a normally distributed population. We used this test used to verify the normality of the residuals which is one of the crucial assumptions for Multivariate Time Series Modelling.

The threshold is set as follows:

* A p-value > 0.05 suggests that the data does not significantly deviate from normality, and the data is considered normally distributed.
* A p-value < 0.05 indicates a significant deviation from normality, rejecting the null hypothesis, and implying that the data may not be normally distributed.

1. Variance Inflation Factors (VIF):

Variance Inflation Factors (VIF) are widely used technique to quantify the extent of multicollinearity in a regression model. Multicollinearity occurs when independent variables are highly correlated, leading to inflated standard errors and unreliable coefficient estimates for regression analysis.

The threshold is set as follows:

* VIF < 5: Indicates low multicollinearity.
* VIF between 5 and 10: Suggests moderate multicollinearity, which may require some consideration.
* VIF > 10: Signals high multicollinearity, typically warranting corrective measures such as removing or combining variables such as taking ratios.

1. Cook's Distance:

Cook's Distance is a measure used to identify influential data points in a regression model. An influential data point is one that, if removed, would change the model's results significantly.

The threshold is set as follows:

* Cook’s Distance > 1 flags highly influential points that might unduly affect the model.
* Cook’s Distance < 1 flags data points that are less likely to be influential and typically do not have a significant impact on the model's results.

# **Summary of Regression Runs (Uzair)**

## As stated in the inception report that we planned to use stepwise process of regression runs. During these runs, we changed the functional forms of the model specification to test the results and hypothesis.

Our generalized regression equation format is as below:

Yt = a0 +β1X1,t + β2X2,t + β3X3,t + β4X4,t + β5X5,t + β6X6,t + β7X7,t + β8X8,t + εt

Y is the dependent / decision variable – Nitrogen Demand

a0 is the constant

B1, B2, B3, B4, B5, B6 B7 B8 are the coefficients of the independent variables

X1, X2, X3, X4, X5, X6 X7 X8 are the independent variables

ℇt is the error term to capture the difference between the projected and historical value for all years of database.

* All the regression runs are made for 29 years database. We ran regression runs on 28-year database to assess the 2021-2022 year’s impact since it is one of the outlier years given by Cook’s distance at the end of the first regression run. Though the results of the 28-year model is not reported but we investigated and reviewed the year 2008-2009, 2014-2015 and 2021-2022 in detail as these were considered influential data points.
* We ran the regression runs using the updated definition of Agricultural Labor Productivity as defined in the first section of Data Variables in the report rather than defined in the Inception report
* In total, three regression runs were made, focusing on models that incorporated both 6-variable and 8-variable specifications. Different model specifications were tested in, including taking ratios for Agriculture Credit with Agric GDP PKR ultimately leading to the development of the optimal model in the third run. Next section will discuss the model coefficients and interpretations with emphasis on third regression run evaluation.
* We decided to lock the 6-variable model for the third regression run as our final model. Therefore, we also present the line of best analysis with ht **w**

## Regression Results Run 1 (Uzair)

Function Equation:

Insert Equation here

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Table 1. Estimates of The First Regression Run | | | | | | | |
| **Variables** | **Units** | **6 Variables** | | | **8 Variables** | | |
| **Coefficients** | **t-values** | **VIF** | **Coefficients** | **t-values** | **VIF** |
| **N – Price** | PKR - MT | -0.0052 | -0.89 | 10.04 | -0.014 | -0.82 | 79.99 |
| **Water** | MAF | 34.83\*\* | 2.24 | 1.65 | 20.79 | 1.06 | 2.54 |
| **Cropped Acreage** | Mn Ha | 137.75 | 1.21 | 2.084 | 101.85 | 0.8 | 2.56 |
| **Agric GDP** | Bn PKR | 0.49 | 1.18 | 93.00 | 0.51 | 1.21 | 94.27 |
| **Agric Credit** | Mn PKR | -0.00016 | -0.31 | 17.34 | 0.00019 | 0.33 | 22.94 |
| **Labour Productivity** | PKR/Ha | 1.47 | 0.26 | 49.79 | 2.05 | 0.26 | 95.43 |
| **Input to Output** | Ratio | - | - | - | -1115.58 | -0.81 | 16.71 |
| **N to P price** | Ratio | - | - | - | -1374.44 | -0.99 | 4.29 |
| **R²** |  | 0.846 | | | 0.856 | | |
| **F. Value** |  | 20.11 | | | 14.91 | | |
| **DW. Stats (Probability)** |  | 1.84  (0.22) | | | 1.94  (0.27) | | |
| **White.Stats (Probability)** |  | 2.24  (0.33) | | | 2.00  (0.37) | | |
| **Shapiro.Stats**  **(Probability)** |  | 0.96  (0.37) | | | 0.92  (0.04) | | |

\*Significant at 90 % Confidence level

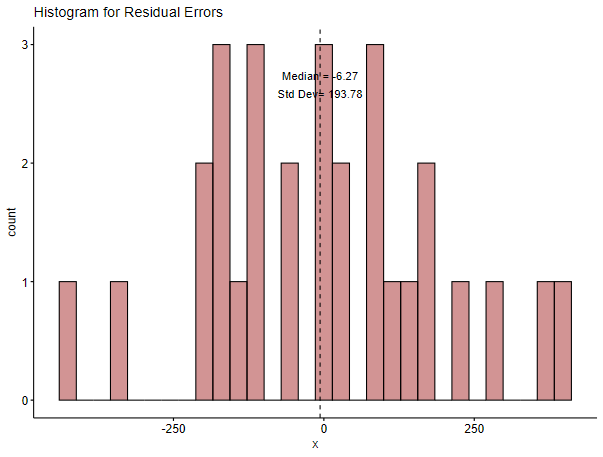
\*\*Significant at 95 % Confidence level

\*\*\*Significant at 99 % Confidence level

1. Nitrogen Price (N – Price):
   1. Coefficients for both models (6 variables and 8 variables) are negative, indicating that a higher price of nitrogen leads to reduced nitrogen consumption. However, the coefficients are not statistically significant, as the t-values are -0.89 and -0.82, both below the threshold for 90% significance Confidence Interval. This suggests that nitrogen price does not have a substantial impact on nitrogen consumption in this model.
   2. The Variance Inflation Factor (VIF) for Nitrogen Price is extremely high in the 8-variable model (79.99), indicating multicollinearity, meaning that Nitrogen price is highly correlated with other independent variables in the model.
2. Water Availability:
   1. In the 6-variable model, water Availability has a positive and statistically significant coefficient (34.83), with a t-value of 2.24, showing that water has a significant positive effect on nitrogen consumption. This means that each additional million Acre Feet (MAF) increase in water availability leads to approximately 34.83 more tons of Nitrogen Consumption (Nutrient per thousand tons), assuming all other factors remain constant.
   2. In the 8-variable model, however, the coefficient decreases to 20.79 and becomes statistically insignificant (t-value 1.06), indicating that the inclusion of additional variables like Input to Output ratio and N to P price ratio may reduce the impact of water availability on nitrogen demand.
3. Agricultural GDP (in Bn PKR):
   1. The coefficient for Agric GDP is positive but not statistically significant in both models (t-values 1.18 and 1.21), suggesting a weak relationship between nitrogen consumption and agricultural GDP in PKR.
4. Agricultural Credit:
   1. In the 6-variable model, the coefficient for Agricultural Credit is negative (-0.00016), and in the 8-variable model, it becomes positive (0.00019). Neither coefficient is statistically significant (t-values -0.31 and 0.33), suggesting that agricultural credit does not have a meaningful impact on nitrogen consumption.
   2. The VIF values (17.34 and 22.94) are elevated, which suggests that Agriculture Credit has high multicollinearity in this model.
5. Labour Productivity:
   1. Both models show small, statistically insignificant coefficients for Labour Productivity (t-values 0.26), indicating that this variable does not significantly impact nitrogen consumption.
   2. The high VIF in the 8-variable model (95.43) suggests multicollinearity.
6. Input to Output Ratio and N to P Price Ratio (only in the 8-variable model):
   1. Both these variables have negative coefficients, implying an inverse relationship with nitrogen consumption, but neither is statistically significant. The t-values (-0.81 and -0.99) are below the significance threshold.
7. R-squared values for both models are relatively high (0.846 and 0.856), indicating that the models explain around 85% of the variance in nitrogen consumption. This suggests that the models have a good fit but there is still a room for improvement in further iterations.
8. Durbin-Watson Statistics (1.84 and 1.94) show that there is no serious autocorrelation in the residuals.
9. White and Shapiro statistics suggest no issues with heteroscedasticity and normality in the 6-variable model. However, the Shapiro stat in the 8-variable model (0.92, p=0.04) indicates a deviation from normality at a 95% confidence level.

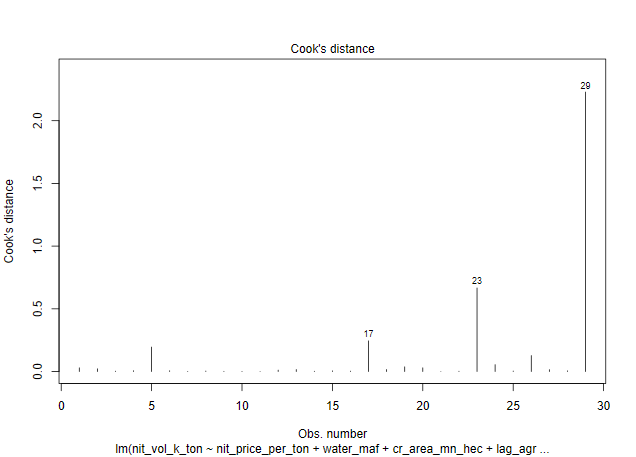
Problems in the the first Model:

1. Agricultural GDP (PKR vs. USD):
   * High VIF values (93.00 and 94.27) suggest multicollinearity, indicating that agricultural GDP is correlated with other variables in the model. This raises the need for exploring GDP in other units (e.g., USD). The problems for Agriculture GDP (PKR) variable have already been discussed in detail in Data Variables section of the report. The Agriculture GDP variable effects four out of six variables due to foreign exchange value resetting in the economy for 5 times in the data base.
2. Agricultural Credit Negative Coefficient:
   * The negative sign in of Agriculture Credit in the 6 variable model is against the economic rationale. As the availability of Agriculture credit increases the farmers purchasing power for Nitrogen Consumption increases. This economic rationale is violated in the current model leading us to explore further alternatives of using this variable in the analysis by either taking ratios or interaction with other variables.
3. Multicollinearity:
   * Many of the VIF values, particularly for Nitrogen Price, Agric GDP, Labour Productivity, and Agricultural Credit, are very high, indicating multicollinearity. This means that these variables are likely highly correlated with each other, which undermines the reliability of the coefficient estimates. Addressing multicollinearity by removing or transforming variables could improve model robustness for next run.
4. Non-Normality (8-Variable Model):
   * The Shapiro-Wilk statistic indicates non-normality in the residuals of the 8-variable model. This suggests a violation of one of the key assumptions of linear regression, where residuals are expected to follow a normal distribution. Non-normal residuals can bias the standard errors of the coefficients, affecting the accuracy of confidence intervals. As a result, the confidence intervals may become too wide or too narrow, leading to incorrect conclusions about the significance of the model’s predictors. The accompanying graph further confirms the non-normal distribution of residuals, highlighting the potential for unreliable inferences in the regression results.



1. Influential Data Points:

6 and 8 variable model results showed that both 2008-09, 2014-15 and 2021-22 were outlier years on the fitted regression models. The problem can be visualized in the following graph



Cook's distance values indicate the influence of each observation on the fitted regression model. Most of the observations in the graph have a Cook’s distance near zero, meaning they had little influence on the model.

However, observations 17 (2008-09), 23 (2014-15) and particularly 29 (2021-22) have high Cook's distance values. The desired threshold for the Cook’s distance is 1 and observations exceeding this threshold are considered influential data points. Observation 29 has an exceptionally large influence, exceeding a value of 2, which suggests it could be an influential or outlier point. These exceptional kinks in the data points can be explained by exchange rate depreciation. Pak rupee depreciated by 25% in 2008-09 compared to 2007-08. Similarly, it depreciated by 11% in 2021-2022. This lead to distortion of the Economic variables particularly Agriculture GDP, Agriculture Credit, Labour Productivity and Nitrogen Prices, impacting 4 out 6 variables directly in the model.

Regression Results Run 2 (Uzair)

Function Equation:

Insert Equation here

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Table 2. Estimates of The Second Regression Run | | | | | | | |
| **Variables** | **Units** | **6 Variables** | | | **8 Variables** | | |
| **Coefficients** | **t-values** | **VIF** | **Coefficients** | **t-values** | **VIF** |
| **N – Price** | PKR – MT | -0.014\*\*\* | -2.76 | 10.89 | -0.03\* | -1.96 | 85.28 |
| **Water** | MAF | 16.25 | 1.19 | 1.95 | 10.24 | 0.63 | 2.64 |
| **Cropped Acreage** | Mn Ha | 103.37 | 1.12 | 2.11 | 119.69 | 1.16 | 2.57 |
| **Agric GDP** | Bn PKR | 75.40\*\*\* | 3.73 | 13.92 | 77.72\*\*\* | 3.54 | 15.82 |
| **Agric Credit** | Mn PKR | -0.000091 | -0.26 | 13.71 | 0.00006 | 0.13 | 20.66 |
| **Labour Productivity** | PKR/Ha | 5.40\* | 1.99 | 17.39 | 10.04\*\* | 1.78 | 72.94 |
| **Input to Output** | Ratio | - | - | - | -1240.14 | -1.11 | 16.62 |
| **N to P price** | Ratio | - | - | - | -42.62 | -0.036 | 4.81 |
| **R²** |  | 0.90 | | | 0.905 | | |
| **F. Value** |  | 32.81 | | | 23.91 | | |
| **DW Stats (Probability)** |  | 2.24  (0.79) | | | 2.32  (0.87) | | |
| **White Stats (Probability)** |  | 3.95  (0.14) | | | 3.79  (0.15) | | |
| **Shapiro Stats**  **(Probability)** |  | 0.98  (0.84) | | | 0.99  (0.96) | | |

\*Significant at 90 % Confidence level

\*\*Significant at 95 % Confidence level

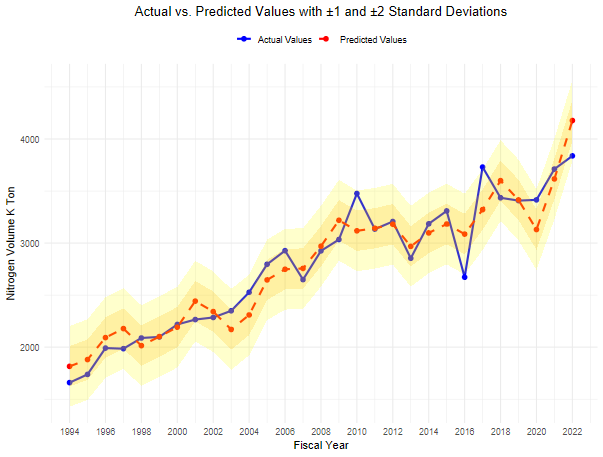
\*\*\*Significant at 99 % Confidence level

1. Nitrogen Price:
   1. In both models, the price of nitrogen has a negative and “significant” effect on nitrogen consumption. In the 6-variable model, the coefficient is -0.014 with a t-value of -2.76 significant at 99% Confidence Interval, while in the 8-variable model, the coefficient increases to -0.03 with a t-value of -1.96 significant at 90% Confidence Interval. This shows that nitrogen consumption is more sensitive to price changes when additional variables are included, though multicollinearity increases in the 8-variable model (VIF 85.28).
   2. In the first regression run, the coefficient for Nitrogen Price was not statistically significant. The improvement in significance in this second regression suggests that the model is better specified, capturing the impact of nitrogen price on consumption more clearly. This means based on the 2021-2022 data, if the price of nitrogen increases by 1%, starting from 83,173 (PKR/MT), it would lead to a decrease in nitrogen consumption by 11.26 (NKT) assuming all other factors remain constant.
2. Agriculture GDP USD:
   1. Agricultural GDP is highly significant in both models. In the 6-variable model, the coefficient is 75.40 (t-value 3.73) significant at 99% Confidence Level, and in the 8-variable model, it is 77.72 (t-value 3.54) significant at 99% Confidence Level. The relationship between agricultural output and nitrogen consumption remains strong and consistent, though the VIF value of 13.92 and 15.82 indicates some multicollinearity in both models (higher in the 8-variable model).
   2. In the first regression run, Agricultural GDP (measured in PKR) was not statistically significant, and high VIF values indicated potential multicollinearity, which suggested that the variable was highly correlated with other predictors. However, in the second regression, using Agricultural GDP measured in USD, the coefficient became highly significant, and the multicollinearity issue was substantially reduced. Specifically, the VIF value decreased by 85%, from 93.00 in the first run to 13.92 in the second run. This improvement enhances the robustness and reliability of the coefficient, making the model more accurate in capturing the effect of Agricultural GDP on nitrogen consumption.
3. Agriculture Credit Million PKR:
   1. Agricultural Credit shows a negative and insignificant coefficient in the 6-variable model (-0.000091, t-value -0.26) and a slight positive, but still insignificant, coefficient in the 8-variable model (0.00006, t-value 0.13). The negative sign in the 6-variable model persists, but the overall relationship remains weak in both the models.
4. Labour Productivity:
   1. In the 6-variable model, Labour Productivity has a positive and significant coefficient of 5.40 (t-value 1.99) significant at while in the 8-variable model, the effect becomes stronger with a coefficient of 10.04 (t-value 1.78) significant at 95% of Confidence Interval. The impact of labor productivity on nitrogen consumption grows by 46% when more variables are included, though multicollinearity increases significantly in the 8-variable model (VIF 72.94).
   2. In the first regression run, Labour Productivity was not significant, and its VIF value was 49.79 (more than 5 times the accepted threshold of 10). In the second run, the variable became significant at 95% Confidence Interval, indicating a better model specification and highlighting its importance in explaining nitrogen demand.
5. Model Fitness:
   1. The R² values of 0.90 (6 variables) and 0.905 (8 variables) show an improvement in model fit compared to the first regression (0.846 and 0.856). This indicates that the models explain approximately 90% of the variation in nitrogen consumption, suggesting that the second regression provides a better fit.
   2. Both 6 and 8 variables models are highly significant, with F-values of 32.81 (6-variable) and 23.91 (8-variable), showing that the independent variables explain nitrogen consumption effectively.
   3. The DW stats (2.24 and 2.32) indicate no autocorrelation issues, which is an improvement over the first regression run.

Problems in the Second Model:

1. Multicollinearity Issues:
   1. The VIF value for Nitrogen Price increases to 85.28 in the 8-variable model, indicating a serious multicollinearity problem. This suggests that Nitrogen Price is highly correlated with other independent variables, making it difficult to isolate its individual effect.
   2. The VIF value for Labour Productivity has an improvement dropping from 95.43 to 72.94 in the 8-variable models, suggesting that the Labour productivity is still correlated with other variables in the model.
   3. Although the coefficient for Agricultural GDP is highly significant, its VIF values are still elevated at 13.92 and 15.82), suggesting that it still suffers from some degree of multicollinearity, though the issue has dropped from 93.00 to 13.92 in the second regression run.
2. Negative Sign in Agricultural Credit:
   1. In both the 6-variable and 8-variable models, the coefficient for Agricultural Credit is negative and insignificant. This negative sign is unexpected, as one would assume that increased credit would promote higher nitrogen consumption. This result suggest that the variable needs some tweaking to support the economic rationale.

In summary, the second regression run provides significant improvements in the model's fitness, with several important coefficients becoming statistically significant, indicating better specification. However, multicollinearity remains an issue, and the negative sign in Agricultural Credit requires further investigation or potential model adjustments. The plot below shows the actual and predicted values for the eight variables estimated model.



Regression Results Run 3 (Daniyal)

* Explanation of Coefficients (6 variables)
* Brief discussion (8 variables)
* Elasticities explanation (6 variables)
* Graphical Presentation of Results
* In depth analysis for 6 vars
* 8 vars have same 3 vars significant there. 1% increase in R2 for 8 variables only.
* Residuals bell curve more normally distributed for 6 variables
* Cooks distance, residual vs fit graph
* To check for the versality of the results, readers are encouraged to view the diagrams mentioned in the Excel Sheet for third run.

Conclusion (Daniyal)

* Therefore, we decided to to lock in as our base case and regression.
* Sheer complexity to update values for 8 variables due to ratios so we chose 6 variables.
* Then show Line of Best Fit (6 variables) with explanation of CAGR and how well the model is predicting the trend.