

# Linear Regression Final Project

STAT 371

Instructor: Dr. Schonlau

Presentation by: Dhruv Trivedi

## Dataset Summary: Gujarat Dairy Farms

- **Objective:** The dataset was collected to analyze dairy farm productivity and economic factors in Gujarat.
- **Collection Method:** A survey was disseminated via Google Forms to roughly 500 dairy farmers through the start-up Nityam's network, yielding approx. 300 responses (60% response rate).
- **Content:** The dataset includes variables on milk production, costs, revenue, and government support satisfaction.
- **Confidentiality:** Respondent anonymity was strictly maintained, with no personal identifiers collected

### **Variables List and type:**

- Number of Cows on the Farm: Numerical data representing the count of cows.
- Number of Buffaloes on the Farm: Numerical data representing the count of buffaloes.
- Location of Farm in Gujarat: Categorical data indicating the farm's location.
- Average Daily Milk Production per day (in litres): Numerical data on daily milk production.
- Milk Collection Centre Affiliation: Categorical data showing the affiliated milk collection center.
- Yearly Expenditure on Animal Health (in INR): Numerical data indicating monthly spending on animal health.
- Yearly Income from Selling Manure (if applicable, in INR): Numerical data showing income from selling manure.
- Primary Feed for Livestock: Categorical data about the primary feed type.
- Satisfaction with Government Support: Numerical rating of satisfaction with government support.
- Approximate Monthly Operating Costs (in INR): Numerical data on monthly operating costs.
- Approximate Monthly Revenue (in INR): Numerical data indicating monthly revenue.
- Use of Automation in Farming Operations: Categorical data indicating if automation is used.
- Number of Family Members and/or Employees Working at the Farm: Numerical data on the count of family members or employees working on the farm.

# Cleaning and Prepping Data:

- Loading and Initial Assessment: The dataset from Gujarat farms was initially loaded and examined to understand its structure and contents. This included various variables related to livestock, milk production, financial metrics, and categorical variables like location, satisfaction with government support, and milk collection centers.
- Handling Categorical Variables:
  - **Location of Farm:** The variable 'Location of Farm' was transformed into dummy variables using one-hot encoding, with 'Vadodara' as the baseline (most frequent category). This reduces multicollinearity in regression models by representing categories as separate binary variables.
  - **Satisfaction with Government Support:** This variable was first categorized into three groups ('Satisfaction\_1\_4', 'Satisfaction\_5\_7', 'Satisfaction\_8\_10') based on satisfaction scores. Later, these categories were converted to dummy variables with 'Satisfaction\_1\_4' (most frequent) as the baseline and the original variable was removed.
  - **Milk Collection Centre:** Similar to 'Location of Farm', this was transformed into dummy variables, with 'Kwality Limited' as the baseline.
  - **Primary Feed for Livestock and Use of Automation:** Both variables were converted to binary dummy variables, simplifying them for regression analysis.

## Data Cleaning:

- **Removing NA Entries:** Rows containing NA entries were identified and removed. This step ensured that the dataset used for regression would not have missing values, which can distort regression analysis results.
- **Normalization of Time-Related Variables:** Variables with yearly and monthly frequencies were converted to daily frequencies to align with 'Average Daily Milk Production (litres)', which is a daily measure. This step is crucial for regression analysis as it standardizes the time scale across all variables, allowing for more accurate and interpretable coefficients.

Remaining observations: 292/300

## Why have we mapped categorical variables ?

- As, x-variables need not be continuous, to deal with the type of data types of variables in linear regression we convert non-numerical categories into a numerical format, creating a clear binary distinction that avoids implying any natural order among categories. This binary coding makes model coefficients more interpretable and simplifies the model by allowing it to estimate distinct effects for each category.
- “To avoid collinearity as response from farmers were from a list of given options”
- “The baseline category is the left-out category, and the interpretation of an indicator variable is relative to the baseline category.”

Response Variate - "Average Daily Milk Production per day (in litres)"

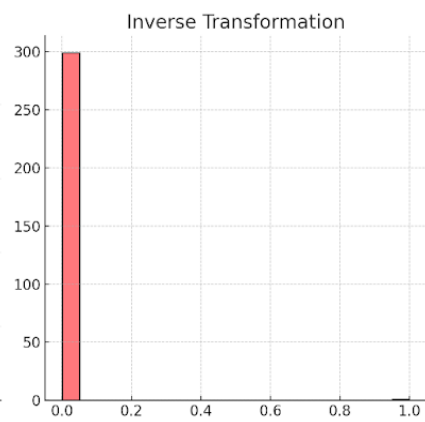
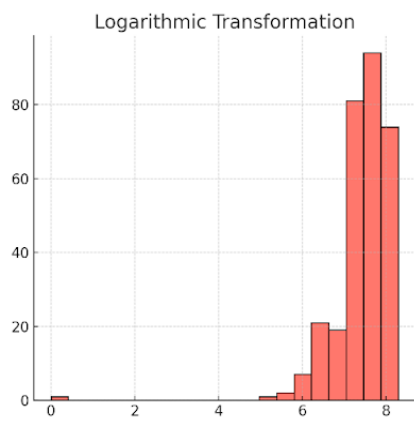
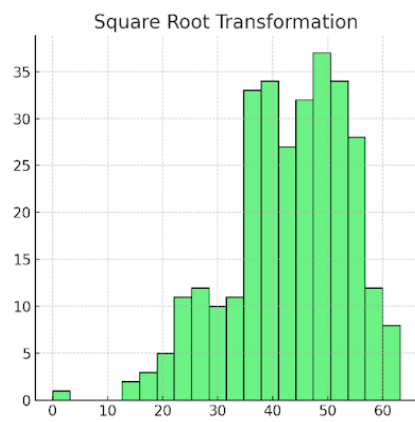
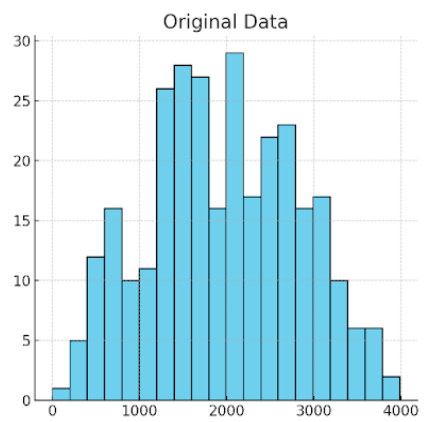
### Normality Check?

- The histogram of daily milk production suggests that the data may not be perfectly normally distributed, as indicated by the shape of the distribution. Appears to be relatively normally distributed with a slight skew.

Transformation:

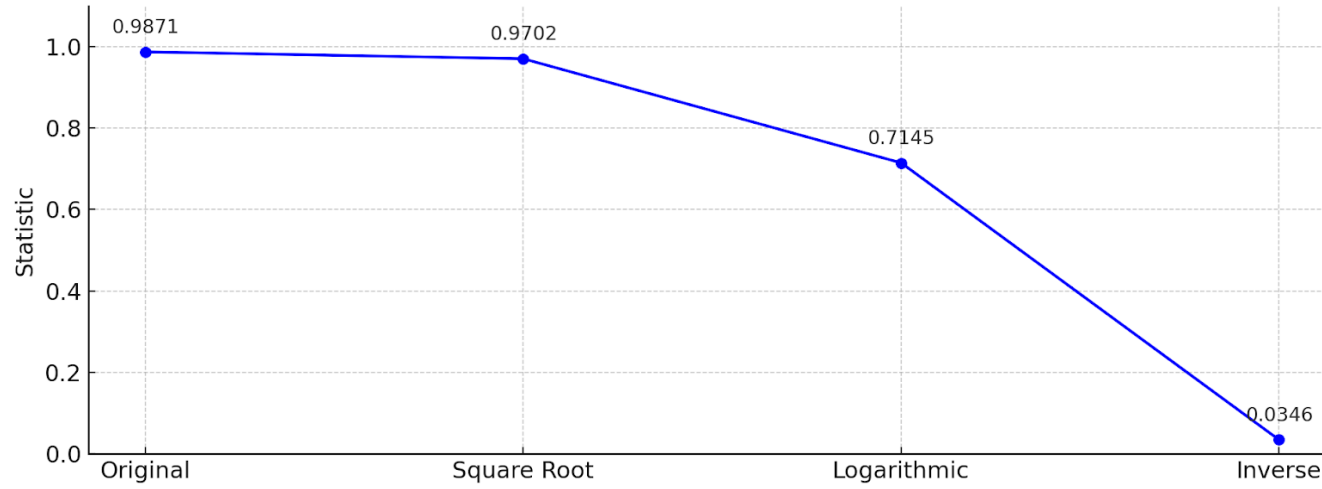
- Square Root Transformation ( $\sqrt{\text{Milk Production}}$ ): Moderates the skew slightly, but still maintains a similar shape to the original data.
- Logarithmic Transformation ( $\log(\text{Milk Production})$ ): Significantly alters the distribution, showing a more pronounced skew.
- Inverse Transformation ( $1/(\text{Milk Production})$ ): Leads to a highly skewed distribution, very different from the original data.

*-> The response variable (here, daily milk production) is not normally distributed and cannot be successfully transformed into a normal distribution so further analysis would be based on the result here that the response variate isn't following normal distribution.*

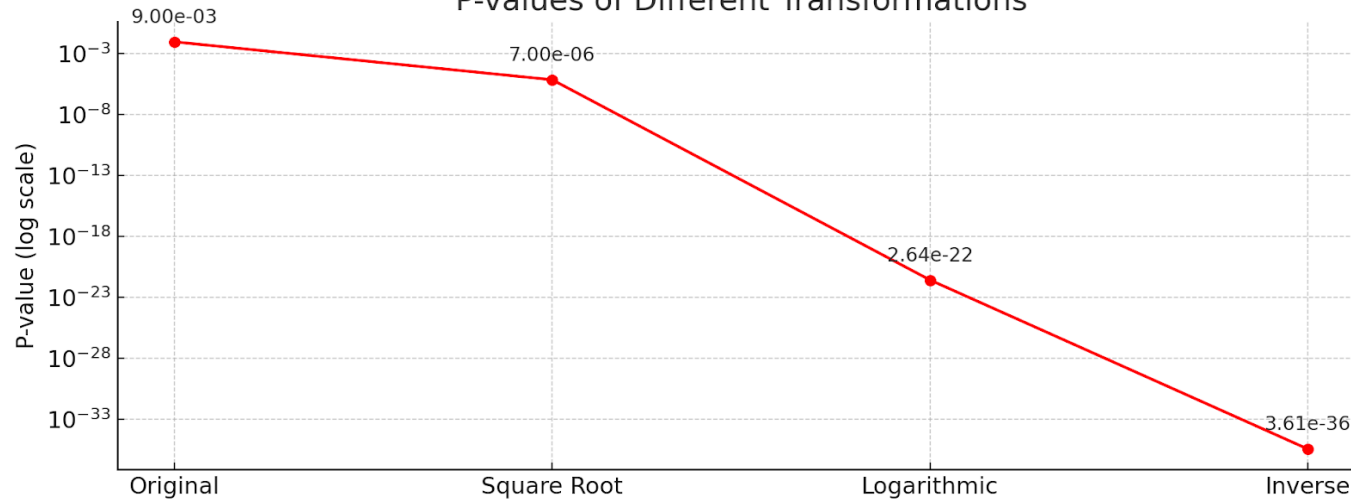




Statistics of Different Transformations



P-values of Different Transformations



- Initial Model/Full Model:

$$y = \alpha + \beta_1(\text{number\_of\_cows}) + \beta_2(\text{number\_of\_buffaloes}) + \beta_3(\text{number\_of\_family\_members\_employees\_working\_at\_farm}) + \beta_4(\text{daily\_expenditure\_on\_animal\_health\_inr}) + \beta_5(\text{daily\_income\_from\_selling\_manure\_inr}) + \beta_6(\text{daily\_operating\_costs\_inr}) + \beta_7(\text{daily\_revenue\_inr}) + \beta_8(\text{ahmedabad}) + \beta_9(\text{rajkot}) + \beta_{10}(\text{surat}) + \beta_{11}(\text{jamnagar}) + \beta_{12}(\text{aavin}) + \beta_{13}(\text{amul}) + \beta_{14}(\text{dudhsagar dairy}) + \beta_{15}(\text{dynamix dairy}) + \beta_{16}(\text{karnataka co-operative milk federation}) + \beta_{17}(\text{selling privately to consumers}) + \beta_{18}(\text{mother dairy}) + \beta_{19}(\text{orissa state cooperative milk producers federation}) + \beta_{20}(\text{parag milk foods ltd}) + \beta_{21}(\text{verka}) + \beta_{22}(\text{natural plants}) + \beta_{23}(\text{Satisfaction\_5\_7}) + \beta_{24}(\text{Satisfaction\_8\_10}) + \beta_{25}(\text{use\_of\_automation}) + \epsilon'$$

- $y$  is the average daily milk production in litres.
- $\alpha$  is the intercept of the model.
- Each  $\beta$  represents the coefficient of the corresponding variable in the model, indicating the expected change in the response variable ( $y$ ) for a one-unit change in the predictor variable, holding all other predictors constant.
- $\varepsilon$  represents the error term of the model, capturing the variation in  $y$  not explained by the predictors.


# Regression Results:

- The Ordinary Least Squares (OLS) regression analysis of the dataset produced the following statistical results:
- Dependent Variable: Average Daily Milk Production (litres)
- R-squared: 0.586 - This implies that approximately 58.6% of the variance in the dependent variable (daily milk production) can be explained by the independent variables in the model. Adjusted R-squared: 0.545 - This is a modified version of R-squared that has been adjusted for the number of predictors in the model. It provides a more accurate measure of the model's explanatory power.
- F-statistic: 14.41 - This value tests the overall significance of the regression model.
- The associated Prob (F-statistic) is very small ( $3.70e-37$ ), indicating that the overall model is statistically significant.
- Coefficients: Number of Cows and Number of Buffaloes show significant positive coefficients, indicating that an increase in their numbers is associated with an increase in milk production. Daily Expenditure on Animal Health (INR) and Daily Revenue (INR) also show significant positive relationships with milk production.
- Daily Income from Selling Manure (INR) shows a significant negative relationship.
- Other variables, such as Use of Automation, Satisfaction with Government Support categories, and various location-based dummy variables, were not statistically significant in this model.


**\*\*Potential Multicollinearity: The condition number is quite large, indicating potential multicollinearity issues. This might require further investigation to ensure the model's reliability.\*\***

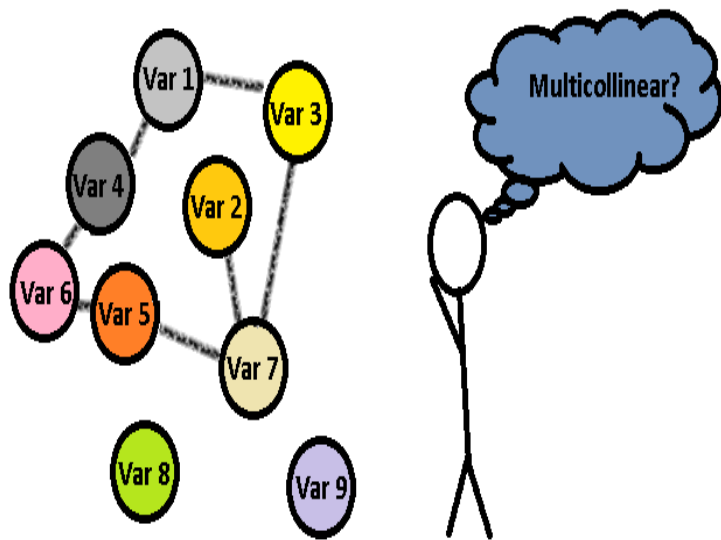


## OLS Regression Results



<b>Dep. Variable:</b>	Average Daily Milk Production (litres)	<b>R-squared:</b>	0.585
<b>Model:</b>	OLS	<b>Adj. R-squared:</b>	0.546
<b>Method:</b>	Least Squares	<b>F-statistic:</b>	15.03
<b>Date:</b>	Tue, 28 Nov 2023	<b>Prob (F-statistic):</b>	1.02e-37
<b>Time:</b>	19:45:03	<b>Log-Likelihood:</b>	-2251.7
<b>No. Observations:</b>	292	<b>AIC:</b>	4555.
<b>Df Residuals:</b>	266	<b>BIC:</b>	4651.
<b>Df Model:</b>	25		
<b>Covariance Type:</b>	nonrobust		





### *Why need to address Multicollinearity?*

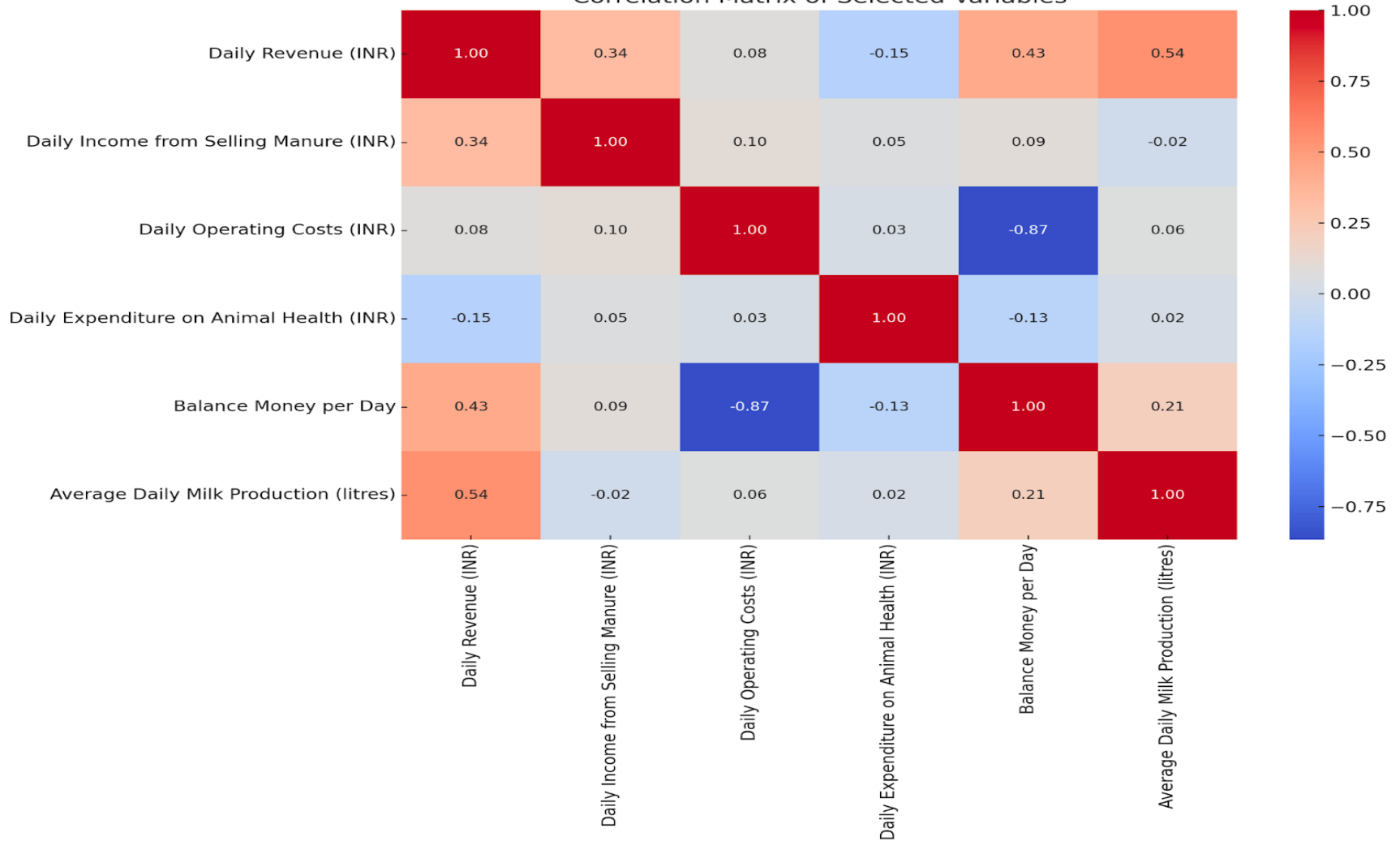
- Multicollinearity is the occurrence of high intercorrelations among two or more independent variables in a multiple regression model.
- Multicollinearity can lead to skewed or misleading results when a researcher or analyst attempts to determine how well each independent variable can be used most effectively to predict or understand the dependent variable in a statistical model.
- In general, multicollinearity can lead to wider confidence intervals that produce less reliable probabilities in terms of the effect of independent variables in a model.

Daily Revenue (INR) shows a VIF of 22.66, which is significantly higher than the common **threshold of 10**, indicating strong multicollinearity.

	<b>variable</b>	<b>VIF</b>
<b>22</b>	Daily Revenue (INR)	22.626124
<b>0</b>	Number of Cows	8.568805
<b>20</b>	Daily Income from Selling Manure (INR)	6.150645
<b>1</b>	Number of Buffaloes	5.952573
<b>19</b>	Daily Expenditure on Animal Health (INR)	5.561436
<b>21</b>	Daily Operating Costs (INR)	5.468870
<b>2</b>	Number of Family Members/Employees Working at ...	4.499684
<b>4</b>	jamnagar	2.141800
<b>6</b>	surat	2.062678
<b>18</b>	Use_of_Automation	1.964859
<b>5</b>	rajkot	1.954350
<b>8</b>	amul	1.908705
<b>12</b>	mother dairy	1.897252
<b>15</b>	selling privately to consumers	1.894949
<b>17</b>	natural plants	1.887057
<b>14</b>	parag milk foods ltd	1.864444
<b>3</b>	ahmedabad	1.761126
<b>7</b>	aavin	1.747232
<b>24</b>	Satisfaction_8_10	1.734666
<b>23</b>	Satisfaction_5_7	1.664132
<b>13</b>	orissa state cooperative milk producers federa...	1.646243
<b>9</b>	dudhsagar dairy	1.599509
<b>11</b>	karnataka co-operative milk federation	1.546725
<b>10</b>	dynamix dairy	1.471665
<b>16</b>	verka	1.389482



Correlation Matrix of Selected Variables



Utilizing correlation matrix saw that Daily Revenue (INR) was highly correlated to other fiscal variables so the Variance Inflation Factor (VIF) for the full model with the newly created 'Balance Money per Day' variable (which combines 'Daily Revenue (INR)', 'Daily Income from Selling Manure (INR)', 'Daily Operating Costs (INR)', and 'Daily Expenditure on Animal Health (INR)') shows the following:

- The VIF for 'Balance Money per Day' is 1.44, indicating that this new variable does not contribute to multicollinearity in the model.
- The highest VIFs are now observed in 'Number of Buffaloes' (5.02), 'Number of Family Members/Employees Working at the Farm' (4.39), and 'Number of Cows' (4.22). These values suggest moderate multicollinearity but are considerably lower than the threshold of 10.
- All other variables have VIF values well below 5, indicating minimal concerns regarding multicollinearity.
- Concluding, by the creation of the 'Balance Money per Day' variable has effectively addressed the previous multicollinearity issues related to the individual financial components, leading to a more stable and reliable regression model.

The OLS regression results for the full model with the new 'Balance Money per Day' variable and without the individual fiscal variables are as follows:

- R-squared: 0.558 - This indicates that about 55.8% of the variance in the dependent variable (daily milk production) is explained by the model.
- Adjusted R-squared: 0.520 - This adjusted measure, which accounts for the number of predictors, is 52.0%.
- Key Points:
- 'Balance Money per Day' Variable: This new variable is not statistically significant in the model ( $p$ -value  $> 0.05$ ). Its coefficient is -0.0126, suggesting a minor and insignificant effect on daily milk production.
- Other Significant Predictors: 'Number of Cows' and 'Number of Buffaloes' remain significant predictors of daily milk production, both with positive coefficients.
- Model Fit: While the model's R-squared is slightly lower than the previous model (0.586), it still explains a substantial part of the variance in milk production.

Overall, while the 'Balance Money per Day' variable helped address multicollinearity issues, it does not appear to be a significant predictor of daily milk production in this model. The model's explanatory power (as measured by R-squared) is slightly lower than in the previous model but remains substantial.

## Is Location an important criteria here?

- To test if the coefficients for the location variables 'Ahmedabad', 'Rajkot', 'Surat', and 'Jamnagar' are all equal to zero in the regression model, performing a hypothesis test.
- Testing the null hypothesis that all these coefficients are equal to zero against the alternative hypothesis that at least one of them is not equal to zero.
- The results of the F-test comparing the full model (including the location variables 'Ahmedabad', 'Rajkot', 'Surat', and 'Jamnagar') against the reduced model (excluding these location variables) are as follows:
- F-Statistic: 1.010 P-Value: 0.402 df\_diff: 4.0 (degrees of freedom difference between the models)

*The test result indicates that there isn't enough statistical evidence to suggest that the average daily milk production in 'Ahmedabad', 'Rajkot', 'Surat', and 'Jamnagar' is significantly different from 'Vadodara'. In other words, these locations, relative to 'Vadodara', do not show a significant difference in terms of their contribution to the variance in milk production explained by the model.*



## Does Satisfaction with Government play role in Y?

- F-test comparing two models: a full model that includes 'Satisfaction\_5\_7' and 'Satisfaction\_8\_10', and a reduced model that excludes these variables.
- The null hypothesis (H0) is that both coefficients are equal to zero (implying no significant difference from the baseline 'Satisfaction\_1\_4'), and the alternative hypothesis (H1) is that at least one of the coefficients is not equal to zero.
- F-Statistic: 1.406 P-Value: 0.237
- Interpretation in Context of Baseline: The test result indicates that, in terms of average daily milk production, there is no significant difference between farms with satisfaction levels 5-7 or 8-10 and those with satisfaction levels 1-4. In other words, these higher satisfaction levels, relative to the baseline (satisfaction levels 1-4), do not show a significant difference in their contribution to the variance in milk production explained by the model.

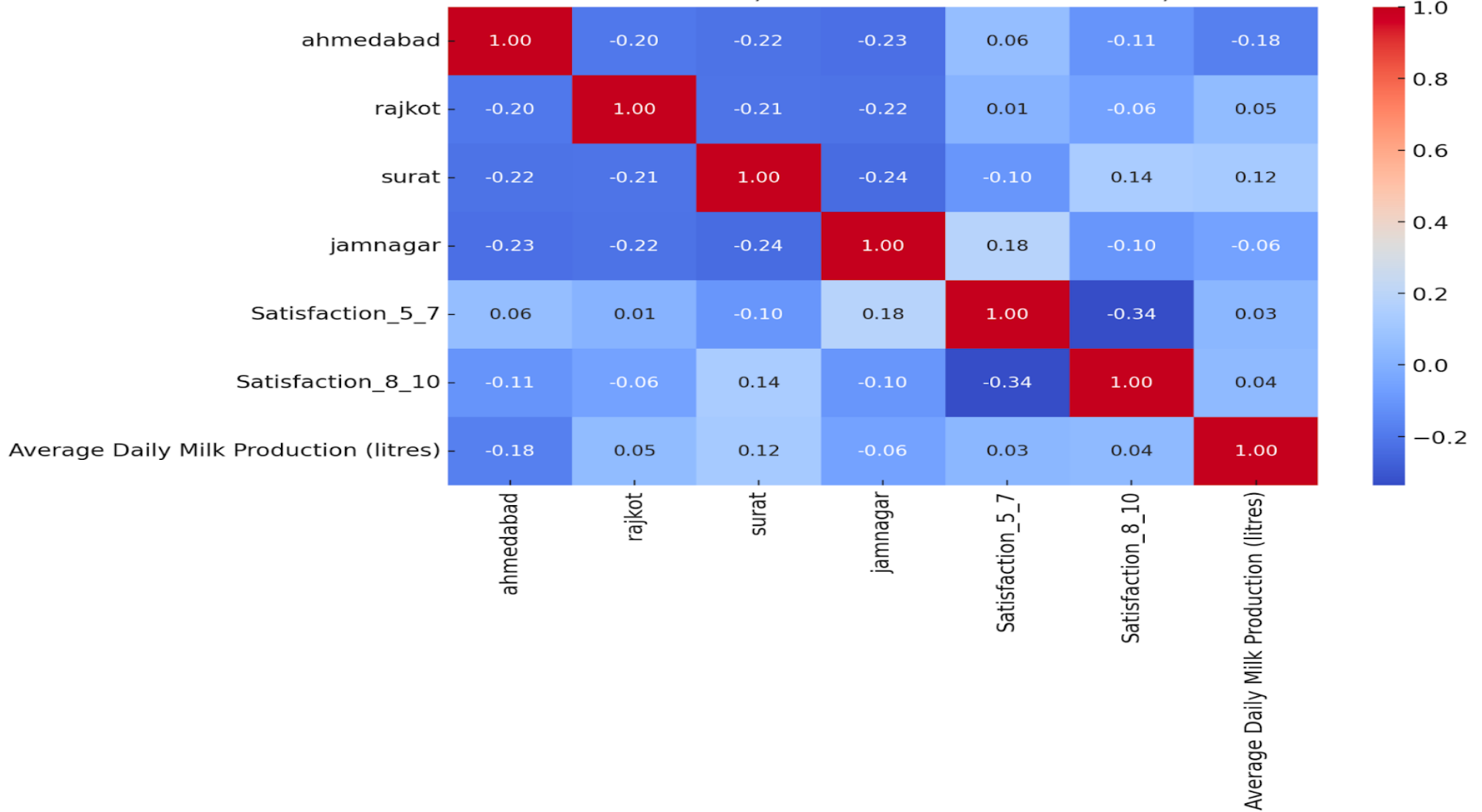
## Is milk production dependent on *Location and Satisfaction Level with government?*

- **Location Variables and Milk Production:** The weak correlations between the location variables ('Ahmedabad', 'Rajkot', 'Surat', 'Jamnagar') and daily milk production suggest that these locations, relative to the baseline 'Vadodara', do not have a strong linear relationship with milk production. This indicates that the impact of these specific locations on milk production is not significantly different from 'Vadodara'.
- **Government Satisfaction Categories:** The government satisfaction categories 'Satisfaction\_5\_7' and 'Satisfaction\_8\_10' also exhibit low correlation with daily milk production. This implies that, relative to the baseline category 'Satisfaction\_1\_4', these levels of satisfaction do not show a strong direct linear relationship with milk production.
- **Inter-Variable Correlation:** The low correlations among the location variables and between these variables and the satisfaction categories suggest that these variables do not strongly relate to each other in a linear manner.
- **Overall Model Interpretation:** The correlation analysis, in conjunction with the earlier F-test results, indicates that while these variables may have some influence, they do not appear to be major predictors of daily milk production in the dataset, especially when compared to the baseline categories ('Vadodara' for location and 'Satisfaction\_1\_4' for government satisfaction).

- “There is lot of difference between correlation and causation”

-> Correlation indicates a relationship, whereas Causation implies one variable directly influences another.

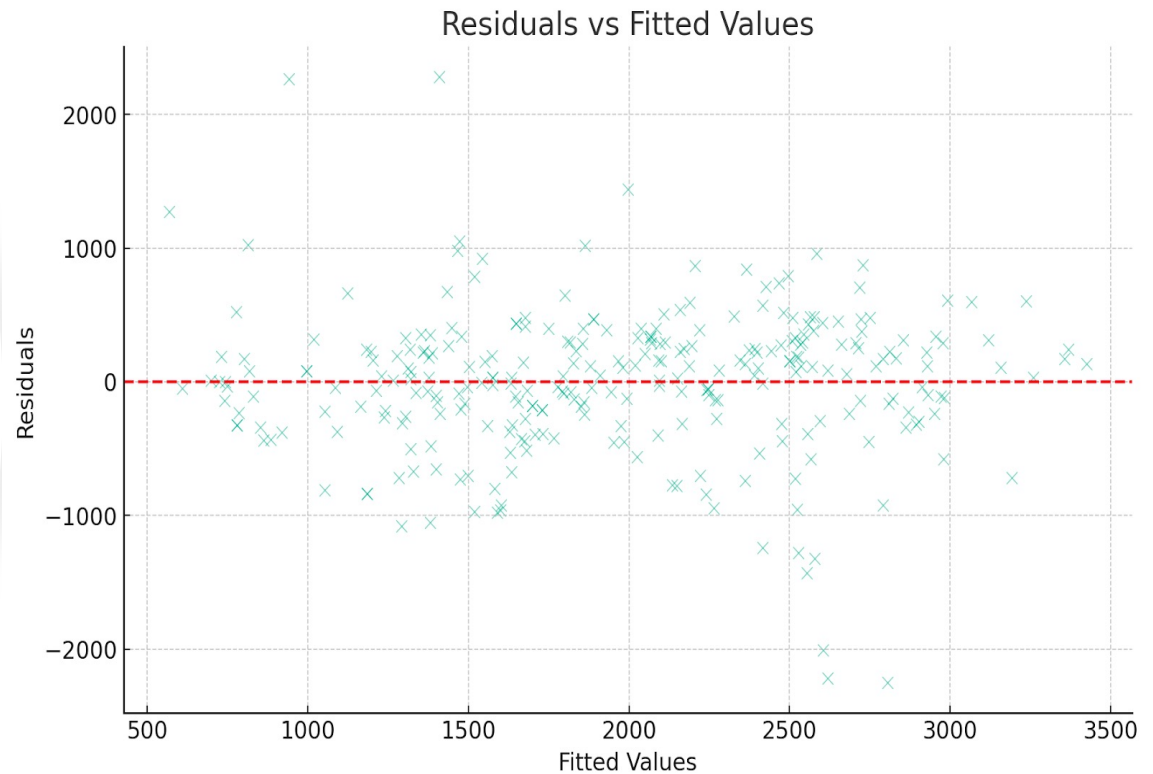
Correlation Matrix for Location, Government Satisfaction, and Milk Production



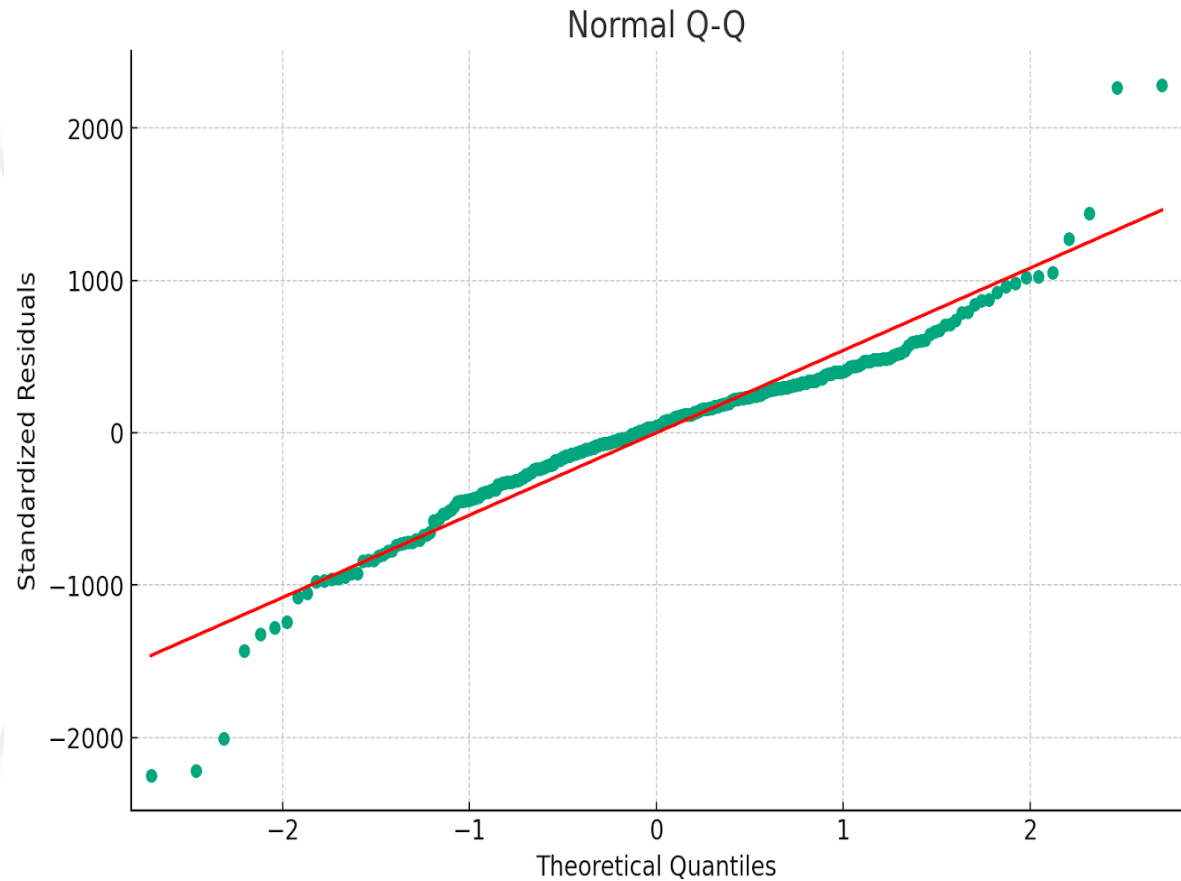


## Diagnostic Plot for Full Model

*The residuals do not appear to have a distinct pattern, and the spread of residuals seems consistent across the range of fitted values, suggesting that the assumption of constant variance holds.*



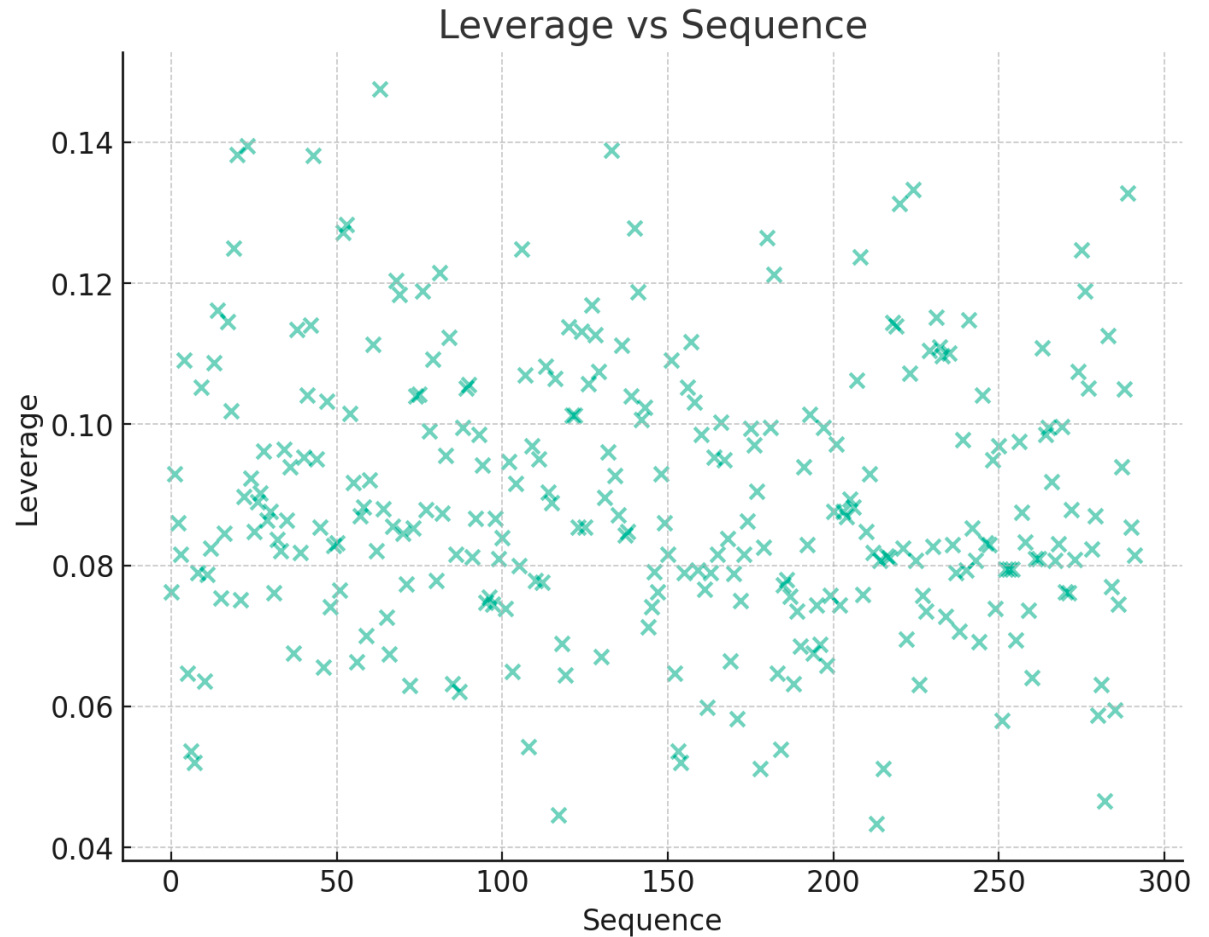
The QQ plot displays the standardized residuals against the theoretical quantiles of a normal distribution. In an ideal scenario, if the residuals are normally distributed, the points should lie approximately along the reference line. Deviations from this line indicate departures from normality. Here, we can see with some **presence of outliers** there is a **systematic up and down** w.r.t line so can assume that here Gaussian Assumption isn't holding.



The Leverage vs Sequence plot identifies observations with high leverage, which can have a disproportionate influence on the parameter estimates.

---

Observations with high leverage stand out from the rest of the data. The plot shows the leverage of each observation against its sequence in the dataset. There do not appear to be any points with excessively high leverage, meaning there may not be any particular observations unduly influencing the model.



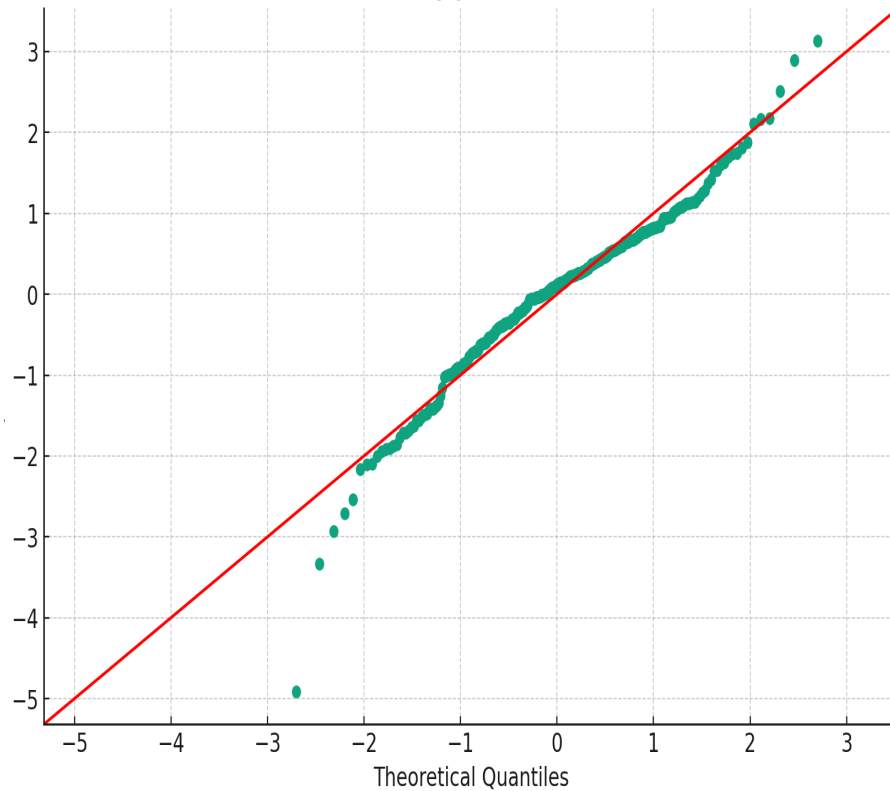
This plot shows the studentized residuals versus the fitted values. It is used to detect outliers in the Y-space ("regular" outliers). The red line at zero represents the expected value if the model's assumptions hold.

Points that deviate significantly from this line, especially those with studentized residuals greater than an absolute value of 4, can be considered outliers. In the plot generated, there are some points with high studentized residuals, suggesting the presence of outliers in the dataset.

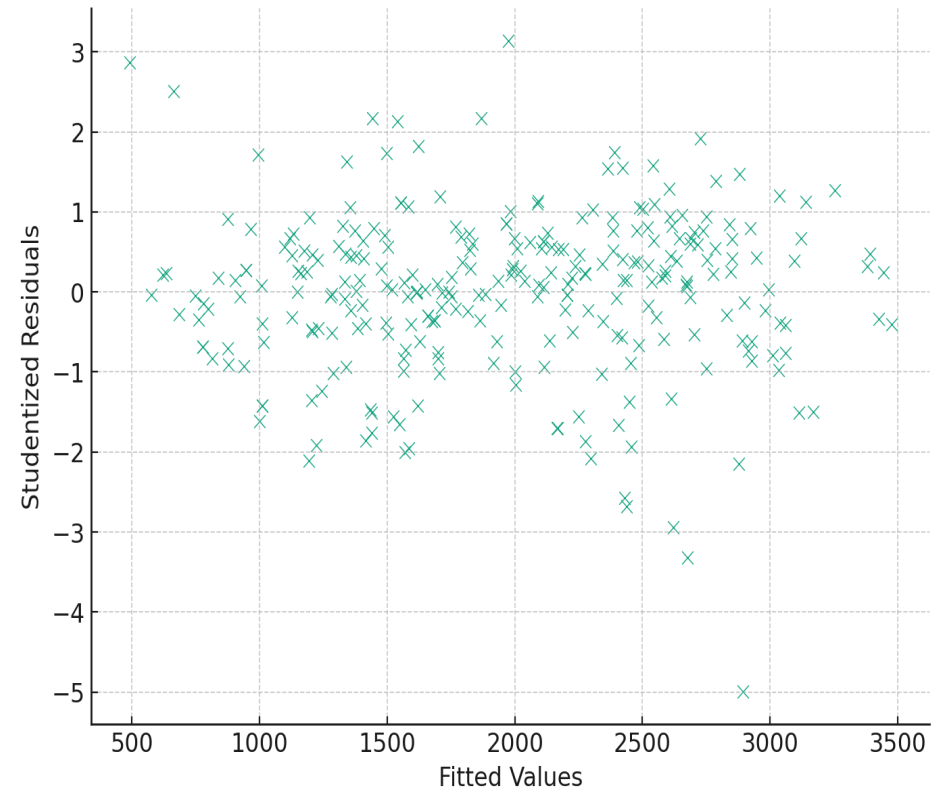


# After removing Outliers:

QQ Plot



Studentized Residuals vs Fitted Values



**Residuals vs Fitted Values still has 1 outlier and QQ plot has ignorable outliers but the constant up and down pattern is still concern**

# "Modern" Forward Selection with AIC method for building regression model

Start with the Null Model: This model includes no predictors and just the intercept. It serves as our baseline.

Iteratively Add Variables: For each step ( $k=1, \dots, p$ ), where  $p$  is the number of potential predictors, we'll identify the single variable that, when added to the model, reduces the AIC the most.

If adding the best candidate variable does not reduce the AIC, we'll stop the process.

The null model has been fitted, and its Akaike Information Criterion (AIC) is approximately 4762.48. This will serve as our baseline for comparison.

- > Identifying single variable reduces the AIC the most when added to this model.
- > The variable that reduces the Akaike Information Criterion (AIC) the most when added to the null model is "**Number of Cows.**" The AIC is reduced to approximately 4596.33, which is a significant decrease from the null model's AIC of 4762.48.
- Here's a summary of the model's results:
- Dependent Variable: Average Daily Milk Production (litres)
- Independent Variable: Number of Cows
- R-squared: 0.438, indicating that around 43.8% of the variability in milk production is explained by the number of cows.
- Coefficients: The intercept is around 1005.49, and the coefficient for "Number of Cows" is approximately 3.82. This suggests that each additional cow is associated with an increase of about 3.82 litres in average daily milk production.

The variable that further reduces the Akaike Information Criterion (AIC) the most when added to the current model (which already includes "Number of Cows") is "**Number of Buffaloes.**" The AIC decreases to approximately 4548.93 with the addition of this variable.

- Dependent Variable: Average Daily Milk Production (litres)
- Independent Variables: Number of Cows and Number of Buffaloes
- R-squared: 0.525, indicating that around 52.5% of the variability in milk production is explained by these two variables.
- Coefficients:
  - Intercept: ~614.48
  - Number of Cows: ~3.61 (each additional cow is associated with an increase of about 3.61 litres in average daily milk production)
  - Number of Buffaloes: ~2.74 (each additional buffalo is associated with an increase of about 2.74 litres in average daily milk production)



## Iterating through the process:

The final model now includes

"Number of Cows,"

"Number of Buffaloes,"

"Daily Expenditure on Animal Health (INR),"

"Dynamix dairy," "Daily Revenue (INR),"

"Daily Income from Selling Manure (INR)," and

"natural plants" as independent variables.

*The AIC for this model is approximately 4533.35, indicating a slight reduction from the previous model.*

Dependent Variable: Average Daily Milk Production (litres)

Independent Variables: Number of Cows, Number of Buffaloes, Daily Expenditure on Animal Health (INR), dynamix dairy, Daily Revenue (INR), Daily Income from Selling Manure (INR), natural plants

R-squared: 0.565, suggesting that about 56.5% of the variability in milk production is explained by these variables.

Coefficients: Intercept:  $\sim 263.49$

Number of Cows:  $\sim 3.13$  (each additional cow is associated with an increase of about 3.13 litres in average daily milk production)

Number of Buffaloes:  $\sim 2.41$  (each additional buffalo is associated with an increase of about 2.41 litres in average daily milk production)

Daily Expenditure on Animal Health (INR):  $\sim 1.58$  (suggesting a positive association between expenditure on animal health and milk production)

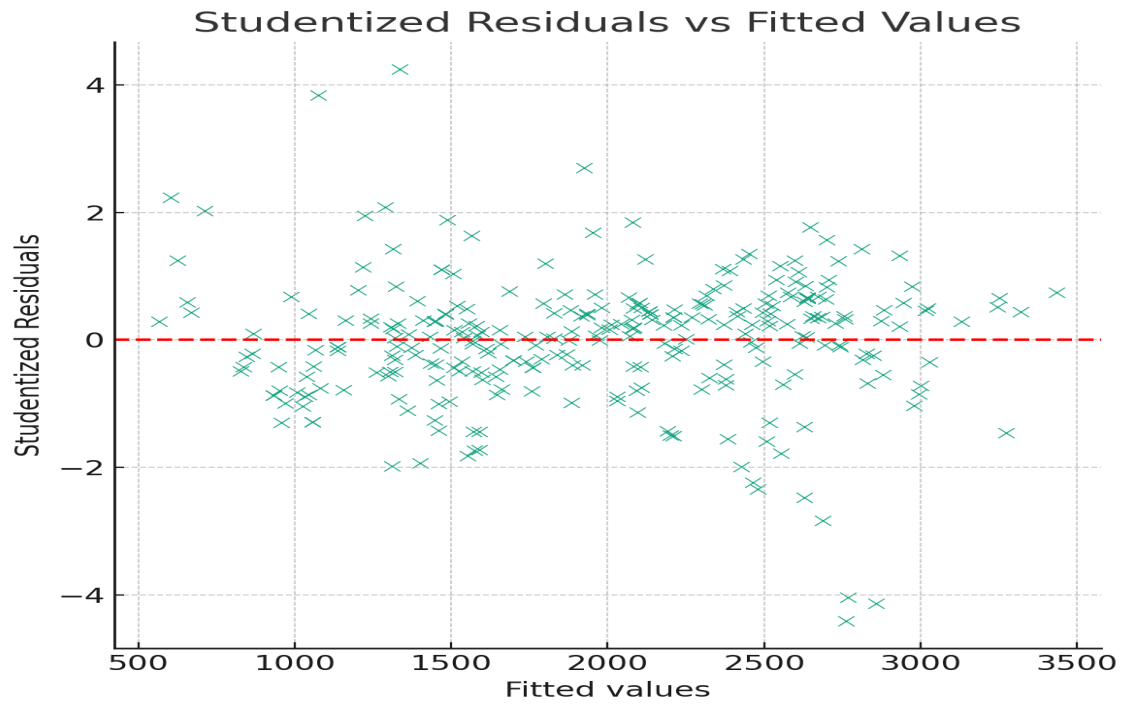
dynamix dairy:  $\sim -344.78$  (indicating a negative association with milk production, compared to the baseline category for milk collection centres)

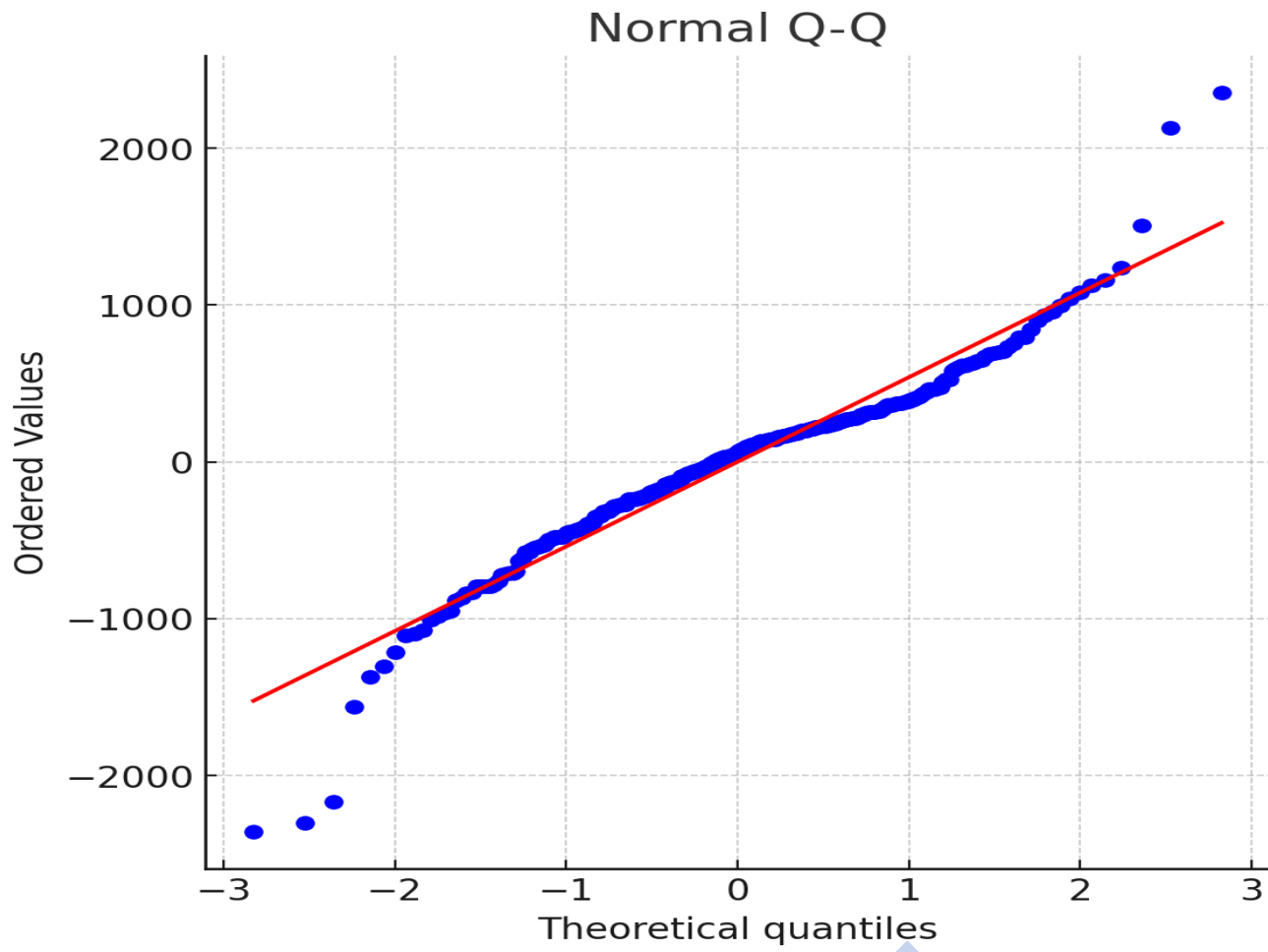
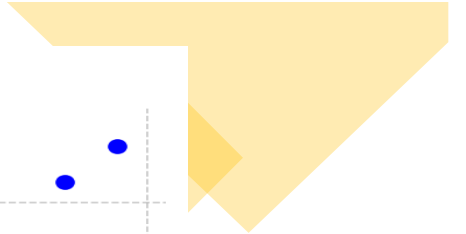
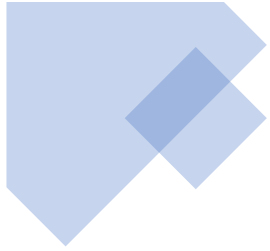
Daily Revenue (INR):  $\sim 0.14$  (indicating a positive association between daily revenue and milk production)

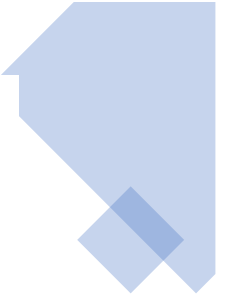
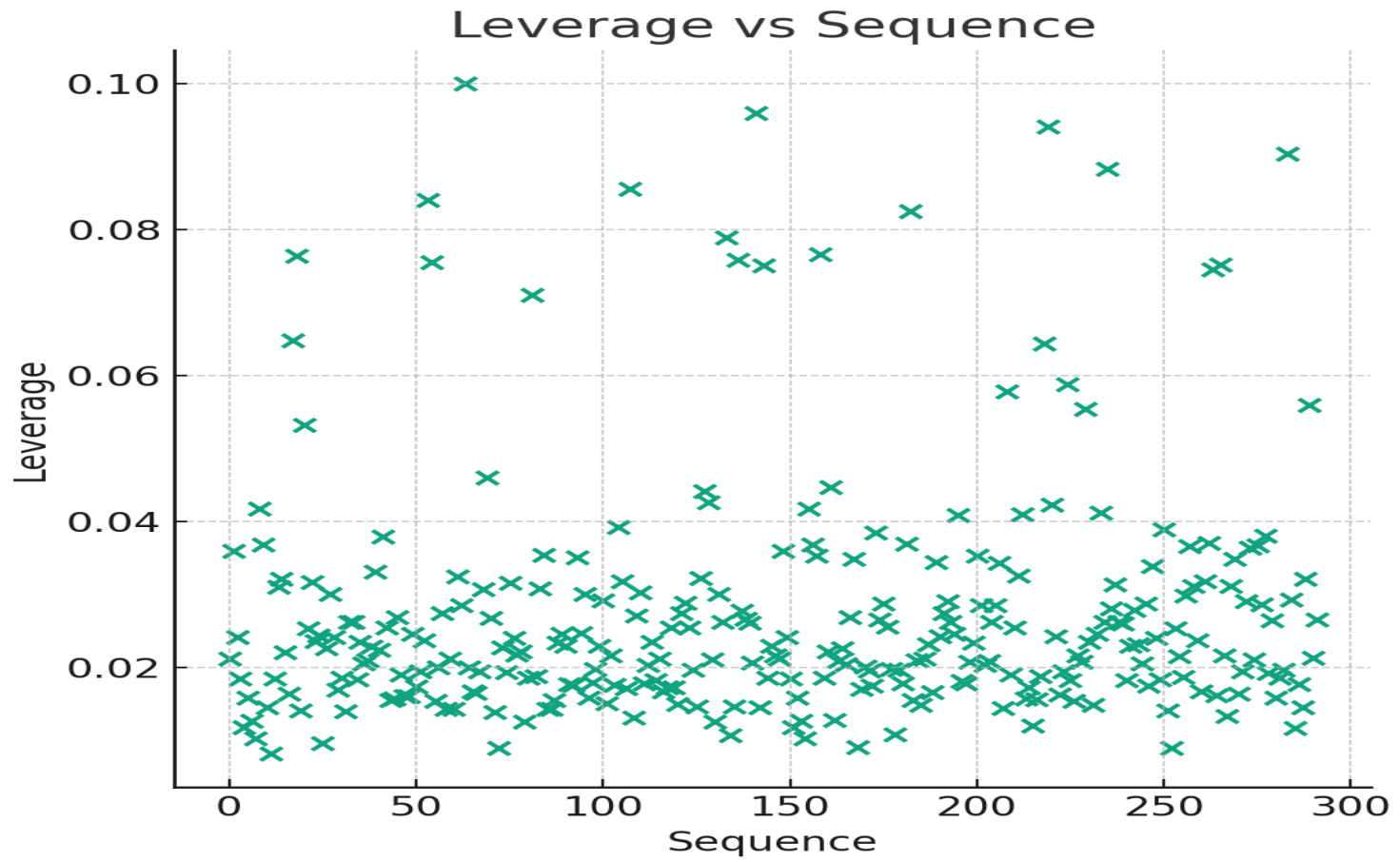
Daily Income from Selling Manure (INR):  $\sim -2.39$  (indicating a negative association between income from selling manure and milk production)

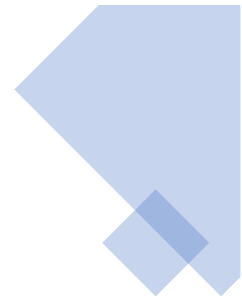
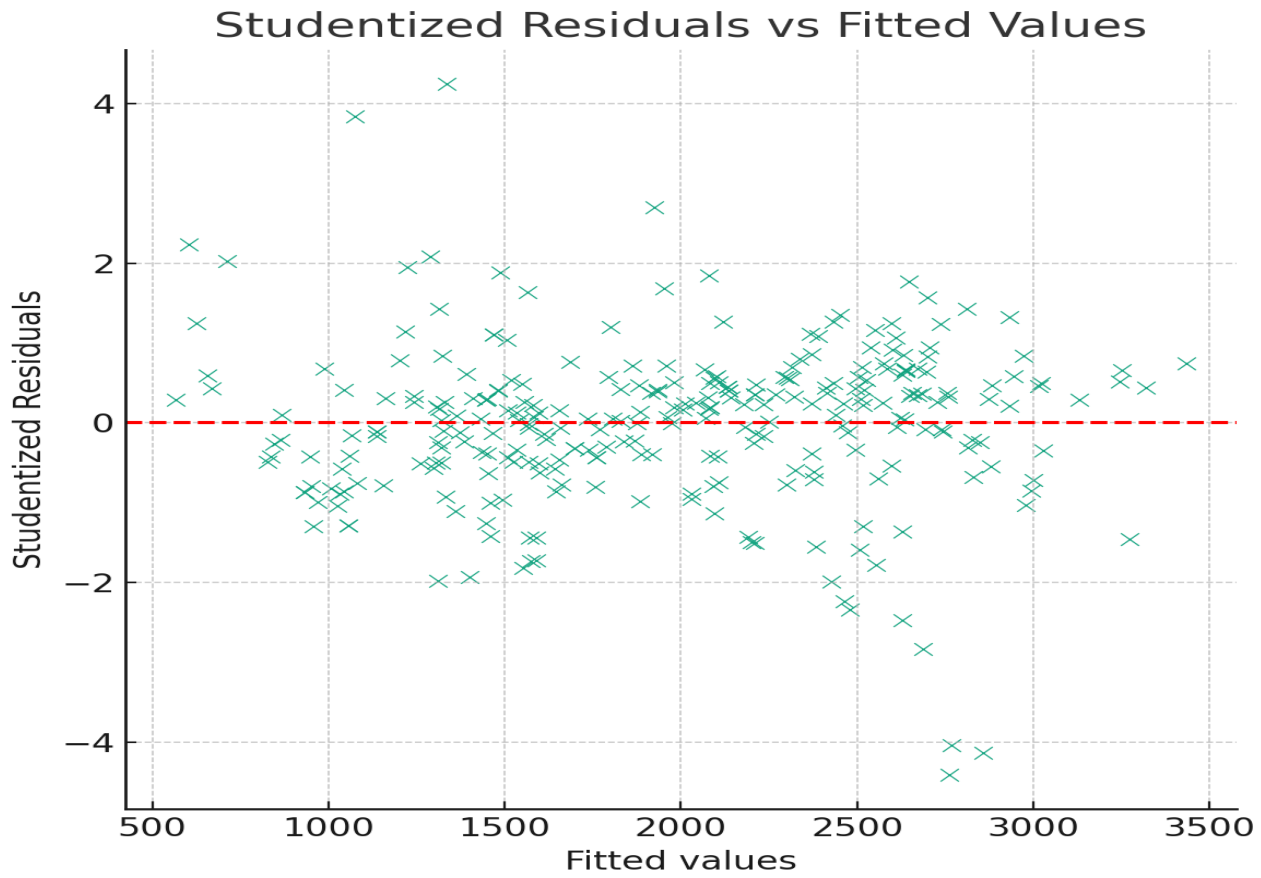
natural plants:  $\sim 119.72$  (indicating a positive association with milk production)

# Diagnostic Plots









## **Multicollinearity Analysis:**

Variance Inflation Factor (VIF) values for each predictor in the model:

Number of Cows: VIF = 1.84

Number of Buffaloes: VIF = 1.18

Daily Expenditure on Animal Health (INR): VIF = 1.05

dynamix dairy: VIF = 1.02

Daily Revenue (INR): VIF = 2.33

Daily Income from Selling Manure (INR): VIF = 1.25

natural plants: VIF = 1.01

“This indicates that each predictor provides unique information to the model, and the reliability of the regression coefficients is not adversely affected by multicollinearity.”

## ***Forward Selection Why?***

**Simplicity:** Begins with no variables and adds them one by one, making the model easier to understand.

**Reduces Overfitting:** By including fewer variables, it can prevent overfitting, especially in datasets with many features.

**Computationally Efficient:** More efficient for large datasets as it evaluates fewer models than the full model approach.

**Identifies Key Predictors:** Helps in identifying the most significant variables for the model.

**Good for Exploratory Analysis:** Useful in exploring which variables have the most predictive power.



- **Model Overview:** Model explains 56.5% of the variability in daily milk production, which is a significant portion considering the complexity of agricultural systems. Key factors include the number of cows and buffaloes, daily expenditure on animal health, revenue streams, and more.
- **Key Findings Livestock Counts Matter:**
  - Each additional cow and buffalo significantly boosts milk production by 3.13 and 2.41 liters, respectively.
  - **Investment in Animal Health Pays Off:** Increased spending on animal health positively correlates with higher milk production.
  - **Revenue Implications:** Daily revenue has a positive but modest impact on milk production.
  - **Interesting Insights:** The negative impact of income from selling manure suggests a possible trade-off between resource allocation for milk production vs. manure sales.
  - The negative association with 'dynamix dairy' indicates specific operational or environmental factors affecting production at these centers. Natural plants' positive impact suggests beneficial environmental or dietary factors.

- Implications for Dairy Farming Strategic Investment:
  - Encourage investment in livestock health and appropriate feed (including natural plants) to maximize milk production.
  - Balanced Resource Allocation: Consider the trade-offs in resource allocation, especially in terms of revenue generation from alternate sources like manure.

Thank you for your precious  
time!