Cairo University
Faculty of Economics and Political Science
Statistics Department

Determinants of Milk Production Efficiency: A Fixed Effects Panel Data Analysis of Spanish Dairy Farms

Prepared by
Ezz Eldin Ahmed Mohamed
Mohamed Amir Mohamed
Abdulrahman Mostafa Kamel

Under the supervision of

Dr. Marwah Sabry
Professor of Statistics

TA. Caroline Sherif
Assistant Lecturer

2025

Table of Contents

Introduction

Milk might appear to be a simple drink, but it is anything but simple. Milk is vital for health, the economy and society. Speaking of health, Milk is a rich source of calcium, vitamin D, protein, and other essential nutrients. These nutrients are crucial for bone health, supporting strong bones, teeth and potentially reducing the risk of osteoporosis. Milk can also contribute to overall health, supporting normal growth and development, releasing energy from foods, and boosting the immune system. Moving to Economy, Dairy farming is a significant economic activity, providing livelihoods for millions of people worldwide generating jobs both on and off-farm. Milk can be processed into various dairy products, creating a diverse range of food products such as various types of cheese and yogurt and it is involved in bakery industries. For society, milk plays a crucial role in cultural practices, being a staple food in many regions and traditions. It is often used in cooking and baking, contributing to culinary diversity. Milk can also be a source of cultural identity, representing a connection to rural heritage and farming communities.

Research Questions

RQ1. What is the impact of core inputs (COWS, LAND, LABOR, FEED) on milk production over time?

RQ2. Are there diminishing or increasing returns to scale in milk production?

RQ3. How does milk production get affected by cows on Land different units?

Data Description

This study utilizes a balanced panel dataset encompassing 247 Spanish dairy farms over a six-year period (1993–1998), resulting in a total of 1,482 observations. The dataset captures a wide array of variables, including the number of cows (Cows), land area in hectares (Land), milk output in liters (Milk), number of workers (Labor), and feed expenses (Feed). It also includes a log-transformed output variable (Yit) and log-transformed input variables (X1–X4), along with squared and interaction terms (X11, X12, …) to capture potential non-linear relationships and input complementarities. Dummy variables representing each year (Year93–Year98) are included to account for time-fixed effects, while TI indicates the number of observations per farm (which is consistently six, due to the balanced panel structure). The data originates from a study conducted by Antonio Álvarez and Luis Arias, focusing on Spanish dairy farm production.

Exploratory Data Analysis (EDA)

1. Descriptive Statistics

	Desc.	Min.	1st Quartile	Median	Mean	3rd Quartile	Max.
COWS	Number of cows in the farm	5	14.12	20	22.12	27	82.3
LAND	Land used for milk production (hectares)	2	8.5	12	12.99	16	45.10
MILK	Milk produced (in liters)	14410	68,569	110,236	131,107	163,262	727,281
LABOR	Number of workers (or labor units)	1	1	2	1.672	2	4
FEED	Feed expenses (in Spanish Pesetas)	3924	25,590	46,029	57,941	73,297	376,732

Table (1): Summary statistics of the dairy farms dataset

Descriptive Statistics Summary:

The dataset includes panel data from 247 farms observed over six years (1993–1998). Key variables reflect the scale and inputs of milk production:

- **COWS**: Ranges 5–82.3 (mean: 22.12, median: 20). Right-skewed, with decimals indicating averages; treated as continuous for the translog model.[1]
- **LAND (hectares)**: Ranges 2–45.1 (mean: 12.99, median: 12). Near-symmetric distribution; moderate land use for most farms.
- **MILK (liters)**: Ranges 14,410–727,281 (mean: 131,107, median: 110,236). Right-skewed, reflecting diverse farm productivity.
- **LABOR (workers)**: Ranges 1–4 (mean: 1.672, median: 2). Low variability; most farms use 1–2 workers, treated as continuous.
- **FEED (Pesetas)**: Ranges 3,924–376,732 (mean: 57,941, median: 46,029). Right-skewed; moderate expenses for most farms.

[1] In agricultural economics and production function analyses, variables like the number of livestock (COWS) are often modeled as continuous to capture the relationship between input quantities and output (MILK) more flexibly. This is standard in datasets like this, where the focus is on modeling production relationships rather than strictly counting discrete units.

Implications: Heterogeneity in inputs/output supports the translog model. Continuous inputs (COWS, LAND, LABOR, FEED) enable nonlinear analysis. Balanced panel aids farm- and time-specific effect estimation.

The variables were inspected for missing values, and none were detected, so no imputation or deletion was necessary.

These descriptive statistics reveal substantial heterogeneity in farm size, input use, and milk production, providing a rich dataset for panel data analysis of milk production efficiency and returns to scale.

---

2. Boxplots and Outlier treatment

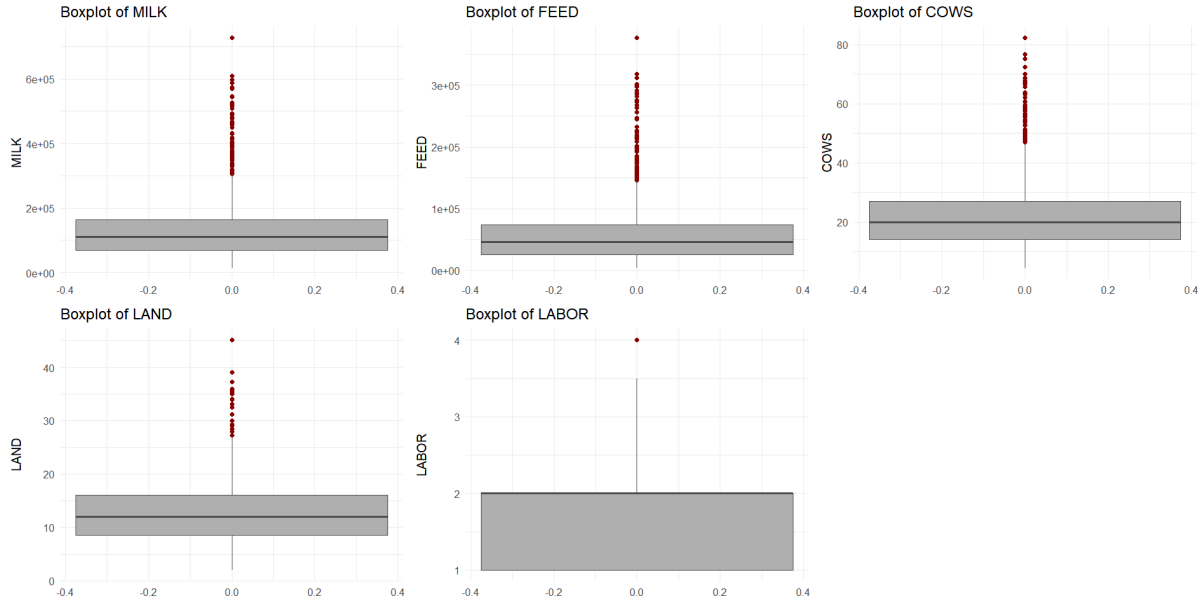

Figure (1): Boxplots of key variables

Boxplots reveal numerous clear outliers in four of the five key variables. To address this, further analysis using the interquartile range (IQR) method was conducted to identify and potentially handle these outliers.

Using the IQR method, the percentage of data points identified as outliers in each key variable is as follows:

- MILK: 80 points (5.40% of total)
- FEED: 78 points (5.26% of total)
- COWS: 60 points (4.05% of total)
- LAND: 58 points (3.91% of total)
- LABOR: 1 point (0.07% of total)

Given that the dataset had log-transformed variables, and the outliers represent real, high or low values, it's a good practice to assess outliers after transformation.

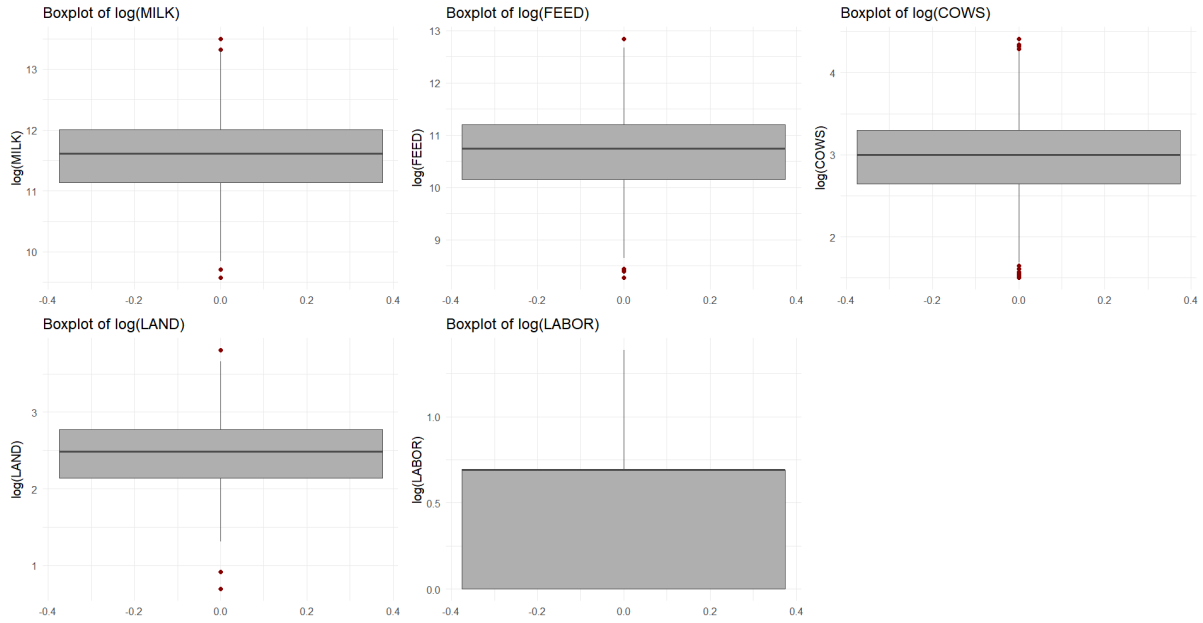

Figure (2): Boxplots of key variables

After applying log transformations to the key variables, the proportion of outliers was minimal, with the highest incidence being 0.88% for X1 (log of cows) and 0.47% for X2 (log of land). The dependent variable (milk production) exhibited just 0.27% outliers. Given the low proportions and the potential importance of extreme observations in capturing production dynamics within the panel data, it was decided to retain these points. Furthermore, robust statistical methods, such as heteroskedasticity-consistent standard errors and robustness checks, were employed to mitigate the potential influence of these outliers without sacrificing valuable information.

3. Histogram Plots

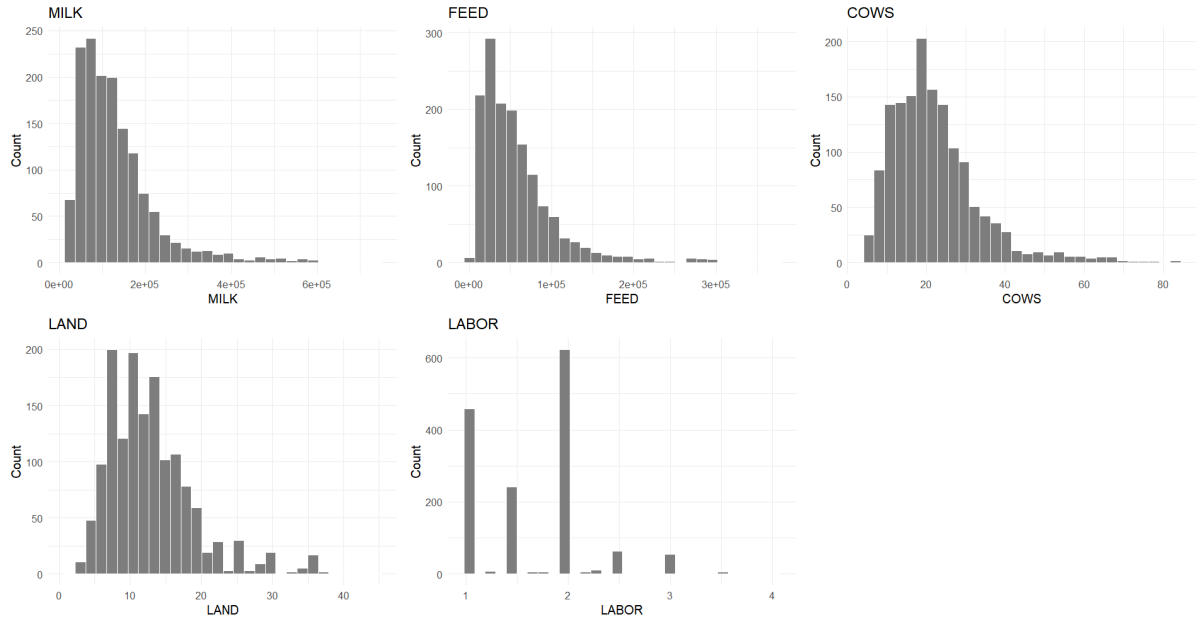

Figure (3): Histogram plots of key variables

The histogram plots reveal a clear right skewness for all variables. To address this skewness and approximate a more normal distribution, it is advisable to apply a logarithmic transformation to all variables.

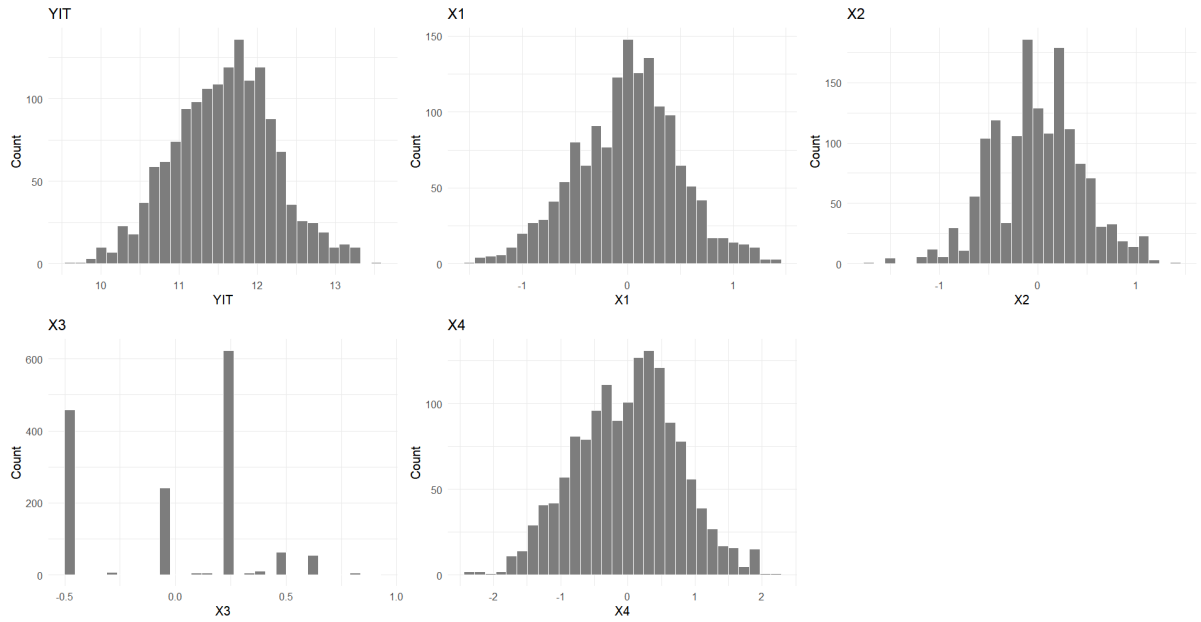

Figure (4): Histogram plots after logging key variables

After applying logarithmic transformations to the key variables, their distributions more closely approximate a normal distribution.

4. Correlation Analysis

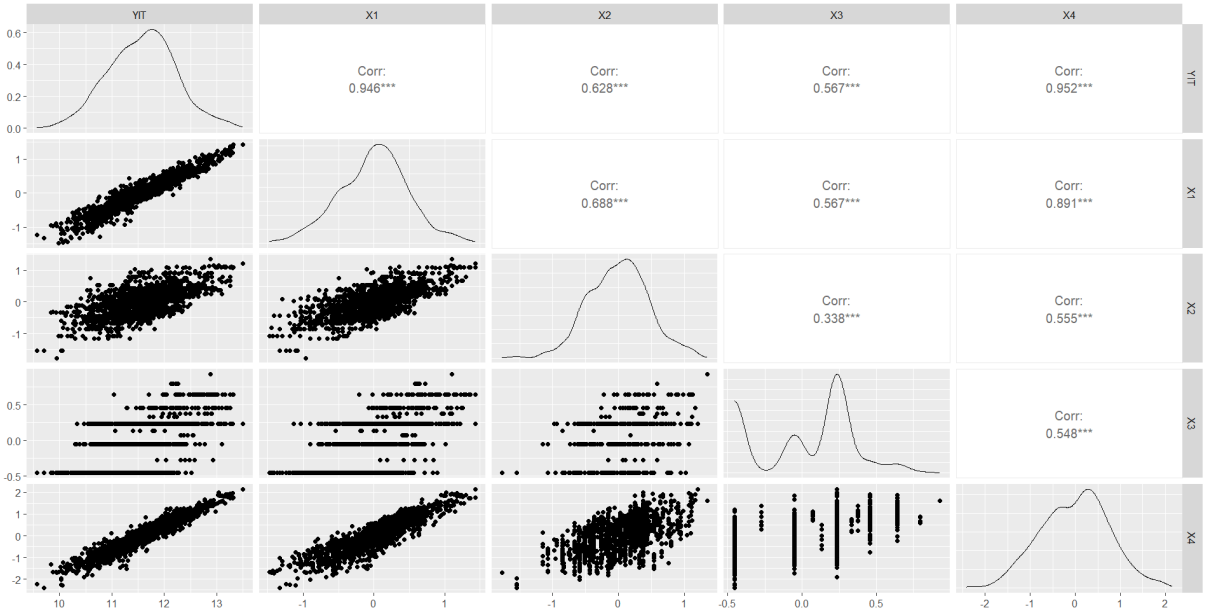

Figure (5): Correlation matrix visualization of long-transformed variables

The relationship between the explanatory variables and the dependent variable, *Milk*, reveals some compelling patterns:

- *Cows* show an exceptionally strong positive correlation with *Milk* production (r = 0.946), indicating that as the number of cows increases, milk output rises almost proportionally.

- Similarly, *Feed* demonstrates an even stronger positive link with *Milk* (r = 0.952), highlighting feed's critical role in driving milk yield.

- *Land* shares a solid positive correlation with *Milk* (r = 0.628), suggesting that larger land areas contribute meaningfully to milk production.

- *Labor* maintains a moderate positive relationship with *Milk* (r = 0.567), indicating workforce size has a noticeable, though less pronounced, impact.

Turning to correlations among the explanatory variables themselves:

- There is a very strong positive association between *Cows* and *Feed* (r = 0.891), reflecting that farms with more cows naturally require more feed.

- Moderate positive correlations exist between *Feed* and both *Land* (r = 0.561) and *Labor* (r = 0.506), hinting at interconnected resource dynamics.

- A strong positive relationship is also evident between *Cows* and *Land* (r = 0.702), showing that larger herds tend to be supported by larger farms.

- *Cows* and *Labor* exhibit a moderate positive correlation (r = 0.688), as more animals generally demand more labor.

- Finally, *Land* and *Labor* have a modest positive link (r = 0.338), reflecting that bigger farms usually require a greater workforce, though less tightly than other pa

---

Econometric Modeling

After exploring the data structure, we proceed to estimate models that quantify the effects of core inputs while accounting for farm-level and temporal variations.

RQ1: What is the impact of core inputs (COWS, LAND, LABOR, FEED) on milk production over time?

We specify the following Cobb-Douglas type production function (See Appendix C) in logarithmic form:

$$log(Milk_{it}) = \beta_0 + \beta_1 log(Cows_{it}) + \beta_2 log(Land_{it}) + \beta_3 log(Labor_{it}) + \beta_4 log(Feed_{it}) + u_{it}$$

Where: i indexes farms, t indexes time and u_{it} is the error term.

Model Selection: Pooled OLS vs. Random Effects Model

To assess the appropriate estimation strategy, we first tested whether a pooled Ordinary Least Squares (OLS) model adequately fits the data or whether panel-specific models better capture unobserved farm heterogeneity.

- **Breusch-Pagan Lagrange Multiplier Test for Random Effects:**

$$\chi^2 = 1582.2,\ df = 1,\ p < 2.2\times10^{-16}$$

The null hypothesis of no significant random effects is rejected, indicating that pooled OLS is not suitable and random effects should be considered.

Model Selection: Fixed Effects Model vs. Random Effects Model

- **Hausman Test:**

$$\chi^2 = 12.721,\ df = 4,\ p = 0.0127$$

Since $p < 0.0127$, we reject the null hypothesis that the random effects estimator is consistent, favoring the Fixed Effects Model (FEM) due to correlation between the regressors

and the individual effects.

Final Model Specification

Given the test results and consistent findings across models, the fixed effects model is preferred and specified as:

$$log(Milk_{it} - \overline{Milk_i}) = \beta_1 log(Cows_{it} - \overline{Cows_i}) + \beta_2 log(Land_{it} - \overline{Land_i}) + \beta_3 log(Labor_{it} - \overline{Labor_i}) + \beta_4 log(Feed_{it} - \overline{Feed_i}) + \mu_i$$

or equivalently:

$$log(Milk_{it}^{\star}) = \beta_1 log(Cows_{it}^{\star}) + \beta_2 log(Land_{it}^{\star}) + \beta_3 log(Labor_{it}^{\star}) + log(\beta_4 Feed_{it}^{\star}) + \mu_i$$

where μ_i is the demeaned error term.

Estimation Results

Using a balanced panel of 247 farms over 6 years (1,482 observations), the within (fixed effects) model yields:

```
Coefficients with Cluster-Robust Standard Errors:

t test of coefficients:

   Estimate Std. Error t value Pr(>|t|)
X1 0.662001   0.034367 19.2629  < 2e-16 ***
X2 0.037352   0.017360  2.1516  0.03162 *
X3 0.030399   0.024646  1.2334  0.21765
X4 0.382510   0.017334 22.0674  < 2e-16 ***
```

The model fit is strong with $Adjusted\ R^2 = 0.803$. The F-statistic is highly significant ($F(4, 1231) = 1568.11, p < 0.001$)

Model Diagnosis

Normality

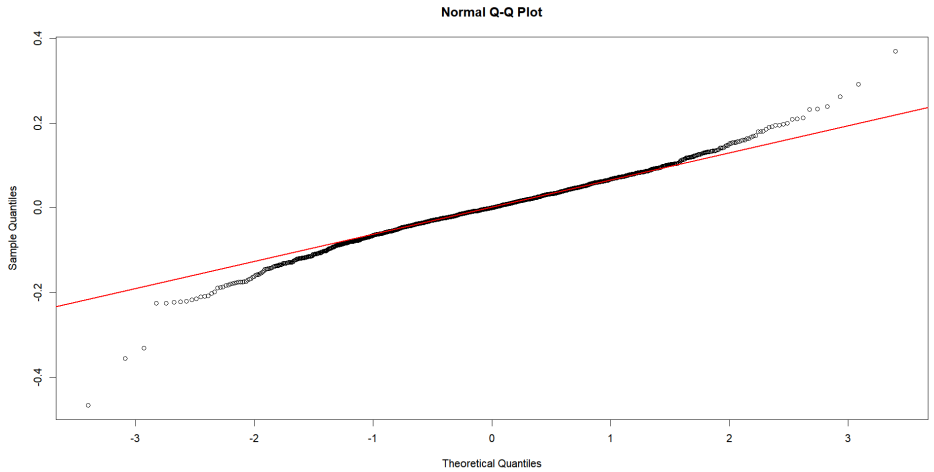

Figure (6): Q-Q plot of residuals

The Q-Q plot of residuals shows mild curvature at the tails, suggesting heavier tails than the normal distribution. However, with a large balanced panel of 247 farms and robust standard errors clustered at the farm level, this deviation is unlikely to materially affect inference.

Heteroskedasticity

- **Breusch-Pagan test for heteroskedasticity**

$$BP = 52.547,\ df = 6,\ p < 1.447\times10^{-9}$$

The test suggests rejecting the null hypothesis, indicating variance isn't constant across residuals.

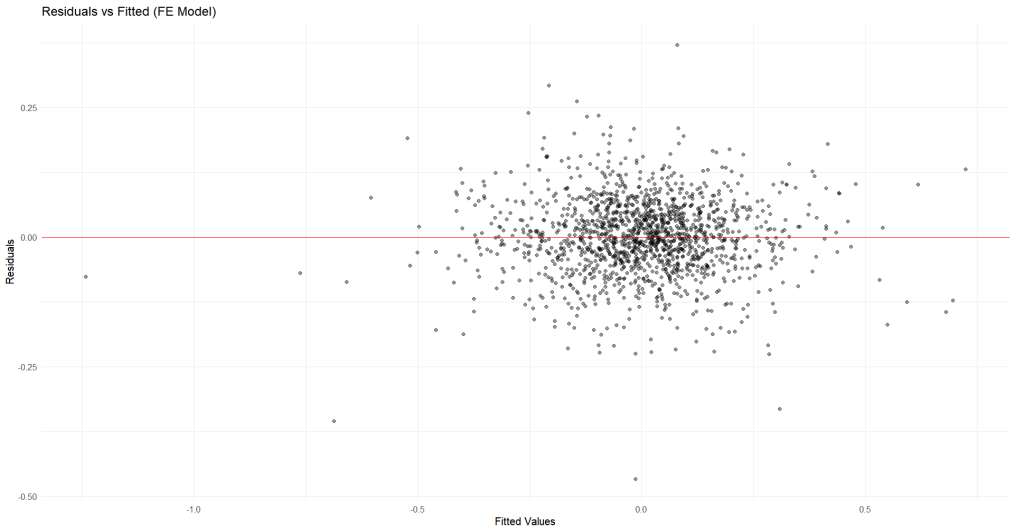

Figure (7): Residual vs fitted values of the FE Model

Serial Autocorrelation

- **Breusch-Godfrey/Wooldridge test**

$$\chi^2 = 191.81,\ df = 4,\ p < 2.2\times10^{-16}$$

The test indicated significant serial correlation within farms over time, suggesting that the model errors are temporally correlated. This is a panel-specific form of autocorrelation.

Robust standard errors clustered by farm were estimated using HC3 (See Appendix B) to account for heteroskedasticity and within-farm correlation, resulting in slightly more conservative inference without changing the overall significance of key predictors.

Multicollinearity:

A VIF value exceeding **10** is commonly considered indicative of severe multicollinearity (Kutner et al., 2005). In this case, all VIF values are below 7 for all variables, indicating acceptable multicollinearity levels.

For detailed R output, see Appendix E.

Interpretation

- The coefficients represent output elasticities: a 1% increase in the number of cows leads to approximately 0.66% increase in milk production, while a 1% increase in feed expenses leads to about 0.38% increase.

- Land has a small but statistically significant positive effect.

- Labor does not significantly contribute to explaining milk production variations in this dataset.

- The sum of the elasticities exceeds one ($RTS = 0.662 + 0.037 + 0.030 + 0.383 = 1.112$), indicating **increasing returns to scale** in milk production across farms.

RQ2. Are there diminishing or increasing returns to scale in milk production?

In RQ2, a Translog production function was utilized to explore the nature of returns to scale in milk production, allowing for the estimation of input interaction effects and non-linearities that extend beyond the Cobb-Douglas specification of RQ1.

Model Specification

The Translog function was specified as:

$$log(Milk_{it}) = \beta_0 + \beta_1 log(Cows_{it}) + \beta_2 log(Land_{it}) + \beta_3 log(Labor_{it}) + \beta_4 (Feed_{it})$$

$$+ \frac{1}{2}\beta_5 (Cows_{it})^2 + \frac{1}{2}\beta_6 (Land_{it})^2 + \frac{1}{2}\beta_7 (Labor_{it})^2 + \frac{1}{2}\beta_8 (Feed_{it})^2 + u_{it}$$

where μ_i is the demeaned error term

See Appendix (D) on the rationale of multiplying quadratic terms by a factor of $\frac{1}{2}$.

Model Selection: Pooled OLS vs. Random Effects Model

Once again, the lagrange-multiplier test is used to choose between pooled OLS or REM

- **Breusch-Pagan Lagrange Multiplier Test for Random Effects:**

$$\chi^2 = 1588.9,\ df = 1,\ p < 2.2\times10^{-16}$$

While the model appears to fit well, the Breusch-Pagan test indicates significant random effects ($p < 2.2\times10^{-16}$), suggesting that Pooled OLS is inappropriate.

Model Selection: Fixed Effects Model vs. Random Effects Model

- **Hausman Test:**

$$\chi^2 = 21.853,\ df = 6,\ p = 0.001288$$

Since $p < 0.05$, we reject the null hypothesis of consistent random effects. This supports the **Fixed Effects Model (FEM)**, suggesting that the individual effects correlate with the regressors.

Final Model Specification

We specify the within estimator model as follows after demeaning :

$$log(Milk_{it}^{\star}) = \beta_1 log(Cows_{it}^{\star}) + \beta_2 log(Land_{it}^{\star}) + \beta_3 log(Labor_{it}^{\star}) + log(\beta_4 Feed_{it}^{\star})$$

$$+ \frac{1}{2}\beta_5 (Cows_{it}^{\star})^2 + \frac{1}{2}\beta_6 (Land_{it}^{\star})^2 + \frac{1}{2}\beta_7 (Labor_{it}^{\star})^2 + \frac{1}{2}\beta_8 (Feed_{it}^{\star})^2 + u_{it}$$

Estimation Results

```
Coefficients with Cluster-Robust Standard Errors:

t test of coefficients:

     Estimate Std. Error t value Pr(>|t|)
X1   0.668284   0.035620 18.7614  < 2e-16 ***
X2   0.037382   0.016949  2.2055  0.02760 *
X3   0.020017   0.026350  0.7596  0.44761
X4   0.375652   0.017570 21.3799  < 2e-16 ***
X11  0.082498   0.063727  1.2945  0.19572
X22 -0.067067   0.068860 -0.9740  0.33027
X33 -0.120251   0.192680 -0.6241  0.53268
X44  0.048548   0.021682  2.2391  0.02533 *
```

The model shows strong fit with $Adjusted\ R^2 = 0.807$. The F-statistic is highly significant ($F(6, 1229) = 1077.71, p < 2.2e^{-16}$)

Model Diagnostics

Normality

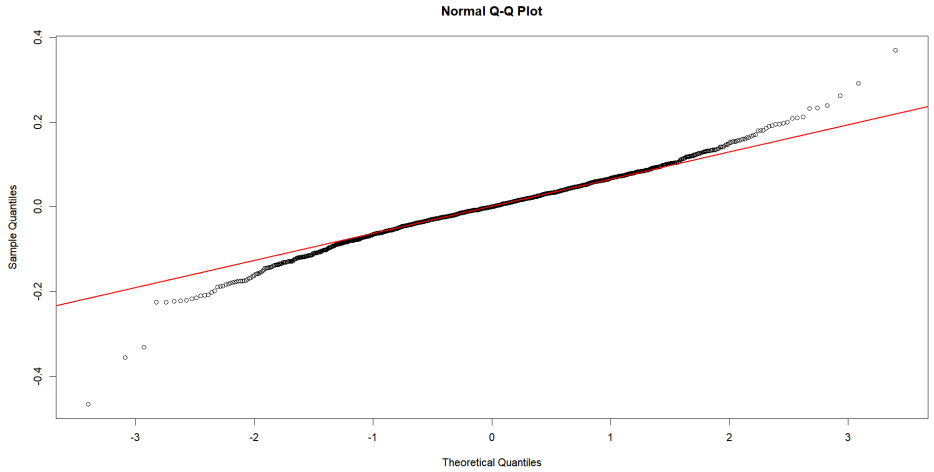

Figure (8): Q-Q plot of residuals

Once again, the Q-Q plot of residuals shows mild curvature at the tails, suggesting heavier tails than the normal distribution. However, with a large balanced panel of 247 farms and robust standard errors clustered at the farm level, this deviation is unlikely to materially affect inference.

Heteroskedasticity

- **Breusch-Pagan test for heteroskedasticity**

$$BP = 55.683,\ df = 8,\ p = 3.252e^{-09}$$

The test suggests rejecting the null hypothesis, indicating variance isn't constant across residuals.

Figure (9): Residual vs fitted values of the FE Model

Serial Autocorrelation

- **Breusch-Godfrey/Wooldridge test**

$$\chi^2 = 195.46,\ df = 6,\ p < 2.2\times10^{-16}$$

The test indicated significant serial correlation within farms over time, suggesting that the model errors are temporally correlated. This is a panel-specific form of autocorrelation.

Robust standard errors clustered by farm were estimated using HC3 (See Appendix B) to account for heteroskedasticity and within-farm correlation, resulting in slightly more conservative inference without changing the overall significance of key predictors.

Multicollinearity

A VIF value exceeding 10 is commonly considered indicative of severe multicollinearity. Since the largest VIF value is ~8.1 for X1, there's no indication of severe multicollinearity.

For detailed R output, see Appendix F.

Interpretation

- **Cows (X1):**
  The coefficient (0.668) is highly significant ($p < 0.001$), indicating that a 1% increase in the number of cows is associated with approximately a 0.67% increase in milk output, controlling for farm-specific effects. This is the strongest positive effect among inputs.

- **Land (X2):**
  Positive and significant at the 5% level (coefficient 0.037), suggesting modest but statistically significant returns to additional land. However, the quadratic term X22 is negative but not statistically significant after clustering (estimate -0.067), hinting at possible diminishing returns at higher land levels, though evidence is weaker here.

- **Labor (X3):**
  The coefficient (0.020) is small and not statistically significant, implying labor changes do not strongly affect milk output in this fixed effects framework.

- **Feed (X4):**
  Positive and highly significant (coefficient 0.376, $p < 0.001$). This shows that improvements in feed quantity/quality have a substantial positive effect on milk production.

Quadratic (Translog) Terms

- **X11 (Cows squared):** Positive coefficient (0.082), marginally significant (p ~ 0.07), suggesting a slight increasing returns to scale for cows, but this is not strongly conclusive.

- **X22 (Land squared):** Negative coefficient (-0.067), not significant after clustering, suggesting potential diminishing returns to land but with weak evidence.

- **X33 (Labor squared):** Negative but insignificant (-0.12), no clear nonlinear effect of labor on milk production.

- **X44 (Feed squared):** Positive and significant (0.049, p < 0.05), suggesting increasing returns to feed input. That is, as feed increases, milk output increases at an increasing rate.

RQ3. How does milk production get affected by cows on Land different units ?

The translog model extends RQ1 and RQ2, adding interaction terms (X12: COWS*LAND, X13: COWS*LABOR, X14: COWS*FEED, X23: LAND*LABOR, X24: LAND*FEED, X34: LABOR*FEED) to examine input complementarities.

Model Specification

The translog transformation was specified as:

$$log(Milk_{it}) = \beta_0 + \beta_1 log(Cows_{it}) + \beta_2 log(Land_{it}) + \beta_3 log(Labor_{it}) + \beta_4 log(Feed_{it}) + \beta_{12} log(Cows_{it}) log(Land_{it}) + \beta_{13} log(Cows_{it}) log(Labor_{it}) + \beta_{14} log(Cows_{it}) log(Feed_{it}) + \beta_{23} log(Land_{it}) log(Labor_{it}) + \beta_{24} log(Land_{it}) log(Feed_{it}) + \beta_{34} log(Labor_{it}) log(Feed_{it}) + u_{it}$$

Model Selection: Pooled OLS vs. Random Effects Model

Once again, the lagrange-multiplier test is used to choose between pooled OLS or REM

- **Breusch-Pagan Lagrange Multiplier Test for Random Effects:**

$$\chi^2 = 1564.4,\ df = 1,\ p < 2.2 \times 10^{-16}$$

While the model appears to fit well, the Breusch-Pagan test indicates significant random effects ($p < 2.2\times10^{-16}$), suggesting that Pooled OLS is inappropriate.

Model Selection: Fixed Effects Model vs. Random Effects Model

- **Hausman Test:**

$$\chi^2 = 23.158,\ df10,\ p = 0.01018$$

Since $p < 0.05$, we reject the null hypothesis of consistent random effects. This supports the **Fixed Effects Model (FEM)**, suggesting that the individual effects correlate with the regressors.

Final Model Specification

Given the test results and consistent findings across models, the fixed effects model is preferred and specified as:

$$log(Milk_{it}) = \beta_1 log(Cows_{it}) + \beta_2 log(Land_{it}) + \beta_3 log(Labor_{it}) + \beta_4 log(Feed_{it}) + \Sigma_{j<k} \beta_{jk} (log(X_j) \cdot log(X_k)) + \alpha_i + \varepsilon_{it}$$

Estimation Results

```
Coefficients with Cluster-Robust Standard Errors:


t test of coefficients:

      Estimate Std. Error t value  Pr(>|t|)
X1   0.6772535  0.0345641 19.5942 < 2.2e-16 ***
X2   0.0298500  0.0172440  1.7310 0.0836959 .
X3   0.0221909  0.0251495  0.8824 0.3777558
X4   0.3715507  0.0172923 21.4864 < 2.2e-16 ***
X12  0.0073343  0.0641251  0.1144 0.9089589
X13 -0.0123264  0.1131761 -0.1089 0.9132892
X14  0.0669212  0.0192307  3.4799 0.0005192 ***
X23  0.0210813  0.0509471  0.4138 0.6791015
X24 -0.0374950  0.0337923 -1.1096 0.2674012
X34  0.0291262  0.0471197  0.6181 0.5366022
```

The model fit is strong with $Adjusted\ R^2 = 0.8065$. The F-statistic is highly significant ($F(4, 1231) = 642.924,\ p < 2.22e^{-16}$)

Model Diagnostics

Normality

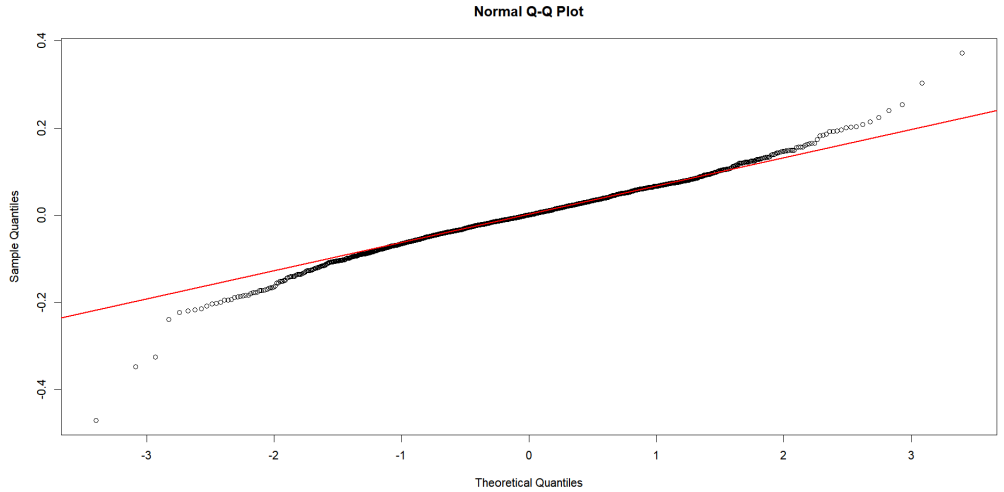

Figure (10): Residual vs fitted values of the FE Model

The Q-Q plot of residuals still shows mild curvature at the tails, suggesting heavier tails than the normal distribution. Usage of robust SEs still alleviates the non-normality of residuals.

Heteroskedasticity

- **Breusch-Pagan test for heteroskedasticity**

$$BP = 55.207,\ df = 8,\ p = 3.252e^{-09}$$

The test suggests rejecting the null hypothesis, indicating variance isn't constant across residuals.

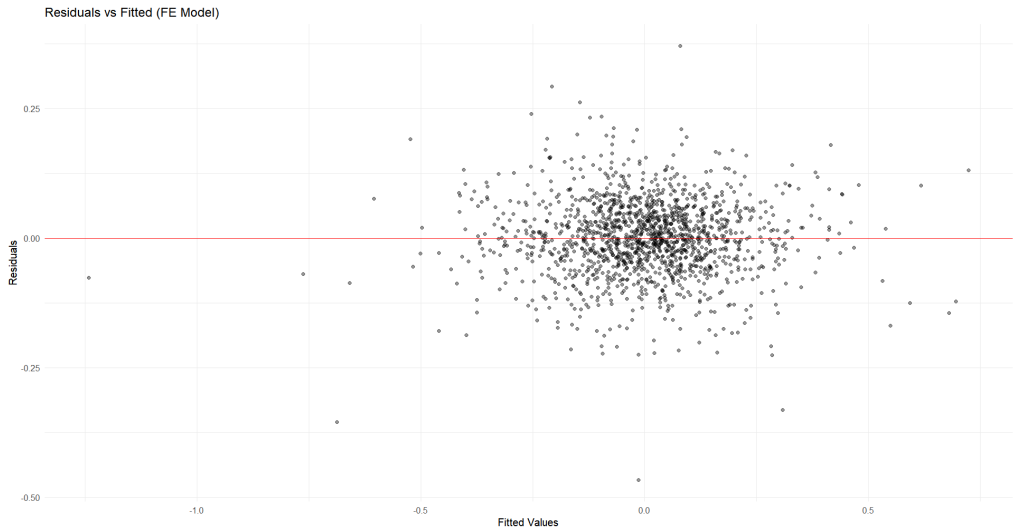

Figure (11): Residual vs fitted values of the FE Model

Serial Autocorrelation

- **Breusch-Godfrey/Wooldridge test**

$$\chi^2 = 191.84,\ df = 6,\ p < 2.2\times10^{-16}$$

The test indicated significant serial correlation within farms over time, suggesting that the model errors are temporally correlated. This is a panel-specific form of autocorrelation.

Robust standard errors clustered by farm were estimated using HC3 (See Appendix B) to account for heteroskedasticity and within-farm correlation, resulting in slightly more conservative inference without changing the overall significance of key predictors.

Multicollinearity

Since the largest VIF value is ~9.4 < 10 for X13, there's no indication of severe multicollinearity.

For detailed R output, see Appendix G.

Interpretation

Key Input Variables:

- **Cows (X1):** A one-cow increase leads to a significant ($p < 2.2e-16$) rise in milk production by 0.677 liters, indicating a strong positive impact.
- **Feed (X4):** Increased feed expenditure (in some monetary unit) significantly ($p < 2.2e-16$) boosts milk production by 0.372 liters, demonstrating a strong positive effect.
- **Land (X2):** A one-hectare increase in land area results in a marginally significant ($p = 0.084$) increase of 0.030 liters in milk production, suggesting a weak positive influence.
- **Labor (X3):** Changes in the number of workers do not significantly ($p = 0.378$) impact milk output in this model, as the coefficient (0.022 liters per worker) is not statistically significant.

Interaction Effects:

- **Cows × Feed (X14):** A significant ($p = 0.001$) positive interaction exists between the number of cows and feed expenses. This indicates a synergistic relationship where the effect of feed on milk production is enhanced by a larger number of cows, and conversely, the effect of having more cows on milk production is amplified with increased feed input.
- **Cows × Land (X12):** No significant interaction ($p = 0.909$) was found between the number of cows and land area, suggesting their effects on milk production are independent.
- **Cows × Labor (X13):** There is no significant interaction ($p = 0.913$) between the number of cows and labor, indicating their effects are independent.
- **Land × Labor (X23), Land × Feed (X24), Labor × Feed (X34):** None of these interaction terms showed statistical significance, suggesting no strong interactive effects between these input pairs.

Conclusions and Policy Recommendations

Our analysis of 247 Spanish dairy farms from 1993 to 1998 shows that milk production is primarily driven by the number of cows and feed expenses, with a 1% increase in either boosting output by about 0.66% or 0.38%, respectively. Land has a minor positive effect, while labor's impact is negligible. Farms exhibit increasing returns to scale, meaning larger operations are more efficient. The synergy between cows and feed further enhances output, and production grew over time, likely due to technological advances. Farmers should focus on expanding herds and optimizing feed quality, while policymakers could offer subsidies for feed or technology adoption to boost dairy efficiency.

Limitations

While this analysis provides valuable insights into the determinants of milk production efficiency among Spanish dairy farms, it is important to acknowledge several limitations that should be considered when interpreting the results:

1. **Unobserved Time-Varying Factors**
   The fixed effects model controls for unobserved farm-specific heterogeneity but does not account for time-varying factors such as weather conditions, technological advancements, input price changes, or policy shifts. These omitted factors may influence both input choices and milk production outcomes, potentially biasing elasticity estimates.

2. **Potential Dynamic Effects**
   The model assumes contemporaneous relationships between inputs and outputs. However, dairy production may involve dynamic adjustments, such as lagged effects of feed quality or changes in herd size. Future research could explore dynamic panel models, such as Arellano-Bond or system GMM, to better capture these temporal dependencies.

3. **Measurement Error in Variables**
   Some variables, including labor input, feed usage, and milk output, may be subject to measurement error due to reporting inaccuracies or aggregation issues. These errors could attenuate coefficient estimates, particularly if measurement error is systematic.

4. **Handling of Skewness and Outliers**
   Log transformations were applied to stabilize variance and reduce skewness, and robust standard errors were used to account for heteroscedasticity. However, a few influential observations remain, which may impact the distribution of residuals and exert leverage on coefficient estimates. Future studies could incorporate robust panel regression techniques or quantile regression to further address these concerns.

5. **Model Specification**
   The Cobb-Douglas and Translog functional forms provide a useful framework for interpreting input elasticities and interactions but rely on specific parametric assumptions. These forms may not fully capture nonlinear relationships or complex interactions among inputs. Nonparametric methods, such as kernel regression, or flexible machine learning techniques could be explored to complement the findings.

6. **Generalizability and Data Scope**
   The analysis focuses exclusively on Spanish dairy farms over a six-year period (1993–1998), which may limit the generalizability of the findings to other contexts, regions, or time periods. Structural changes in the dairy sector, technological developments, and market conditions since 1998 may affect the applicability of the

results today.

7. **Omitted Variables**
   The model does not include potentially relevant variables such as farm management practices, animal health, genetic factors, or environmental variables (e.g., pasture quality, rainfall). The exclusion of these factors could result in biased or incomplete estimates of production efficiency.

Future Research Directions

Future research could:

- Extend the analysis with dynamic panel models to account for temporal dependencies.
- Incorporate additional explanatory variables to capture a more comprehensive picture of farm operations.
- Explore quantile regression or robust regression methods to minimize sensitivity to influential observations.
- Analyze regional differences or farm size heterogeneity to understand context-specific effects.
- Apply nonparametric or machine learning models for greater flexibility in capturing complex data structures.

References

1. Kutner, M. H., Nachtsheim, C. J., Neter, J., & Li, W. (2005). *Applied Linear Statistical Models*. McGraw-Hill.
2. Long, J. S., & Ervin, L. H. (2000). Using heteroscedasticity consistent standard errors in the linear regression model. *The American Statistician*, 54(3), 217–224.
3. MacKinnon, J. G., & White, H. (1985). Some heteroskedasticity-consistent covariance matrix estimators with improved finite sample properties. *Journal of Econometrics*, 29(3), 305–325.

Appendix A: Hypothesis Testing Formulas

1) Hausman Test tests if individual random effect are correlated with the regressors

$$H_0 : \beta_F = \beta_R$$

$$H_1 : \beta_F \neq \beta_R$$

Test statistic: $H = (\beta_F - \beta_R)' (\Sigma_F - \Sigma_R)^{-1} (\beta_F - \beta_R) \sim \chi^2_P$

where F represents Fixed effect model (FEM) and R represents Random effect model (REM).

2) Breusch and Pegan Lagrange Multiplier Test

$$H_0 : \sigma^2_u = 0$$

$$H_1 : \sigma^2_u \neq 0$$

This follows $\chi^2_{(1)}$

3) Breusch-Pagan test for heteroskedasticity

$$H_0 : Homoscedasticity\ is\ present.$$

$$H_1 : Heteroscedasticity\ is\ present.$$

This follows $\chi^2_{(p)}$, $p\ is\ the\ number\ of\ predictors$

Homoscedasticity means that the residuals are distributed with equal variance.

4) Breusch-Godfrey/Wooldridge test

$$H_0 : There\ is\ no\ serial\ correlation\ of\ any\ order\ up\ to\ p.$$

$$H_1 : Serial\ correlation\ of\ any\ order\ up\ to\ p\ is\ present.$$

Appendix B: Clustered Standard Errors

Clustered standard errors are a type of robust standard error estimator designed to correct for intra-cluster correlation, which occurs when observations within the same cluster (e.g., farm) are correlated rather than independent.

Traditional OLS assumes that error terms are independently and identically distributed. However, when data are grouped—such as panel data with repeated observations for the same farms—this assumption is violated, leading to underestimated standard errors and inflated t-statistics, increasing the risk of Type I errors. Clustered standard errors adjust for this by accounting for correlations within clusters while maintaining independence across clusters.

In matrix form, the variance-covariance matrix of the coefficient estimates with clustered standard errors is:

$$Var_{Cluster}(\widehat{\beta}) = (X'X)^{-1}(\sum_{g=1}^{G} X_g'\widehat{u_g u_g}'X_g)(X'X)^{-1}$$

Where:

- g: index for each cluster (e.g., farm),
- G: total number of clusters,
- X_g: matrix of regressors for cluster g,
- $\widehat{u_g}$: vector of residuals for cluster g.

The adjustment permits arbitrary correlation of errors within each cluster while assuming independence across clusters, making inference more robust. The quality of the adjustment improves with a larger number of clusters.

Recommended by **White (1984)** and **Arellano (1987)**, clustered standard errors are particularly suitable for panel data and grouped cross-sectional data where intra-cluster dependence may bias traditional standard error estimates.

Appendix C: The Cobb-Douglas Production Function

The Cobb-Douglas production function is a widely used functional form in economics to represent the relationship between inputs and output in production processes. It takes the general form:

$$Y = A \cdot X_1^{\alpha} \cdot X_2^{\beta} \cdot \ldots \cdot X_n^{\gamma}$$

Where:

- Y is the total output (e.g., milk production),
- A is total factor productivity, capturing technology and efficiency,
- $X_1 \cdot X_2 \cdot \ldots \cdot X_n$ are input factors (e.g., labor, land, feed, number of cows),
- α, β, γ are output elasticities of the respective inputs, measuring the percentage change in output resulting from a 1% change in each input, holding other inputs constant.

Key features:

- Constant Elasticity of Substitution: The Cobb-Douglas function assumes a constant elasticity of substitution equal to 1 between inputs.
- Returns to Scale: The sum of the exponents $(\alpha + \beta + \gamma + \ldots)$ indicates the returns to scale:
    - If the sum equals 1, production exhibits constant returns to scale,
    - If greater than 1, increasing returns to scale,
    - If less than 1, decreasing returns to scale.

Log-linearization:

Taking natural logarithms on both sides transforms the multiplicative model into a linear form suitable for regression analysis:

$$ln(Y) = ln(A) + \alpha ln(X_1) + \beta ln(X_2) + \ldots + \gamma ln(X_n) + \epsilon$$

where ϵ is a random error term.

This log-linear form allows estimation of the elasticities (α, β, γ) using linear regression techniques.

Appendix D: Rationale for Multiplying Squared Terms by 0.5 in the Translog Function

In the translog (transcendental logarithmic) production function, second-order terms—including squares and cross-products of logarithmic inputs—are commonly included to capture curvature (i.e., interactions and nonlinearities). These second-order terms typically take the form:

$$\frac{1}{2}\beta_{ii}(lnX_i)^2$$

The inclusion of the $\frac{1}{2}$ factor is a standard convention in the literature for two primary reasons:

1. **Simplified Differentiation**:
   When taking derivatives of the translog function with respect to lnX_i (e.g., to calculate elasticities or marginal products), the $\frac{1}{2}$ factor ensures a clean cancellation of the constant that arises from the power rule:

$$\frac{d}{dlnX_i}\left(\frac{1}{2}\beta_{ii}(lnX_i)^2\right) = \beta_{ii}(lnXi_i)$$

   Without the $\frac{1}{2}$, the derivative would introduce an unnecessary factor of 2:

$$\frac{d}{dlnX_i}\left(\beta_{ii}(lnX_i)^2\right) = 2\beta_{ii}(lnXi_i)$$

2. **Symmetry and Consistency**:
   The translog function is designed to be a second-order Taylor series approximation of a general function, and symmetry across second-order terms is essential for this approximation to hold. Multiplying squared terms by $\frac{1}{2}$ ensures the function's second derivatives (curvature) align with the Taylor expansion. It also simplifies interpretation, making the contribution of β_{ii} to the elasticity and interaction effects consistent with the cross-product terms.

This convention is widely used in empirical economics and production modeling to ensure interpretability of the model results.

Appendix E: RQ1 Model Output

```
--- Running Pooled OLS ---
Pooling Model

Call:
plm(formula = formula, data = pdata, model = "pooling")

Balanced Panel: n = 247, T = 6, N = 1482

Residuals:
     Min.   1st Qu.    Median   3rd Qu.      Max.
-0.554019 -0.085341  0.010470  0.096479  0.583301

Coefficients:
              Estimate Std. Error   t-value Pr(>|t|)
(Intercept) 11.5774868  0.0036459 3175.5155  < 2e-16 ***
X1           0.5951756  0.0195833   30.3920  < 2e-16 ***
X2           0.0230501  0.0112227    2.0539  0.04016 *
X3           0.0231924  0.0130310    1.7798  0.07532 .
X4           0.4517578  0.0107847   41.8890  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Total Sum of Squares:    613.15
Residual Sum of Squares: 29.096
R-Squared:      0.95255
Adj. R-Squared: 0.95242
F-statistic: 7412.19 on 4 and 1477 DF, p-value: < 2.22e-16

--- Lagrange Multiplier Test for Random Effects ---

  Lagrange Multiplier Test - (Breusch-Pagan)

data:  formula
chisq = 1582.2, df = 1, p-value < 2.2e-16
alternative hypothesis: significant effects


--- Running Fixed Effects Model ---
Oneway (individual) effect Within Model

Call:
plm(formula = formula, data = pdata, model = "within")

Balanced Panel: n = 247, T = 6, N = 1482

Residuals:
       Min.     1st Qu.      Median     3rd Qu.        Max.
```

```
-0.46660674 -0.04125798  0.00012905  0.04517861  0.37014265

Coefficients:
   Estimate Std. Error t-value Pr(>|t|)
X1 0.662001   0.024678 26.8251  < 2e-16 ***
X2 0.037352   0.016133  2.3153  0.02076 *
X3 0.030399   0.023208  1.3099  0.19048
X4 0.382510   0.012017 31.8310  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Total Sum of Squares:    49.745
Residual Sum of Squares: 8.1611
R-Squared:      0.83594
Adj. R-Squared: 0.80262
F-statistic: 1568.11 on 4 and 1231 DF, p-value: < 2.22e-16

--- Running Random Effects Model ---
Oneway (individual) effect Random Effect Model
   (Swamy-Arora's transformation)

Call:
plm(formula = formula, data = pdata, model = "random")

Balanced Panel: n = 247, T = 6, N = 1482

Effects:
                  var std.dev share
idiosyncratic 0.00663 0.08142 0.336
individual    0.01309 0.11440 0.664
theta: 0.721

Residuals:
      Min.    1st Qu.     Median    3rd Qu.       Max.
-0.3728577 -0.0460565  0.0040983  0.0503395  0.4291820

Coefficients:
              Estimate Std. Error   z-value Pr(>|z|)
(Intercept) 11.5774868  0.0076015 1523.0448  < 2e-16 ***
X1           0.6502720  0.0208835   31.1381  < 2e-16 ***
X2           0.0300489  0.0133827    2.2454  0.02475 *
X3           0.0350700  0.0173829    2.0175  0.04364 *
X4           0.3995280  0.0108786   36.7259  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Total Sum of Squares:    93.611
Residual Sum of Squares: 9.8477
R-Squared:      0.8948
```

```
Adj. R-Squared: 0.89452
Chisq: 12563.2 on 4 DF, p-value: < 2.22e-16

--- Hausman Test (FE vs RE) ---

  Hausman Test

data:  formula
chisq = 12.721, df = 4, p-value = 0.01273
alternative hypothesis: one model is inconsistent


--- Model Diagnostics (Fixed Effects Model) ---

Breusch-Pagan Test for heteroskedasticity:

  studentized Breusch-Pagan test

data:  model_fe
BP = 46.396, df = 4, p-value = 2.037e-09


Breusch-Godfrey Test for serial correlation:

  Breusch-Godfrey/Wooldridge test for serial correlation in panel models

data:  formula
chisq = 194.52, df = 6, p-value < 2.2e-16
alternative hypothesis: serial correlation in idiosyncratic errors


Variance Inflation Factors (VIF):
      X1       X2       X3       X4
6.881934 1.968365 1.500692 5.050484

Coefficients with Cluster-Robust Standard Errors:

t test of coefficients:

   Estimate Std. Error t value Pr(>|t|)
X1 0.662001   0.034367 19.2629  < 2e-16 ***
X2 0.037352   0.017360  2.1516  0.03162 *
X3 0.030399   0.024646  1.2334  0.21765
X4 0.382510   0.017334 22.0674  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1


Model Formula: YIT ~ X1 + X2 + X3 + X4
```

```
Coefficients:
      X1       X2       X3       X4
0.662001 0.037352 0.030399 0.382510
```

Appendix F: RQ2 Model Output

```
--- Running Pooled OLS ---
Pooling Model

Call:
plm(formula = formula, data = pdata, model = "pooling")

Balanced Panel: n = 247, T = 6, N = 1482

Residuals:
     Min.   1st Qu.    Median   3rd Qu.      Max.
-0.586295 -0.083853  0.010729  0.095924  0.590496

Coefficients:
              Estimate Std. Error   t-value Pr(>|t|)
(Intercept) 11.5714640  0.0059244 1953.1731  < 2e-16 ***
X1           0.6092926  0.0212154   28.7194  < 2e-16 ***
X2           0.0155737  0.0114341    1.3620  0.17339
X3           0.0265089  0.0133442    1.9865  0.04716 *
X4           0.4451564  0.0115518   38.5356  < 2e-16 ***
X11          0.0719399  0.0381015    1.8881  0.05921 .
X22         -0.0859917  0.0282067   -3.0486  0.00234 **
X33          0.0702396  0.0746922    0.9404  0.34717
X44          0.0077932  0.0153102    0.5090  0.61082
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Total Sum of Squares:    613.15
Residual Sum of Squares: 28.777
R-Squared:      0.95307
Adj. R-Squared: 0.95281
F-statistic: 3738.99 on 8 and 1473 DF, p-value: < 2.22e-16

--- Lagrange Multiplier Test for Random Effects ---

  Lagrange Multiplier Test - (Breusch-Pagan)

data:  formula
chisq = 1589.6, df = 1, p-value < 2.2e-16
alternative hypothesis: significant effects


--- Running Fixed Effects Model ---
Oneway (individual) effect Within Model
```

```
Call:
plm(formula = formula, data = pdata, model = "within")

Balanced Panel: n = 247, T = 6, N = 1482

Residuals:
      Min.    1st Qu.     Median    3rd Qu.       Max.
-0.4711958 -0.0423664  0.0010046  0.0455690  0.3712847

Coefficients:
     Estimate Std. Error t-value  Pr(>|t|)
X1   0.668284   0.025694 26.0095 < 2.2e-16 ***
X2   0.037382   0.016451  2.2723  0.023242 *
X3   0.020017   0.024139  0.8292  0.407148
X4   0.375652   0.012148 30.9233 < 2.2e-16 ***
X11  0.082498   0.045346  1.8193  0.069113 .
X22 -0.067067   0.046752 -1.4345  0.151676
X33 -0.120251   0.136027 -0.8840  0.376859
X44  0.048548   0.015895  3.0543  0.002305 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Total Sum of Squares:    49.745
Residual Sum of Squares: 7.9313
R-Squared:      0.84056
Adj. R-Squared: 0.80756
F-statistic: 808.596 on 8 and 1227 DF, p-value: < 2.22e-16

--- Running Random Effects Model ---
Oneway (individual) effect Random Effect Model
   (Swamy-Arora's transformation)

Call:
plm(formula = formula, data = pdata, model = "random")

Balanced Panel: n = 247, T = 6, N = 1482

Effects:
                   var  std.dev share
idiosyncratic 0.006464 0.080399 0.329
individual    0.013211 0.114939 0.671
theta: 0.7254

Residuals:
     Min.   1st Qu.    Median   3rd Qu.      Max.
-0.381697 -0.044881  0.005033  0.049759  0.432149

Coefficients:
             Estimate Std. Error   z-value Pr(>|z|)
(Intercept) 11.569701   0.010145 1140.4082  < 2e-16 ***
```

```
X1           0.659622   0.021775   30.2933  < 2e-16 ***
X2           0.029313   0.013459    2.1779  0.02942 *
X3           0.033128   0.018017    1.8387  0.06595 .
X4           0.393447   0.011083   35.4991  < 2e-16 ***
X11          0.081787   0.040013    2.0440  0.04095 *
X22         -0.088023   0.036261   -2.4275  0.01521 *
X33         -0.052399   0.104049   -0.5036  0.61454
X44          0.035523   0.014552    2.4411  0.01464 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Total Sum of Squares:    92.225
Residual Sum of Squares: 9.5947
R-Squared:      0.89596
Adj. R-Squared: 0.8954
Chisq: 12685.6 on 8 DF, p-value: < 2.22e-16

--- Hausman Test (FE vs RE) ---

  Hausman Test

data:  formula
chisq = 19.896, df = 8, p-value = 0.01074
alternative hypothesis: one model is inconsistent


--- Model Diagnostics (Fixed Effects Model) ---

Breusch-Pagan Test for heteroskedasticity:

  studentized Breusch-Pagan test

data:  model_fe
BP = 55.683, df = 8, p-value = 3.252e-09


Breusch-Godfrey Test for serial correlation:

  Breusch-Godfrey/Wooldridge test for serial correlation in panel models

data:  formula
chisq = 195.46, df = 6, p-value < 2.2e-16
alternative hypothesis: serial correlation in idiosyncratic errors


Variance Inflation Factors (VIF):
      X1       X2       X3       X4      X11      X22      X33      X44
8.144043 2.060206 1.586803 5.842825 3.268764 1.429776 1.207098 2.630445

Coefficients with Cluster-Robust Standard Errors:
t test of coefficients:
```

```
     Estimate Std. Error t value Pr(>|t|)
X1   0.668284   0.035620 18.7614  < 2e-16 ***
X2   0.037382   0.016949  2.2055  0.02760 *
X3   0.020017   0.026350  0.7596  0.44761
X4   0.375652   0.017570 21.3799  < 2e-16 ***
X11  0.082498   0.063727  1.2945  0.19572
X22 -0.067067   0.068860 -0.9740  0.33027
X33 -0.120251   0.192680 -0.6241  0.53268
X44  0.048548   0.021682  2.2391  0.02533 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1


Model Formula: YIT ~ X1 + X2 + X3 + X4 + X11 + X22 + X33 + X44

Coefficients:
       X1        X2        X3        X4       X11       X22       X33       X44
 0.668284  0.037382  0.020017  0.375652  0.082498 -0.067067 -0.120251  0.048548
```

Appendix G: RQ3 Model Output

```
--- Running Pooled OLS ---
Pooling Model

Call:
plm(formula = formula, data = pdata, model = "pooling")

Balanced Panel: n = 247, T = 6, N = 1482

Residuals:
     Min.   1st Qu.    Median   3rd Qu.      Max.
-0.588858 -0.083821  0.010868  0.094306  0.596591

Coefficients:
              Estimate Std. Error   t-value  Pr(>|t|)
(Intercept) 11.5715680  0.0045683 2533.0362 < 2.2e-16 ***
X1           0.6194472  0.0207712   29.8224 < 2.2e-16 ***
X2           0.0131213  0.0115800    1.1331  0.257355
X3           0.0217640  0.0130273    1.6706  0.095005 .
X4           0.4418058  0.0112271   39.3516 < 2.2e-16 ***
X12         -0.0512480  0.0367136   -1.3959  0.162959
X13          0.1713232  0.0627644    2.7296  0.006416 **
X14          0.0188934  0.0163232    1.1575  0.247274
X23          0.0010964  0.0425795    0.0257  0.979461
X24         -0.0049622  0.0250776   -0.1979  0.843172
X34         -0.0550376  0.0355173   -1.5496  0.121452
---
```

```
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Total Sum of Squares:    613.15
Residual Sum of Squares: 28.665
R-Squared:      0.95325
Adj. R-Squared: 0.95293
F-statistic: 2999.46 on 10 and 1471 DF, p-value: < 2.22e-16

--- Lagrange Multiplier Test for Random Effects ---

  Lagrange Multiplier Test - (Breusch-Pagan)

data:  formula
chisq = 1564.4, df = 1, p-value < 2.2e-16
alternative hypothesis: significant effects


--- Running Fixed Effects Model ---
Oneway (individual) effect Within Model

Call:
plm(formula = formula, data = pdata, model = "within")

Balanced Panel: n = 247, T = 6, N = 1482

Residuals:
       Min.     1st Qu.      Median     3rd Qu.        Max.
-0.47084984 -0.04146630  0.00059321  0.04578499  0.37192349

Coefficients:
      Estimate Std. Error t-value  Pr(>|t|)
X1   0.6772535  0.0252133 26.8609 < 2.2e-16 ***
X2   0.0298500  0.0161623  1.8469    0.0650 .
X3   0.0221909  0.0237588  0.9340    0.3505
X4   0.3715507  0.0121913 30.4767 < 2.2e-16 ***
X12  0.0073343  0.0456679  0.1606    0.8724
X13 -0.0123264  0.0622914 -0.1979    0.8432
X14  0.0669212  0.0167134  4.0040 6.601e-05 ***
X23  0.0210813  0.0465382  0.4530    0.6506
X24 -0.0374950  0.0263523 -1.4228    0.1550
X34  0.0291262  0.0358943  0.8114    0.4173
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Total Sum of Squares:    49.745
Residual Sum of Squares: 7.9613
R-Squared:      0.83996
Adj. R-Squared: 0.80651
F-statistic: 642.924 on 10 and 1225 DF, p-value: < 2.22e-16
```

```
--- Running Random Effects Model ---
Oneway (individual) effect Random Effect Model
   (Swamy-Arora's transformation)

Call:
plm(formula = formula, data = pdata, model = "random")

Balanced Panel: n = 247, T = 6, N = 1482

Effects:
                   var  std.dev share
idiosyncratic 0.006499 0.080617  0.33
individual    0.013168 0.114751  0.67
theta: 0.7243

Residuals:
      Min.    1st Qu.     Median    3rd Qu.       Max.
-0.3750200 -0.0454037  0.0048129  0.0498793  0.4345431

Coefficients:
              Estimate Std. Error   z-value  Pr(>|z|)
(Intercept) 11.5628968  0.0084240 1372.6204 < 2.2e-16 ***
X1           0.6675546  0.0214692   31.0935 < 2.2e-16 ***
X2           0.0231918  0.0134576    1.7233  0.084830 .
X3           0.0334621  0.0175466    1.9070  0.056515 .
X4           0.3906669  0.0110589   35.3261 < 2.2e-16 ***
X12         -0.0168172  0.0398236   -0.4223  0.672811
X13          0.0286812  0.0575346    0.4985  0.618129
X14          0.0543714  0.0154849    3.5112  0.000446 ***
X23          0.0103717  0.0415902    0.2494  0.803069
X24         -0.0268262  0.0242704   -1.1053  0.269028
X34          0.0075693  0.0332961    0.2273  0.820164
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Total Sum of Squares:    92.568
Residual Sum of Squares: 9.6409
R-Squared:      0.89585
Adj. R-Squared: 0.89514
Chisq: 12652.9 on 10 DF, p-value: < 2.22e-16

--- Hausman Test (FE vs RE) ---

  Hausman Test

data:  formula
chisq = 23.158, df = 10, p-value = 0.01018
alternative hypothesis: one model is inconsistent
```

```
--- Model Diagnostics (Fixed Effects Model) ---

Breusch-Pagan Test for heteroskedasticity:

  studentized Breusch-Pagan test

data:  model_fe
BP = 55.207, df = 10, p-value = 2.888e-08


Breusch-Godfrey Test for serial correlation:

  Breusch-Godfrey/Wooldridge test for serial correlation in panel models

data:  formula
chisq = 191.84, df = 6, p-value < 2.2e-16
alternative hypothesis: serial correlation in idiosyncratic errors


Variance Inflation Factors (VIF):
      X1       X2       X3       X4      X12      X13      X14      X23      X24
X34
7.826706 2.118544 1.516223 5.533184 8.351421 9.441274 4.636802 3.644355 8.826090
6.580155

Coefficients with Cluster-Robust Standard Errors:

t test of coefficients:

      Estimate Std. Error t value  Pr(>|t|)
X1   0.6772535  0.0345641 19.5942 < 2.2e-16 ***
X2   0.0298500  0.0172440  1.7310 0.0836959 .
X3   0.0221909  0.0251495  0.8824 0.3777558
X4   0.3715507  0.0172923 21.4864 < 2.2e-16 ***
X12  0.0073343  0.0641251  0.1144 0.9089589
X13 -0.0123264  0.1131761 -0.1089 0.9132892
X14  0.0669212  0.0192307  3.4799 0.0005192 ***
X23  0.0210813  0.0509471  0.4138 0.6791015
X24 -0.0374950  0.0337923 -1.1096 0.2674012
X34  0.0291262  0.0471197  0.6181 0.5366022
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1


Model Formula: YIT ~ X1 + X2 + X3 + X4 + X12 + X13 + X14 + X23 + X24 + X34

Coefficients:
```

```
        X1          X2          X3          X4         X12         X13         X14
X23
 0.6772535   0.0298500   0.0221909   0.3715507   0.0073343 -0.0123264   0.0669212
0.0210813
       X24         X34
-0.0374950   0.0291262
```

Appendix H: R Script

```
# 1. INSTALL AND LOAD REQUIRED PACKAGES -------------------------------------
# Install packages if not already installed
install.packages(c("plm", "ggplot2", "dplyr", "GGally", "lmtest", "pbgtest",
"patchwork"))

# Load required libraries
library(plm)        # Panel data econometrics
library(ggplot2)    # Data visualization
library(dplyr)      # Data manipulation
library(GGally)     # Extended ggplot functionality
library(lmtest)     # Hypothesis testing for linear models
library(car)        # Companion to Applied Regression
library(sandwich)   # Robust covariance matrix estimators
library(patchwork) # Combine multiple ggplot plots
library(magrittr)
library(tidyr)

# Function to detect outliers using IQR
detect_outliers_iqr <- function(data, var) {
  Q1 <- quantile(data[[var]], 0.25, na.rm = TRUE)
  Q3 <- quantile(data[[var]], 0.75, na.rm = TRUE)
  IQR <- Q3 - Q1
  lower <- Q1 - 1.5 * IQR
  upper <- Q3 + 1.5 * IQR
  data$outlier <- ifelse(data[[var]] < lower | data[[var]] > upper, TRUE, FALSE)
  data
}

# 2. LOAD AND INSPECT DATA --------------------------------------------------

# Load dataset from CSV file
data <- read.csv("C:/Users/MSI/Downloads/dairy.csv")

# Quick check of data structure and summary statistics
```

```r
head(data)      # View first few rows
summary(data)   # Summary statistics for all variables

# Check for missing data in all columns
data %>%
  summarise(across(everything(), ~sum(is.na(.)))) %>%
  pivot_longer(everything(), names_to = "Variable", values_to = "NA_Count")


# 3. DATA PREPARATION ------------------------------------------------------

# Convert YEAR to factor for proper categorical treatment
data$YEAR <- as.factor(data$YEAR)

# Define key variables for analysis
key_vars <- c("MILK", "FEED", "COWS", "LAND", "LABOR")
log_key_vars <- c("YIT","X1", "X2", "X3", "X4")

# 4. EXPLORATORY DATA ANALYSIS ---------------------------------------------
## 4.1 Pairwaise Correlation Matrix
ggpairs(data[log_key_vars])

## 4.1 Histograms of Key Variables ----
# Create histograms for all key variables
plot_list <- lapply(key_vars, function(var) {
  ggplot(data, aes_string(x = var)) +
    geom_histogram(bins = 30, fill = "grey50", color = "white") +
    labs(title = var, x = var, y = "Count") +
    theme_minimal()
})

plot_list_log <- lapply(log_key_vars, function(var) {
  ggplot(data, aes_string(x = var)) +
    geom_histogram(bins = 30, fill = "grey50", color = "white") +
    labs(title = var, x = var, y = "Count") +
    theme_minimal()
})

# Combine histograms using patchwork (3 columns)
wrap_plots(plotlist = plot_list, ncol = 3)
wrap_plots(plotlist = plot_list_log, ncol = 3) # Log variables

## 4.2 Boxplots of Key Variables ----
```

```
# Boxplots of original variables
boxplot_list <- lapply(key_vars, function(var) {
  ggplot(data, aes_string(y = var)) +
    geom_boxplot(fill = "gray70", color = "grey30", outlier.colour = "darkred") +
    labs(title = paste("Boxplot of", var), y = var) +
    theme_minimal()
})

# Boxplots of log-transformed variables
boxplot_list_log <- lapply(key_vars, function(var) {
  ggplot(data, aes_string(y = paste0("log(", var, ")"))) +
    geom_boxplot(fill = "gray70", color = "grey30", outlier.colour = "darkred") +
    labs(title = paste("Boxplot of log(", var, ")", sep = ""), y = paste("log(", var,
")", sep = "")) +
    theme_minimal()
})

# Combine and show boxplots in 3 columns
wrap_plots(plotlist = boxplot_list, ncol = 3)
wrap_plots(plotlist = boxplot_list_log, ncol = 3)

## 4.3 Yearly Boxplots ----
# Create boxplots showing distribution by year
boxplot_list <- lapply(key_vars, function(var) {
  ggplot(data, aes(x = factor(YEAR), y = var)) +
    geom_boxplot(fill = "gray70", color = "grey30", outlier.colour = "darkred") +
    labs(title = paste("Boxplot of", var), y = var) +
    theme_minimal()
})

# Combine yearly boxplots (3 columns)
wrap_plots(plotlist = boxplot_list, ncol = 3)

## 4.4 Time Series Plots ----
# Create line plots showing average values over time
lineplot_list <- lapply(key_vars, function(var) {
  data %>%
    group_by(YEAR) %>%
    summarise(avg_value = mean(.data[[var]], na.rm = TRUE)) %>%
    ggplot(aes(x = YEAR, y = avg_value, group = 1)) +
    geom_line(color = "grey70", size = 1) +
    geom_point(color = "steelblue") +
    labs(title = paste("Average", var, "Over Time"),
```

```
         x = "Year", y = paste("Average", var)) +
    theme_minimal()
})

# Combine time series plots (3 columns)
wrap_plots(plotlist = lineplot_list, ncol = 3)

# 5. PANEL DATA ANALYSIS ----------------------------------------------------

## 5.1 Prepare Panel Data Structure ----
# Convert to pdata.frame with FARM and YEAR as panel indices
pdata <- pdata.frame(data, index = c("FARM", "YEAR"))

## 5.2 Time Trend Visualization ----
# Plot average milk production over time
data %>%
  group_by(YEAR) %>%
  summarise(mean_YIT = mean(YIT, na.rm=TRUE)) %>%
  ggplot(aes(x=YEAR, y=mean_YIT)) +
  geom_line(group=1) + geom_point() +
  labs(title="Average Log Milk Production Over Years")

## 5.3 Correlation Analysis ----
# Create correlation plot of input variables
ggpairs(data[,c("X1","X2","X3","X4")])

for (var in log_key_vars) {
  outliers <- detect_outliers_iqr(data, var)
  n_outliers <- sum(outliers$outlier)
  total <- nrow(data)
  pct_outliers <- round(100 * n_outliers / total, 2)

  cat(sprintf("Outliers in %s: %d data points (%.2f%% of total)\n", var, n_outliers,
pct_outliers))
}
# 6. MODEL ESTIMATION and Dignosis
-------------------------------------------------------

# Define translog production function formula

run_panel_models <- function(formula, pdata) {
  #-----------------------------------------------------------------------------
  # Function: run_panel_models
```

```r
# Purpose:  Run and compare panel data models (Pooled OLS, FE, RE) with diagnostics
# Inputs:   formula - model formula
#           pdata   - panel data frame
# Output:   Returns fixed effects model object
#------------------------------------------------------------------------
# ====================== MODEL ESTIMATION ================================

# 1. POOLED OLS MODEL
cat("\n--- Running Pooled OLS ---\n")
model_pooled <- plm(formula, data = pdata, model = "pooling")
print(summary(model_pooled))

# 2. RANDOM EFFECTS TEST (Breusch-Pagan LM test)
cat("\n--- Lagrange Multiplier Test for Random Effects ---\n")
print(plmtest(model_pooled, type = "bp"))

# 3. FIXED EFFECTS MODEL
cat("\n--- Running Fixed Effects Model ---\n")
model_fe <- plm(formula, data = pdata, model = "within")
print(summary(model_fe))

# 4. RANDOM EFFECTS MODEL
cat("\n--- Running Random Effects Model ---\n")
model_re <- plm(formula, data = pdata, model = "random")
print(summary(model_re))

# 5. HAUSMAN TEST (FE vs RE comparison)
cat("\n--- Hausman Test (FE vs RE) ---\n")
print(phtest(model_fe, model_re))

# ===================== MODEL DIAGNOSTICS ================================
cat("\n--- Model Diagnostics (Fixed Effects Model) ---\n")

# 1. RESIDUAL ANALYSIS
fitted_values <- fitted(model_fe)
residuals <- resid(model_fe)

# 1.1 Normality check (QQ plot)
qqnorm(residuals)
qqline(residuals, col = "red", lwd = 2)

# 1.2 Residuals vs fitted plot
print(
```

```
    ggplot(data.frame(fitted = fitted_values, residuals = residuals),
         aes(x = fitted, y = residuals)) +
      geom_point(alpha = 0.4) +
      geom_hline(yintercept = 0, color = "red") +
      labs(title = "Residuals vs Fitted (FE Model)",
           x = "Fitted Values",
           y = "Residuals") +
      theme_minimal()
  )

  # 2. HETEROSKEDASTICITY TEST
  cat("\nBreusch-Pagan Test for heteroskedasticity:\n")
  print(bptest(model_fe))

  # 3. SERIAL CORRELATION TEST
  cat("\nBreusch-Godfrey Test for serial correlation:\n")
  print(pbgtest(model_fe))

  # 4. MULTICOLLINEARITY CHECK
  cat("\nVariance Inflation Factors (VIF):\n")
  print(vif(lm(formula, data = pdata)))

  # 5. ROBUST STANDARD ERRORS
  cat("\nCoefficients with Cluster-Robust Standard Errors:\n")
  print(coeftest(model_fe, vcov = vcovHC(model_fe, type = "HC3", cluster = c("group",
"time"))))

  # ====================== RETURN RESULTS ===================================
  # Return fixed effects model for further analysis
  return(model_fe)
}
# RQ1. What is the impact of core inputs (COWS, LAND, LABOR, FEED) on milk production
over time?
formula_translog <- YIT ~ X1 + X2 + X3 + X4
run_panel_models(formula_translog, pdata)

# RQ2. Are there diminishing or increasing returns to scale in milk production?
formula_translog2 <- YIT ~ X1 + X2 + X3 + X4 + X11 + X22 + X33 + X44
run_panel_models(formula_translog2, pdata)

# RQ3. How do interactions between inputs affect milk production efficiency?
formula_translog3 <- YIT ~ X1 + X2 + X3 + X4 + X12 + X13 + X14 + X23 + X24 + X34
run_panel_models(formula_translog3, pdata)
```