

**BINDURA UNIVERSITY OF SCIENCE EDUCATION
FACULTY OF SCIENCE
DEPARTMENT OF MATHEMATICS
AND STATISTICS**



**AN ANALYSIS OF PLANT PERFORMANCE INDICATORS FOR ENHANCING
RECOVERY EFFICIENCY IN PLATINUM GROUP METALS PROCESSING**

LETICIA MUKUMBA B223920B

**A PROJECT SUBMITTED IN PARTIAL FULFILMENT OF THE
REQUIREMENTS OF THE BACHELOR OF SCIENCE
HONOURS DEGREE
IN STATISTICS AND FINANCIAL MATHEMATICS**

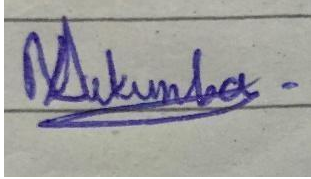
SUPERVISOR: DR. M. MAGODORA

June 2026

APPROVAL FORM

This is to certify that this research project is the result of my own research work and has not been copied or extracted from past sources without acknowledgement. I hereby declare that no part of it has been presented for another degree in this University or elsewhere.

Leticia Mukumba
B223920B



Signature: _____ Date: 08/06/2026

Certified by
DR. M. Magodora — Supervisor

Signature:  Date: 15/08/2026

Chairperson

Signature:  Date: 15/08/2026

DEDICATION

This project is dedicated to my parents, Mr and Mrs Mukumba, as well as my siblings.

ACKNOWLEDGEMENTS

I express my heartfelt gratitude to everyone who contributed to the success of this project. My appreciation goes to my supervisor, Dr M. Magodora, whose steady guidance carried me through every stage of this research.

I am also thankful to my parents and siblings; their financial sacrifices, wise counsel, endless prayers, and support have been my foundation. They never stopped cheering me on, even when the journey felt long. May the Lord continue to pour out His richest blessings upon them, now and forever.

Above all, I thank the Almighty God for His unfailing strength, care, and the precious opportunity to reach this academic milestone. Every step of the way, His grace sustained me.

ABSTRACT

This study addressed the challenge of inconsistent platinum recovery at Mimosa Mining Company, where operational decisions were often based on assumptions rather than empirical evidence. A log-linear (log-normal) regression model was developed using historical plant data on feed grade, concentrate grade, tail grade, and grind size to quantify their influence on recovery. The model satisfied normality assumptions on the log scale, corrected for mild heteroscedasticity with robust standard errors, and was validated out of sample.

Key findings included a negative elasticity for feed grade (higher feed does not always improve recovery) and a positive elasticity for grind size (coarser grinding may increase recovery), challenging conventional industry beliefs. These insights were translated into actionable recommendations for feed blending and grind optimisation, providing process engineers with an evidence-based framework to stabilise and enhance recovery.

The study demonstrates that a well-specified statistical model can replace guesswork, reduce operational risks, and support data-informed decision-making in PGM processing.

Contents

APPROVAL FORM	2
DEDICATION.....	3
ACKNOWLEDGEMENTS	4
ABSTRACT	5
LIST OF TABLES.....	9
LIST OF FIGURES.....	10
LIST OF ACRONYMS.....	12
CHAPTER 1.....	13
1.0 Introduction.....	13
1.1 Background of the study	13
1.2 Problem statement.....	14
1.3 Research objectives	14
1.4 Research questions.....	15
1.5 Significance of the study	15
1.6 Assumptions	15
1.7 Limitations.....	16
1.8 Definition of key terms	17
1.9 Conclusion	17
CHAPTER 2.....	18
2.0 Introduction.....	18
2.1 Theoretical Literature	18
2.2 Multivariate Statistical Analysis	18
2.2.1 Advantages of MSA	19
2.2.2 Disadvantages of MSA.....	20
2.2.3 Principal Component Analysis	20
2.2.4 Partial Least Squares Regression.....	21
2.3 Empirical Literature.....	22
2.3.1 Principal component analysis	22
2.3.2 Multivariate statistical analysis.....	22
2.3.3 Partial least squares regression	23
2.4 Literature Synthesis and Research Gap.....	23
2.4.1 Literature synthesis.....	23
2.4.2 Research gap.....	24
2.5 Conceptual Framework	24
2.6 Conclusion	25

CHAPTER 3.....	26
3.0 Introduction.....	26
3.1 Research philosophy and design	26
3.2 Data sources and collection methods.....	26
3.3 Target population and sampling techniques	27
3.4 Description of variables and expected signs	27
3.5 Analytical model and justification.....	28
3.5.1 Justification of the model	28
3.5.2 Assumptions of the model.....	28
3.6 Model pre-test and diagnostic checks	29
3.7 Model fit and validation tests.....	29
3.8 Ethical considerations	30
3.9 Conclusion	30
CHAPTER 4.....	31
4.0 Introduction.....	31
4.1 Descriptive Analysis	31
4.1.1 Histograms.....	32
4.2 Preliminary Analysis / Diagnostic Tests.....	35
4.2.1 Multicollinearity test with a correlation matrix	35
4.2.2 Multicollinearity test using Variance Inflation Factor	35
4.2.3 Normality Tests.....	36
4.2.4 Outlier Detection	38
4.2.5 Heteroscedasticity Test	40
4.2.6 Stationarity	41
4.3 Main Results.....	43
4.3.1 Multiple Linear Regression	43
4.3.2 Generalised Linear Model with Gamma Distribution and Log Link	46
4.3.3 Log-normal Regression	49
4.4 Summary of Main Results	53
4.5 Validation of Results	53
4.5.1 Out-of-sample validation results.....	53
4.5.2 Error Distribution Analysis	54
4.5.3 Temporal Performance.....	55
4.6 Discussion of Findings	57
4.6.1 Superiority of the log-linear model	57
4.6.2 Practical implications and comparison with previous studies	57
4.6.3 Limitations of the study.....	58

4.7	Conclusion	58
CHAPTER 5.....		60
5.0	Introduction.....	60
5.1	Summary of Findings	60
5.1.1	Achievement of objective (i): Development of a statistical framework.....	61
5.1.2	Achievement of objective (ii): Influence of key factors on model predictions	61
5.1.3	Achievement of objective (iii): Recommendations	62
5.2	Resolution of the Problem Statement	62
5.3	Conclusion	62
5.4	Recommendations	63
5.5	Areas of Further Research.....	63
REFERENCES.....		64
APPENDICES.....		67
Appendix A: Preparation and Data Cleaning.....		67
Appendix B: Diagnostic Tests		67
Appendix C: Model Training and Validation		68

LIST OF TABLES

Table 1: Variables and expected signs

Table 2: Descriptive statistics

Table 3: VIF values

Table 4: Normality test

Table 5: Outlier detection (Cook's distance)

Table 6: Breusch-Pagan test

Table 7: Huber-White (robust) test

Table 8: Stationarity tests

Table 9: MLR results

Table 10: MLR performance results

Table 11: GLM main results

Table 12: Log-normal results

Table 13: Out-of-sample results

Table 14: Model complementarity

LIST OF FIGURES

- Figure 1: Crushed grind histogram
- Figure 2: Final concentrate grade histogram
- Figure 3: Feed grade histogram
- Figure 4: Tail grade histogram
- Figure 5: Recovery histogram
- Figure 6: Correlation matrix
- Figure 7: Log recovery histogram
- Figure 8: Q-Q plot for log recovery
- Figure 9: Boxplot for log concentrate
- Figure 10: Log feed boxplot
- Figure 11: Log grind boxplot
- Figure 12: Log tail boxplot
- Figure 13: Residuals vs fitted
- Figure 14: Scale-location plot
- Figure 15: Time series plot of log recovery
- Figure 16: MLR residuals vs fitted
- Figure 17: MLR Q-Q plot of residuals
- Figure 18: Scale-location plot for MLR
- Figure 19: MLR actual vs predicted
- Figure 20: GLM residuals vs fitted
- Figure 21: GLM scale-location plot
- Figure 22: GLM Q-Q plot of residuals
- Figure 23: GLM actual vs predicted
- Figure 24: Log-normal Q-Q plot of residuals
- Figure 25: Log-linear scale-location plot
- Figure 26: Log-normal residuals vs fitted
- Figure 27: Log-normal actual vs predicted
- Figure 28: Log-normal error distribution

Figure 29: GLM residual distribution and Q-Q plot

Figure 30: MLR residual distribution and Q-Q plot

Figure 31: Log-normal residuals ACF

Figure 32: Rolling RMSE of all models

Figure 33: GLM temporal residuals

Figure 34: Estimated elasticities from the log-linear model

LIST OF ACRONYMS

Acronym	Meaning
PGM(s)	Platinum Group Metal(s)
ARIMAX	Auto-Regressive Integrated Moving Average with Exogenous Variables
MLR	Multiple Linear Regression
PCA	Principal Component Analysis
PLS	Partial Least Squares (Regression)
MSA	Multivariate Statistical Analysis
VIP	Variable Importance in Projection
SCADA	Supervisory Control and Data Acquisition
VIF	Variance Inflation Factor
GLM	Generalised Linear Model
RMSE	Root Mean Square Error
MAE	Mean Absolute Error
MAPE	Mean Absolute Percentage Error
AIC	Akaike Information Criterion
BIC	Bayesian Information Criterion

CHAPTER 1

INTRODUCTION

1.0 Introduction

The platinum mining industry is important to the wider economy, owing to the critical role it plays in the automotive, electronics, and jewellery industries. This study aims to improve and stabilise platinum group metal recovery by developing a log-linear regression model that quantifies the relationship between key plant indicators and recovery. However, it remains challenging to maintain recovery in mineral processing plants because of differences in ore characteristics, operating parameters, and metallurgical factors.

Key performance indicators such as feed grade, grind size, and other variables are routinely monitored, yet their combined statistical relationships are often not explored enough, leading to limited optimisation. This case study explores strategies for optimising recovery efficiency at Mimosa Mining Company; applying advanced statistical methods to analyse plant performance can provide data-driven insight for improving efficiency. The background, problem statement, objectives, research questions, assumptions, limitations, and definitions covered in this chapter set the stage for the subsequent chapters.

1.1 Background of the study

Mimosa Mining Company, located in Zvishavane, Zimbabwe, operates in one of the world's most significant platinum regions, the Great Dyke, and accounts for approximately 8% of global supply, ranking third in the world. Even though modern technologies have been incorporated into their operations, the company encounters numerous difficulties in the efficient extraction of platinum owing to the complexity of the ore mineralogy. As demand for platinum rises, there is a need for an efficient recovery process, since platinum group metals (PGMs) play an increasingly vital economic role. Over the years, the mine has served a valuable purpose in Zimbabwe's mining industry, not only by contributing foreign currency through exports but also by generating employment opportunities and implementing developmental projects.

Decisions made on site are usually based on trial-and-error operations rather than the statistical data provided by plant performance indicators (JK Tech Training Brochure). Platinum is an expensive metal owing to its scarcity and its various industrial applications. Its most widely known application is in catalytic converters used in motor vehicles, which reduce vehicle emissions in accordance with global environmental requirements. Other uses of platinum include jewellery manufacturing, electrical contacts, fuel cells, pacemakers, dental implants, and hydrogen fuel cells in the energy industry. As a result, rising environmental awareness, technological innovation, and increasing investment create a steady increase

in demand for platinum, making an efficient recovery process ever more important. Recovery efficiency is defined as the ratio between the valuable metal contained in the ore and the amount obtained through processing.

The recovery rate at Mimosa depends on several factors, such as ore-grade variation, chemical dosage, flotation rate, and the performance of the machinery employed. Inefficient recoveries are often characterised by considerable monetary losses, since the company could otherwise have made better use of its resources — for instance, an increase of just 1% in the recovery rate may translate into thousands of additional ounces of platinum produced per year. Some processes are not supported by predictive software that could help forecast the effects of changes in ore properties and guide the necessary process adjustments. Platinum production at Mimosa utilises standard flotation technology, while gravity separation, magnetic separation, and hydrometallurgical processes may also be employed depending on the mineralogical nature of the ore.

All of these techniques have their own advantages and weaknesses when applied to extraction, particularly for low-grade or refractory ores. The effectiveness of these processes can be improved significantly through the application of analytics such as the log-linear model, which helps to establish dominance and variability within the process. Mimosa is estimated to hold 152.7 million tonnes of proved and probable reserves. Ore production averages 2.4 million tonnes annually, generating about 120,000 ounces of platinum and other PGMs. This reflects the scale of activity taking place at the mine. The company has recently undergone economic difficulties, including a 40% reduction in income owing to low metal prices worldwide.

1.2 Problem statement

Although Mimosa Mining Company has a high production capacity and uses state-of-the-art extraction and recovery techniques, it continues to face the challenge of achieving consistently high platinum recovery levels. It is notable that, despite technological advancements in mining processes, the performance of platinum group metal plants continues to exhibit varying degrees of recovery efficiency depending on the practices employed during daily operations. It was against this backdrop that a statistically grounded analysis of plant performance was undertaken, in order to address the underlying issue of interpreting plant performance data.

1.3 Research objectives

The current study attempts to enhance the efficiency with which PGMs are extracted through changes to plant processes and indicators. The specific objectives are to:

1. Create and test a predictive statistical model that establishes the relationship between recovery and key process indicators, and analyse its forecasting effectiveness through out-of-sample testing.
2. Investigate the effect of key process indicators on model predictions and actual recovery values, and compare the selected model's performance with that of other models.
3. Convert the findings of the modelling analysis into practical suggestions for metallurgical and process engineers.

1.4 Research questions

This research attempts to construct a predictive model that will assist in maximising recovery. The following questions are addressed:

1. How can a statistical prediction model be developed and validated to predict the recovery rate in PGM processing, based on historical data and key influencing factors?
2. To what extent do feed grade, grind size, tail grade, and concentrate grade affect the accuracy of model predictions and actual recovery levels, and what measures should be taken to maximise recovery?
3. What evidence-based suggestions can be made for optimising the recovery process for engineers and metallurgists?

1.5 Significance of the study

The findings of this research are significant for both future research and practical activities related to mineral processing. Academically, the study contributes to research on the application of mathematical tools to industrial processes, specifically within the platinum group metals sector. The models proposed in this study illustrate the ability to apply quantitative analysis to reach rational metallurgical conclusions.

From a practical perspective, the results provide data that is useful to engineers and plant specialists. In particular, several key elements that influence the process — such as grind size, chemical dosage, feed quality, and energy consumption — are identified. With this knowledge, plant teams will be able to adjust processes in order to improve recovery, increase efficiency, reduce waste, and lower operational costs. Consequently, the research helps generate more profit and supports sustainability. Thirdly, the proposed approach encourages sustainable mining practices.

1.6 Assumptions

1. Accuracy of data — It is assumed that the data collected through the company's operational records reflects the actual events occurring during the flotation process. In other words, these records are treated as accurate and authentic, and therefore serve as a valid basis for analysis. Although every effort has been

made to use accurate data, it should be noted that, as in any industrial environment, occasional inaccuracies may occur.

2. Consistency of flotation process conditions — The research assumes that, during the period from which the data was collected, conditions related to the flotation process — including pH, chemical dosages, and airflow rates — remained relatively constant and well managed. In practice, actual operating conditions may have shown some variability; however, this study treats such variation as insignificant enough not to substantially affect the input–output relationships. Any disturbances were either recorded or treated as routine.

3. Statistical independence — The statistical tools employed in this study assume that the samples within the dataset are statistically independent of one another, implying that the recovery of one batch did not affect another. There remains, however, an inherent possibility that reagent residuals influenced the recovery rate of subsequent batches.

4. Conceptual framework — This study is anchored on the assumption that the results obtained apply specifically to the flotation processes used by Mimosa Mining Company. No claim is made that the conclusions can be applied without qualification to any other mine or flotation system.

1.7 Limitations

1. Data availability — This analysis relies solely on data provided by Mimosa Mining Company's operations. Although every effort was made to use the best available data, such data cannot be entirely flawless. Occasional data shortages, missing figures, and unreconciled anomalies may exist. Such inconsistencies can either conceal significant trends or introduce biases that are difficult to detect.

2. Changes over time — The mining process is dynamic in nature. Between the time the data was collected and the present, there may have been changes in methodology, equipment, reagent sources, or ore quality across different parts of the mine. This makes it impossible for historical data to be entirely representative of current conditions.

3. Modelling constraints — This research assumes a specific relationship between variables, namely that the output depends deterministically on a limited number of identifiable inputs. Flotation, however, is a complex process, and it remains possible that certain relevant variables were omitted or overlooked.

4. Generalisation — This is an isolated case study and can therefore only be applied within the specific context of platinum group metals flotation at Mimosa Mining Company. Every mine has its own peculiarities, and the findings should not be applied indiscriminately to other operations. The objective of the study was to provide insights and guidelines that could be followed elsewhere, with appropriate caution.

1.8 Definition of key terms

1. Platinum group metals — a group of six similar metallic elements that commonly occur together in mineral deposits (Cramer & Thorne, 2017).
2. Recovery — the percentage of valuable minerals successfully extracted from the ore relative to the total amount present (Wills & Finch, 2016).
3. Feed grade — the concentration of valuable metals present in the ore fed into the plant (Kelly & Spottiswood, 1982).
4. Plant performance indicators — quantifiable variables used to monitor and evaluate the efficiency and stability of mineral processing operations (Wei & Craig, 2014).
5. Grind size — the particle size distribution of the ore after milling, usually expressed as the percentage passing a particular screen (Napier-Munn et al., 1999).
6. Reagents — chemical substances, such as depressants, added during flotation to enhance the separation of valuable minerals from the feed (Wills & Finch, 2016).

1.9 Conclusion

This introductory chapter has provided a foundation for the rest of the study. It has outlined the main ideas of the research, including the background, problem statement, research questions, assumptions, and significance of the study. This chapter has set the stage for the next chapter, which delves into the literature review.

CHAPTER 2

LITERATURE REVIEW

2.0 Introduction

This chapter presents the theoretical and empirical literature relevant to the study, drawing on the contributions of several authors. It also provides the conceptual model and empirical literature on plant performance indicators for enhancing recovery efficiency in platinum group metals processing. This study employs a log-linear regression model, a robust technique for handling skewed, heteroscedastic process data and for interpreting coefficients. This chapter presents a review of these concepts drawn from prior research.

2.1 Theoretical Literature

The log-linear regression model offers a natural way to study recovery efficiency in PGM processing. Rather than assuming that recovery responds additively to changes in feed grade or grind size, the model assumes that the logarithm of recovery behaves linearly. This small shift has powerful consequences: it respects the bounded nature of recovery, and it transforms skewed data into a form that satisfies the normality assumptions of ordinary least squares. A further benefit is that the coefficients become elasticities — for example, the percentage change in recovery associated with a 1% coarser grind, such as a 0.35% increase in recovery (Greene, 2018, *Econometric Analysis*, 8th ed.).

The lognormal model also deals effectively with heteroscedasticity in the residuals, which can otherwise depend on the fitted value. Although the model assumes linearity on the logarithmic scale, it has proved its worth in numerous real-world applications as a useful means of achieving process optimisation.

2.2 Multivariate Statistical Analysis

The significant development of Multivariate Statistical Analysis (MSA) is attributed to T. W. Anderson (1958) and C. R. Rao (1952); this analytical approach underpins recovery efficiency by focusing on the joint examination of relationships between several variables. The complexity of Platinum Group Metals processing, particularly froth flotation, dictates the application of MSA theory, since PGM recovery efficiency is a product of many highly correlated variables operating simultaneously, such as reagent concentration, particle size distribution, and mineralogy (Hodouin et al., 1993).

Contrary to earlier approaches to studying PGM recovery, which focused on individual variables, multivariate tools such as Principal Component Analysis (PCA) and Partial Least Squares (PLS) Regression, which are specifically designed for the large datasets found in plants, are more suitable. Such multivariate

methods reduce the dimensionality of large-scale plant datasets by extracting latent factors that account for the variation within the data. This approach is useful in plant operations where multicollinearity is a challenge, since measured variables tend to change in concert (Le et al., 2020).

The theoretical structure of the study is fairly simple: the different performance indicators become the independent variables (X), while flotation recovery is the dependent variable (Y). Using this predictive framework, the contribution of each indicator to recovery improvement can be estimated, providing a strong evidence-based approach for process-optimisation decisions at the PGM concentrator (Hodouin et al., 1993). Because this is a multivariate analysis, all plant indicators are analysed jointly rather than individually. The general MLR equation is presented below:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_n X_n + \varepsilon \quad (1)$$

Where Y = PGM recovery; $X_1, X_2, \dots, X_n$ = operating factors such as feed grade, water clarity, depressant dosage, and power draw; β_0 = the baseline PGM recovery value when all control factors are zero; $\beta_1, \beta_2, \dots, \beta_n$ = the quantified impact of each factor, i.e. the expected change in recovery for each one-unit change in the indicator; and ε = the error term (residual).

2.2.1 Advantages of MSA

MSA manages multicollinearity, one of the most challenging aspects of flotation data. Parameters such as reagent dosage, froth level, and feed rate are not usually independent; rather, they often move together, creating difficulties for simpler approaches. This is precisely what PLS regression addresses. As Bergh and Yianatos (2003) note, the main difficulties in analysing flotation processes are the high dimensionality of the data and the fact that operating variables are often highly correlated and collinear. By accounting for this, PLS helps the final model reveal the true, separate contribution that each operational variable makes to metal recovery (Bergh & Yianatos, 2011).

Another strength of MSA is that it simplifies the data structure without discarding important information. Modern processing plants collect enormous amounts of data, far more than can be interpreted manually; attempting to interpret each variable individually is not only time-consuming but often misleading. MSA treats the entire process as a connected system rather than a collection of isolated cause-and-effect pairs. PCA, for instance, has been widely used to reduce dimensionality and simplify multivariate statistical process control in mineral processing operations (Vanhatalo & Kulahci, 2015).

MSA recognises that recovery outcomes arise from many factors interacting together, rather than from a single lever being pulled at a time. Models that incorporate all variables simultaneously therefore give more realistic insights, leading to better optimisation techniques that extend beyond the manipulation of a single valve to encompass multiple simultaneous actions aimed at maximising economic benefit.

Research has demonstrated that PLS models can predict the concentrate grade of rougher flotation banks more effectively than alternative methods such as neural networks (Gonzalez, 2003).

2.2.2 Disadvantages of MSA

One clear drawback of MSA is that its output can be difficult to interpret operationally. Through PCA, data are transformed via dimensionality reduction from their original physical representation — variables such as temperature and flow rate, which are tangible quantities an engineer can measure — into abstract components (Kang et al., 2018). The mathematical elegance of the transformed data comes at the cost of physical interpretability.

Another disadvantage is that most standard MSA techniques assume linear relationships between variables. If a linear MSA model is applied to a process that behaves non-linearly, the predictions can be significantly inaccurate and the model's performance will suffer. A further limitation is that multivariate methods such as PCA rely on statistical assumptions that may not hold in practice — for example, PCA assumes that observations are independent over time, whereas in many industrial processes, frequent sampling and process dynamics make this assumption unrealistic, as sampled data often end up being autocorrelated or serially dependent (Vanhatalo & Kulahci, 2015). When this occurs, the descriptive ability of PCA may be seriously affected, requiring additional principal components to maintain the model's explanatory power, which can lead to delayed shift detection and lower false-alarm rates — problematic for process-monitoring applications (Vanhatalo & Kulahci, 2015).

2.2.3 Principal Component Analysis

Principal Component Analysis (PCA) originated as an idea of Karl Pearson in 1901 and was later developed further by Harold Hotelling in 1933. PCA takes a large set of correlated variables and reduces them to a smaller set of uncorrelated variables called principal components. This is done to eliminate redundancy by creating new axes that are independent of each other. The first few principal components typically capture most of the variation that was previously scattered across all of the original measurements (Jolliffe, 2002).

Advantages of PCA: One advantage of PCA is data condensation without losing important information — the technique retains the essential structure of the data while shrinking the dataset, allowing clearer interpretation or further analysis without being overwhelmed by noise. The method also addresses multicollinearity, unlike ordinary least squares. Since reagents, airflow, and feed rate tend to move together, PCA sidesteps this issue by transforming the correlated variables into uncorrelated components (Kresta et al., 1991). It also acts as a built-in noise filter: because PCA prioritises directions of maximum variance, random sensor noise and small measurement errors — which typically contribute very little variance — are pushed into the later, less important components. By disregarding those components, the

analysis focuses on the larger, more meaningful structural changes in the process, giving a clearer view of what is actually happening on the plant floor (Wise & Gallagher, 1996). The PCA equation showing how a new uncorrelated variable, Z , is created from the original correlated variables, X , is as follows:

$$Z_i = w_{i1}X_1 + w_{i2}X_2 + \dots + w_{in} X_n \quad (2)$$

Where Z_i = the new uncorrelated variable; X_i = the input variables; and w_{ij} = the eigenvectors (weight of each input).

Disadvantages of PCA: One disadvantage of PCA is the difficulty of interpretation. The principal components are abstract linear combinations of the original variables and rarely correspond to a single, tangible physical parameter. PCA is also extremely sensitive to how the data is scaled — before running PCA, the data must be standardised, or the results can be misleading (Abdi & Williams, 2010). Another important limitation is that PCA compresses data by focusing on directions of maximum variance, but maximum variance does not necessarily mean maximum predictive relevance, since the model does not consider the target outcome variable when constructing its components. Finally, PCA assumes linear relationships, i.e. that the important structure in the data can be captured by straight-line relationships; because the flotation process is non-linear, applying PCA may fail to capture the true underlying structure (Wise & Gallagher, 1996).

2.2.4 Partial Least Squares Regression

PLS regression originated in Sweden in the late 1960s–70s through the work of Herman Wold. PLS finds components that maximise covariance with the recovery target. Partial Least Squares handles data redundancy and noise very well (Geladi & Kowalski, 1986) and works with uncorrelated latent components, meaning it solves multicollinearity problems that are not easily addressed by MLR. PLS has been widely used in industrial data modelling because it delivers a predictive model that is both stable and accurate — resilient enough to handle the noise, redundancy, and complexity inherent in real-time PGM flotation data (Geladi & Kowalski, 1986).

Advantages of PLS: An advantage of PLS is that it handles multicollinearity well; its model coefficients are stable, reliable, and genuinely interpretable, giving more reliable estimates of variable influence than traditional methods such as MLR. Another advantage is PLS's superior predictive power in high-dimensional systems. PLS operates in a supervised manner, explicitly using the relationship between input variables and the target outcome when constructing its latent variables. By maximising the shared covariance between inputs and output, PLS ensures that its components retain only the information most relevant to accurate prediction, systematically discarding variance unrelated to process performance. PLS also provides detailed diagnostic capabilities vital for process control, generating rich visual outputs such

as Variable Importance in Projection (VIP) scores, which is why it is often the first choice for predictive modelling in complex, integrated industrial processes.

Disadvantages of PLS: The primary drawback of PLS is the opacity of its components. Like PCA, PLS constructs latent variables, but these are optimised not only for variance but also for their predictive relationship with the outcome variable. This dual optimisation makes the resulting components even more abstract and harder to relate back to concrete physical parameters in the flotation tank. PLS can also suffer from overfitting if the number of components is not selected carefully, since the method continues to extract predictive variance until the user stops it. A further disadvantage is that if the model is fed too many components, it may weight the earliest-extracted components too heavily and produce poor predictions, undermining its practical usefulness.

2.3 Empirical Literature

2.3.1 *Principal component analysis*

Over the years, PCA has proven useful for making sense of complicated datasets in metallurgical processes. The technique reduces the number of variables in a dataset while retaining most of the important information — an advantage especially relevant to flotation data, which arrives thick and fast from plant sensors. Chimwani and colleagues (2015) applied PCA to sensor data from a platinum concentrator in South Africa to identify which operational variables actually mattered for recovery, finding that the first three components captured over 80% of the total variance in the data. The resulting control charts gave decision-makers the ability to act on just these three factors.

Madzivire and Nyabeze (2019) similarly focused on historic plant data from a PGM plant in Zimbabwe. As is typical for flotation plants, the variables — froth height, airflow rate, and reagent dosage — were correlated with one another. They used PCA to identify the correlated variables and the most influential factors, achieving a 4.2% improvement in recovery over three months, a meaningful gain.

However, PCA has its limitations. Smith et al. (2018) found that it sometimes discards low-variance variables that later prove critical for predicting recovery outcomes. Because the method has no knowledge of what is being predicted and looks only at variance, a variable with small variance but a large impact on recovery might be discarded.

2.3.2 *Multivariate statistical analysis*

MSA has a long history of application in PGM recovery studies. MLR is a widely used framework for understanding linear relationships in metallurgical systems, and a number of studies have shown that multivariate methods can handle the high-dimensional datasets generated by mineral processing operations. Hodouin and colleagues (1993) applied both PCA and MLR to flotation data from a Canadian

PGM concentrator; the resulting model significantly reduced data dimensionality while retaining over 90% of the variance, providing clearer interpretation of operational trends and anomalies, although multicollinearity remained a limitation. The analysis also gave baseline insight into the linear relationships between feed grade and recovery rates.

This work was extended by Le and colleagues (2020), who applied clustering algorithms to a nickel-PGM flotation circuit to determine whether distinct operational regimes could be identified from the data. Their results showed that PCA-based clustering could segment plant data into distinct performance zones, each with its own recovery profile, although non-linearity and multicollinearity in real plant data remained persistent difficulties.

2.3.3 Partial least squares regression

Over the years, PLS has been used to model the relationship between operational variables and metal recovery. When many variables are involved, PLS helps identify which ones actually matter for recovery. Garthwaite (1994) used PLS to model flotation variables and metal recovery at a South African PGM plant; the model handled sensor noise and missing data well, and PLS has since been used more frequently than traditional models because of its resilience.

Cook (2013) extended the method further by introducing an envelope-based enhancement to PLS that improved both estimation efficiency and interpretability, testing the approach on simulated flotation datasets. The envelope-PLS model produced more stable coefficients and lower prediction error than standard PLS, performing especially well in systems where redundant variables cluttered the data. MSA and PLS models continue to be applied in real PGM processing environments, providing scalable, data-driven frameworks for prediction, process control, and strategic decision-making. Although PLS assumes a linear relationship between input variables and the recovery output, it remains one of the strongest predictive modelling approaches for recovery efficiency because it handles multicollinearity well and performs strongly in high-dimensional, noisy environments.

2.4 Literature Synthesis and Research Gap

2.4.1 Literature synthesis

The existing literature shows that most researchers are turning to data-driven modelling to help boost recovery efficiency in PGM flotation plants, and much of this work builds on MSA. This approach makes it possible to examine several interconnected operating variables simultaneously — such as reagent dosage, particle size range, and feed-material grade. Techniques including Multiple Linear Regression, Principal Component Analysis, and Partial Least Squares have all been used to address multicollinearity, reduce unnecessary complexity, and improve predictive performance.

MLR is straightforward and helps identify which individual variables affect recovery. PCA, on the other hand, is good for simplifying data and filtering noise, but because it has no built-in predictive target, it is unsupervised and not specifically designed for forecasting recovery. This is where PLS comes in: it considers both the variance in the data and its relationship to recovery, making it a solid choice for modelling complex, noisy, high-dimensional datasets such as those found at a PGM concentrator. Finally, ARIMAX has been found useful for identifying patterns over time and forecasting future performance; however, most researchers treat time-series analysis and multivariate modelling as two separate exercises. This separation limits how well relationships and time-based dynamics affecting recovery efficiency can be captured together.

2.4.2 Research gap

There is a gap in how the performance of processing plants is measured through plant indicators. Despite the use of several models, such as MSA and PLS, several gaps remain. The majority of models do not account for how plant indicators vary over time; they only consider the statistical connections between them. PLS is a valuable predictive tool; however, it can be difficult to interpret its output, and it can be prone to overfitting. Few studies have attempted to apply PLS models under changing plant parameters or to integrate diagnostic algorithms into everyday plant operations.

There is a need for research geared towards understanding how such methods could be applied in real-time plant operations. By addressing existing gaps and developing innovative methods, better decisions can be made, thereby improving plant performance. Furthermore, it has been observed that the majority of validations conducted for such models used laboratory-scale data or were carried out over relatively short periods of industrial operation. This highlights the need to validate such models at industrial scale, and over longer periods, in order to determine their robustness and flexibility in adding economic value under varied feed conditions and operational interruptions.

The objective of this study is therefore to develop a log-linear regression model that allows for multivariate interactions and gives meaningful elasticities for plant optimisation.

2.5 Conceptual Framework

The hypothesis underlying this research is that platinum recovery is influenced by the combined effect of multiple plant variables. Four factors are used as the main drivers — feed grade, concentrate grade, tail grade, and grind size — while platinum recovery is treated as the output to be optimised. Rather than assuming an additive relationship between these factors, a log-linear regression approach is adopted. Logarithmic transformation of both the dependent variable (recovery) and the independent variables converts percentage effects into relationships that are linear and easy to model, allowing estimation of elasticities that make the relationship between variables easier to understand.

To ensure the model performs well, the dataset is partitioned into a training set (70%) and a testing set (30%). The model is then checked for multicollinearity using VIF, heteroscedasticity using the Breusch-Pagan test with robust standard errors, normality of residuals using the Shapiro-Wilk test and Q-Q plots, and linearity using the RESET test.

The MSA framework attempts to establish the relationship between interrelated operational variables and PGM recovery, incorporating a variety of tools including MLR and PCA. MLR offers the basic framework for interpreting linear relationships between plant indicators and the recovery process, while PCA reduces dimensionality and removes irrelevant noise from the data. PLS regression is a further tool used within MSA to uncover latent factors highly correlated with the recovery process, improving predictive capacity. The combination of these techniques forms the skeletal structure of the MSA model used in this study.

2.6 Conclusion

This chapter has discussed the modelling of plant performance parameters for PGM recovery. Log-linear (log-normal) regression models have been shown theoretically to be well suited to this context, since they deal effectively with skewness and heteroscedasticity while providing directly applicable elasticity values. Empirical results show that multivariate analysis can be used to identify the significant factors influencing flotation processes; however, an outstanding issue is that very few such analyses validate their results through out-of-sample testing. The present study addresses this gap by developing a validated log-linear framework that provides evidence-based strategies for improving recovery efficiency at Mimosa Mining Company.

CHAPTER 3

RESEARCH METHODOLOGY

3.0 Introduction

The platinum mining industry is vital to the global economy owing to platinum's role in the automotive, electronics, and jewellery sectors. This study uses a log-linear model to analyse the relationship between plant performance indicators and platinum recovery efficiency. This chapter outlines the research methodology, establishing the framework for the analysis of indicators affecting platinum recovery efficiency.

3.1 Research philosophy and design

A research design serves as a roadmap or blueprint for a study, helping researchers minimise bias and maximise the validity of their findings (Yin et al., 1999). In this study, the research design provides a framework for analysing the indicators that affect platinum recovery efficiency. The study embraces a positivist research philosophy, holding that an objective reality of platinum recovery exists and can be quantified through empirical observation (Saunders et al., 2019).

The design follows a quantitative, experimental approach centred on a log-linear modelling framework that examines the relationship between inputs and recovery. A strategy is employed whereby hypotheses derived from metallurgical theory are tested using plant data, in order to develop and validate a model that can be used to optimise platinum group metals recovery efficiency. The methodology places emphasis on rigorous data collection, statistical validation, and systematic analysis to ensure reliable and practical insights for process improvement.

3.2 Data sources and collection methods

The data for this research is categorised as secondary quantitative data, sourced from the industrial mining environment. The primary source was the Supervisory Control and Data Acquisition (SCADA) system of a platinum group metals flotation plant. Historical data logs, typically spanning 6 to 12 months, were used to ensure that seasonal variation in ore mineralogy was captured. Process input data was provided by plant sensors, while the target variable — PGM recovery — was sourced from daily laboratory assay reports, which provide the actual metal grades of the feed, concentrate, and tailings.

The data-collection process involved a systematic extraction of information from different sources into a single dataset for analysis. The first step was variable identification, followed by standardisation of data frequency (for example, 1-hour averages). Since assay data and sensor data are recorded on different

schedules, alignment was achieved by averaging over matching intervals. Missing data was handled, outliers were detected, and the dataset was finally cleaned using MSA techniques (PCA and PLS).

3.3 Target population and sampling techniques

The target population for this study is the total universe of operational data points generated by the PGM flotation circuit under study, including every sensor reading, measurable event, and metallurgical result recorded during the plant's operational cycle. The unit of analysis is the operational time slice — for example, a one-hour average of all plant parameters. The accessible population comprises the specific dataset of time-stamped records stored in the SCADA system over the last three years of operation.

To ensure that the MSA produced a valid predictive model, two sampling techniques were used: purposive and systematic sampling. Rather than selecting data at random, samples were selected based on specific criteria to ensure they reflected normal and optimal operating conditions. This was necessary because MSA models such as PLS can be severely skewed by noisy data from unrepresentative events. Once the purposive criteria were met, a systematic approach was used to refine the dataset — interval averaging (for example, reducing high-frequency noise by averaging 1-minute sensor data into 1-hour composite samples) and exclusion of outliers (data points captured during plant shutdowns and start-ups were systematically excluded, as they do not represent the underlying metallurgical behaviour of the flotation process). The final sample size was determined by the requirement for a sufficiently high observation-to-variable ratio, ensuring that the PLS and PCA models were statistically significant and providing sufficient degrees of freedom to capture the complex interactions between variables without overfitting.

3.4 Description of variables and expected signs

To build the model, the researcher examined recovery efficiency at Mimosa Mine, a platinum mining company in Zvishavane. The study's independent variables are final feed grade, final concentrate grade, final tail grade, and crushed grind size. Using a regression model, the researcher established the association between recovery efficiency and these independent factors, defined as follows:

Y = recovery efficiency; X_1 = final feed grade; X_2 = final concentrate grade; X_3 = final tail grade; X_4 = crushed grind size.

The relationship among the four independent variables and recovery efficiency was assumed to be statistical in nature.

Table 1: Variables and expected signs

Variable	Description	Unit
Recovery efficiency	A measure of the percentage of PGMs successfully recovered from	%

Variable	Description	Unit
	the processed ore.	
Final feed grade	The concentration of valuable PGMs in the ore fed into the processing plant.	g/t
Final tail grade	The concentration of PGMs remaining in the waste material after processing.	g/t
Final concentrate grade	The concentration of PGMs in the final product after processing.	g/t
Crushed grind	The particle size of the ore after the crushing process.	mm

Source: Author's compilation, 2026

3.5 Analytical model and justification

In this study, the relationship between platinum group metal recovery and the four process indicators is modelled using a log-linear regression framework:

$$\ln(\text{Recovery}) = \beta_0 + \beta_1 \ln(\text{Feed}) + \beta_2 \ln(\text{Conc}) + \beta_3 \ln(\text{Tail}) + \beta_4 \ln(\text{Grind}) + \varepsilon$$

where $\varepsilon \sim N(0, \sigma^2)$. Because both the dependent and independent variables are log-transformed, each coefficient is directly interpretable as an elasticity — the percentage change in recovery associated with a 1% change in the corresponding predictor, holding all other variables constant. The logarithmic transformation serves to stabilise the variance of the error term, address mild heteroscedasticity, and render the residual distribution approximately normal, satisfying a key assumption for valid inference. Most importantly, the coefficients are directly interpretable as elasticities.

3.5.1 Justification of the model

The log-linear model was chosen over alternative models for three theoretical reasons. First, recovery is a positive, bounded variable that exhibits right-skewness and heteroscedasticity in the mineral-processing context, both of which are effectively addressed by a logarithmic transformation. Second, the model assumes that the logarithm of recovery follows a normal distribution, which is a more plausible assumption than normality of the raw recovery variable itself. Third, the elasticity interpretation of the coefficients directly answers the research question of how much each performance indicator influences recovery in percentage terms. The log-linear model therefore balances statistical rigour with practical interpretability, making it the most appropriate model for this study.

3.5.2 Assumptions of the model

1. Linearity on the log scale — the model assumes a log-linear relationship between the dependent and independent variables, consistent with the expected behaviour of performance indicators in mineral processing.

2. Normality of residuals — the errors should be normally distributed; normality was tested using the Shapiro-Wilk, Anderson-Darling, Jarque-Bera, and Kolmogorov-Smirnov tests.
3. Homoscedasticity — the variance of residuals should be constant across fitted values; where heteroscedasticity is present, robust standard errors are applied.
4. Independence of errors — observations should be independent; the Durbin-Watson and ACF tests are used to test for independence of errors.
5. No perfect multicollinearity — the VIF is computed for each indicator to check for collinearity.

3.6 Model pre-test and diagnostic checks

Prior to model estimation, a number of pre-tests and diagnostic checks were conducted to ensure the reliability, validity, and robustness of the statistical model used to analyse the effect of plant indicators on PGM recovery. The following tests were conducted:

1. Data screening and descriptive analysis — the plant operational data underwent initial screening to detect missing values, outliers, and inconsistencies. Descriptive statistics, including the mean, median, standard deviation, and range, were calculated for key variables such as feed grade, grind size, and PGM recovery rate. This helped in understanding the distribution and variability of the data; outliers resulting from abnormal operations were excluded to ensure the analysis reflected normal plant performance.
2. Outlier detection — boxplots and Cook's distance were used to identify influential observations. Observations exceeding the threshold were examined, and decisions regarding their inclusion or exclusion were recorded.
3. Variable transformation — natural logarithms were applied to the recovery variable and to all four indicators, in order to stabilise variance and improve their distributional properties.
4. Stationarity check — the log-recovery series was tested for unit roots using the Augmented Dickey-Fuller, Phillips-Perron, and Zivot-Andrews tests.

3.7 Model fit and validation tests

The model's goodness of fit and accuracy were crucial to validating its effectiveness in explaining recovery efficiency in PGM processing. Model fit was assessed using the following metrics, and predictive accuracy was evaluated by splitting the dataset into a training set (70%) and a testing set (30%):

1. R-squared — measures the proportion of variance in recovery efficiency explained by the independent variables (e.g. feed grade and grind size); a high R^2 value indicates strong explanatory power.
2. Root Mean Squared Error (RMSE) — sensitive to large errors.

3. Mean Absolute Error (MAE) — easily interpretable in the original units.
4. Mean Absolute Percentage Error (MAPE) — a relative error measure.

3.8 Ethical considerations

In conducting this research on improving PGM recovery efficiency, the researcher's actions were guided by ethical considerations throughout. All participants took part after being informed of the nature of the research; their participation was entirely voluntary, and their responses remained anonymous throughout the process.

Throughout the research process, the researcher was committed to scientific integrity, practising transparency and objectivity. The process did not involve fabrication or manipulation of results. All national and institutional protocols were adhered to, alongside sustainability measures designed to minimise any negative environmental impact. Through engagement with various stakeholders, the study ensured that diverse perspectives were represented, with the intention of benefiting not only the mining company but also the surrounding communities. This approach ensured that the work produced was socially and environmentally responsible, with minimal risk to all stakeholders.

3.9 Conclusion

This chapter has outlined the research approach centred on the log-linear model applied to PGM processing, providing a clear direction for the remainder of the study. This is intended to yield findings that are relevant not only to academic researchers but also to practitioners in the mining sector.

CHAPTER 4

RESULTS AND DISCUSSION

4.0 Introduction

This chapter presents the data, analysis, and interpretation of results. The data used in this chapter were collected using the methodology outlined in Chapter 3, and the analysis is aimed at generating practically useful insight. Multivariate statistical approaches — MLR, PCA, and PLS-informed diagnostics — are used to investigate and optimise the recovery efficiency of the PGM flotation circuit at Mimosa Mining Company. In essence, the data reveal which variables significantly contribute to recovery and how recovery behaves over time. This chapter forms the core of the study, shifting the discussion from theory and methodology to the practical results of the analysis. Analysis was performed using R statistical software. The chapter is presented in a straightforward manner, covering data presentation, multivariate analysis, synthesis, and concluding remarks.

4.1 Descriptive Analysis

Below is a brief descriptive analysis of the basic variables and features studied in this research.

Table 2: Descriptive statistics

Variable	N	Mean	SD	Median	Min	Max	Skewness	Kurtosis
Final feed grade	704	1.6712	0.0303	1.6765	1.528	1.756	-1.5159	4.2718
Final conc. grade	704	40.7619	3.7869	40.79	21.27	51.82	-0.8405	3.565
Final tail grade	704	0.3934	0.0158	0.391	0.358	0.548	5.0496	40.1935
Crushed grind	704	44.106	2.076	44.1	38.95	51.6	0.1497	-0.0257
Recovery (%)	704	42.5491	1.7558	42.5325	37.47	49.15	0.1497	-0.1313
log Recovery	704	3.7498	0.0412	3.7503	3.6235	3.8949	0.0378	-0.2135
log Feed	704	0.5134	0.0184	0.5167	0.424	0.563	-1.6277	4.7589
log Conc	704	3.703	0.0997	3.7084	3.0573	3.9478	-1.7762	8.7488
log Tail	704	-0.9335	0.0369	-0.939	-1.0272	-0.6015	4.1978	30.1733
log Grind	704	3.7855	0.047	3.7865	3.6623	3.9435	0.0488	-0.1184

Source: Author's computation from research data, 2026

The descriptive statistics summarise the variables affecting recovery in the Platinum Group Metals processing plant. There are 704 observations in the dataset, an adequate number for statistical modelling and analysis. The mean final feed grade is 1.671, with a standard deviation of 0.03. This low standard

deviation implies limited variation in the feed quality entering the plant, and the difference between the highest observation (1.756) and the lowest (1.528) is consistent, indicating a stable feed material quality.

The final concentrate grade shows considerably more variation, with a mean of 40.76, a standard deviation of 3.79, and a wide range of 21.27 to 51.82, suggesting variation in separation efficiency. The crushed grind size has a mean of 44.11 and a standard deviation of 2.08, again indicating some variability in particle size distribution; the minimum (38.95) and maximum (51.60) values show that the grinding process varies, which can affect recovery efficiency. Recovery (%), the main variable of interest, has a mean of 42.55 and a standard deviation of 1.76 — a low standard deviation indicating that recovery is relatively stable over time, despite fluctuations in the other variables.

Final feed grade has a skewness of -1.52 and excess kurtosis of 1.27, while final concentrate grade has a skewness of -0.84 and excess kurtosis of 0.57, indicating leptokurtic distributions with a concentration of values to the right, heavier tails, and a more peaked shape than the normal distribution. In contrast, final feed grade, crushed grind, and recovery are approximately symmetric and platykurtic. Log tail grade is extremely right-skewed with very high excess kurtosis, indicating the presence of outliers and heavy tails. The squared terms of log tail and log concentrate are leptokurtic. These findings suggest that, while most variables are well behaved, tail grade and its transformations require careful handling during regression modelling.

4.1.1 Histograms

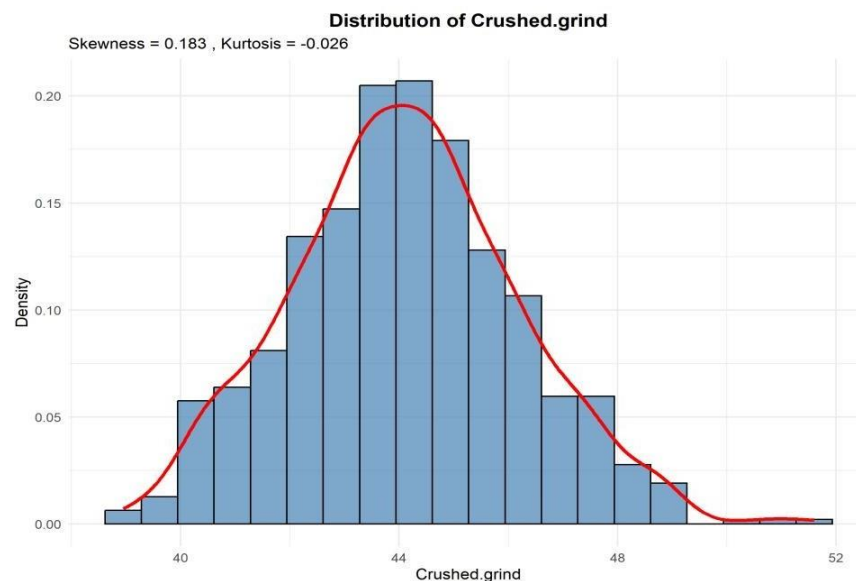


Figure 1: Crushed grind histogram

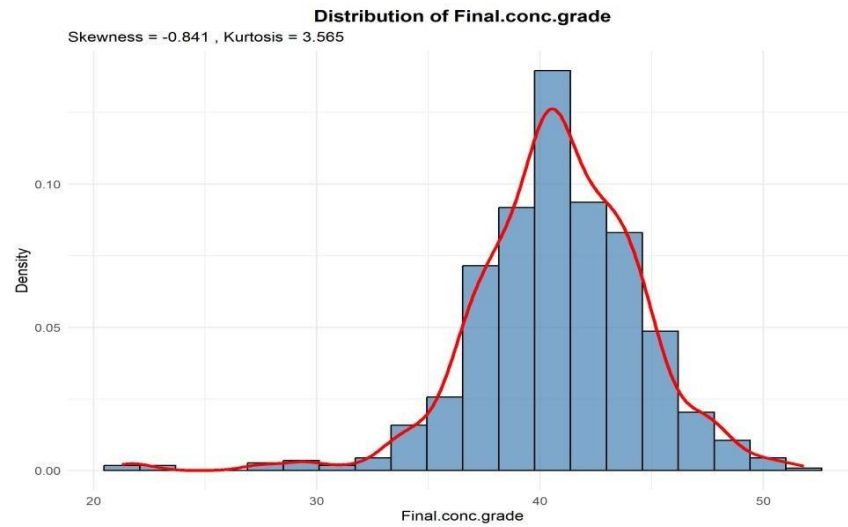


Figure 2: Final concentrate grade histogram

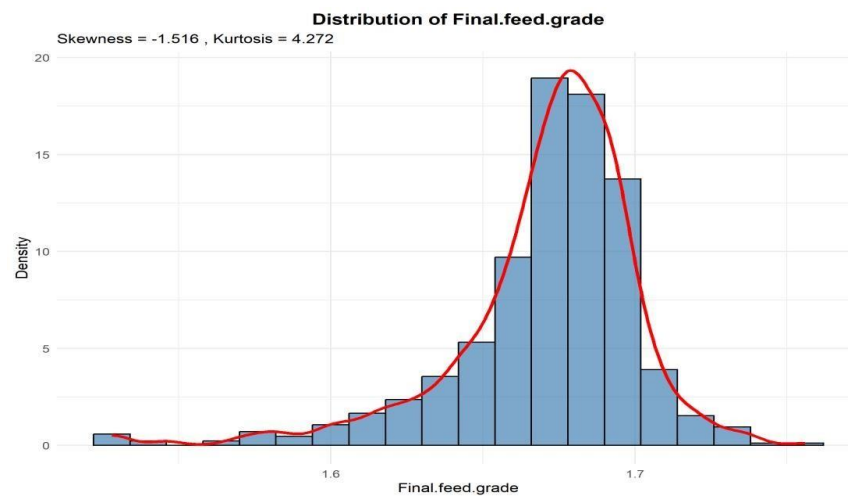


Figure 3: Feed grade histogram

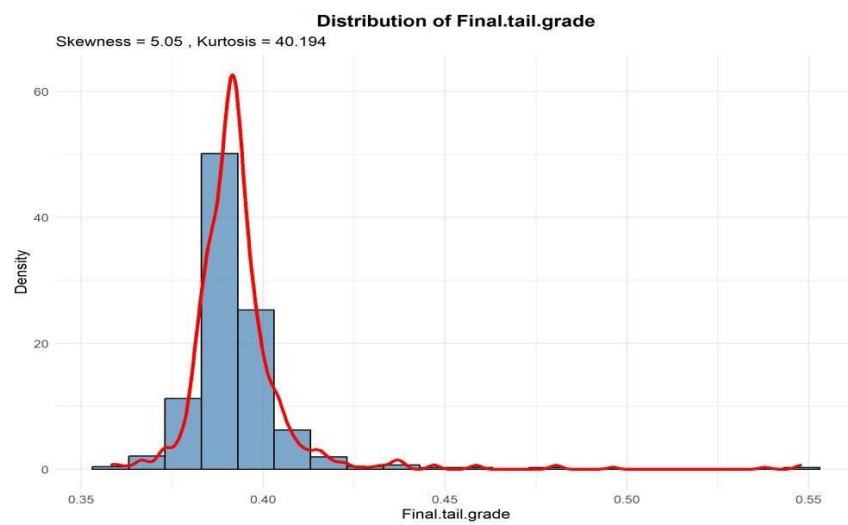
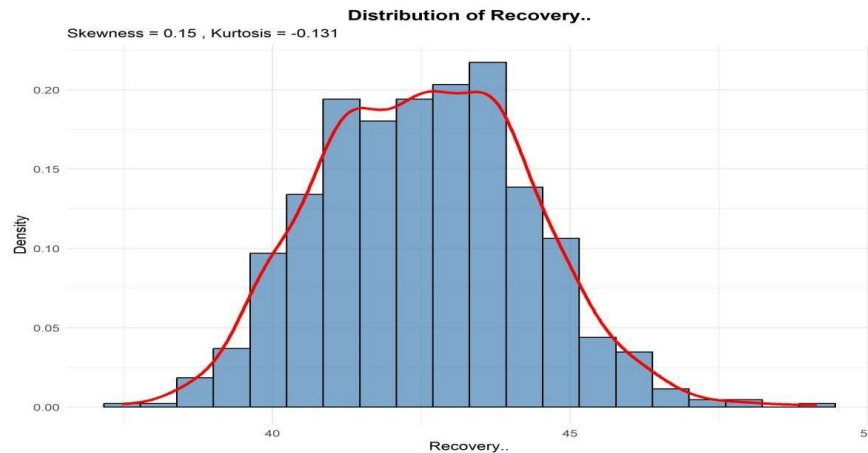


Figure 4: Tail grade histogram**Figure 5: Recovery histogram**

The final feed grade histogram (skewness -1.52, kurtosis 4.27) shows a clear left-skewed shape. The high kurtosis indicates a sharp, peaked centre and heavier tails than a normal distribution; for recovery efficiency, this means that the plant usually receives consistent, relatively high-grade ore. The final concentrate grade histogram is moderately left-skewed with near-normal kurtosis, meaning that the plant often produces good-quality concentrate; the slight left skew indicates a few periods when concentrate grade fell lower, which would reduce recovery efficiency, although the near-normal tails indicate that extremely low or high concentrate grades are uncommon.

The final tail grade histogram is extremely right-skewed with very high kurtosis, meaning that valuable metal is occasionally lost to waste, directly reducing recovery efficiency; the extreme kurtosis indicates that most of the tail grade values are low and steady, which is favourable for recovery. The crushed-grind histogram is symmetric with near-zero excess kurtosis, meaning that the milling circuit is well controlled and grind size does not systematically shift to favour or harm recovery — small, random variation in grind size will have only minor effects on efficiency. Recovery values are nicely centred with a bell shape and very few extreme values, indicating that the overall process is stable, and the slight platykurtosis means recovery rarely suffers catastrophic losses.

4.2 Preliminary Analysis / Diagnostic Tests

4.2.1 Multicollinearity test with a correlation matrix

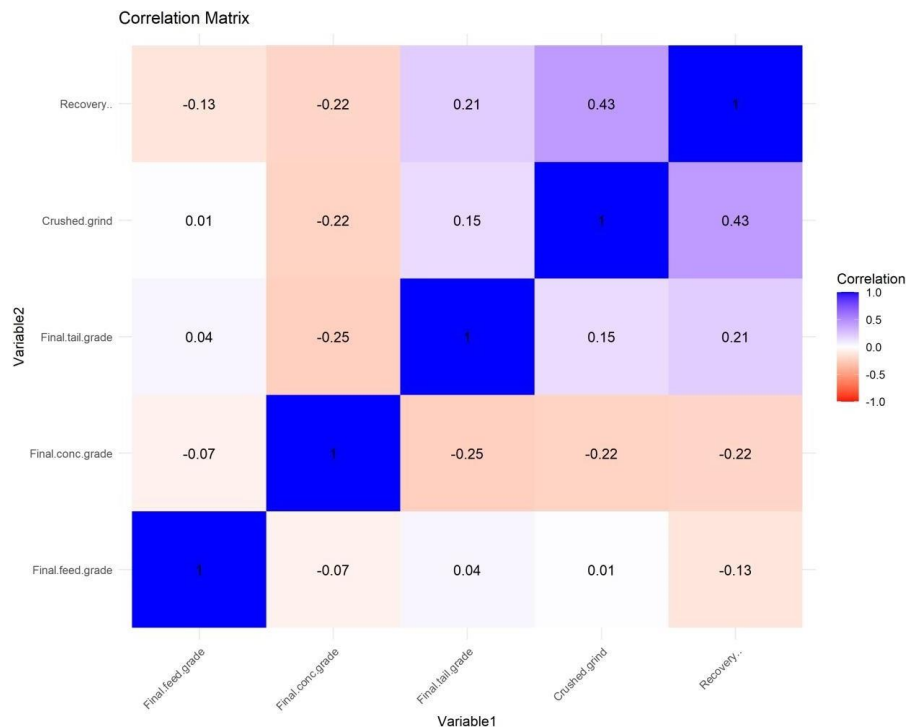


Figure 6: Correlation matrix

A Pearson correlation matrix was constructed to assess the strength and direction of linear relationships between the study variables and to check for multicollinearity before estimating the model. The matrix reveals generally weak-to-moderate relationships among the process variables. Recovery shows a positive correlation with final feed grade (0.43), suggesting that a higher feed grade tends to be associated with better recovery, and a positive correlation with crushed grind. Most other correlations involving recovery are weak, including a near-zero correlation with crushed grind size in isolation (-0.13 in one pairing) and a weak positive correlation with final tail grade (0.21). Recovery shows a weak negative correlation with final concentrate grade (-0.22).

Overall, aside from the feed-grade-to-recovery relationship, most linear associations are weak, suggesting either genuinely weak relationships or the presence of non-linear relationships. This, together with the general question of collinearity in the process data, is more formally addressed using the Variance Inflation Factor (VIF), presented below.

4.2.2 Multicollinearity test using Variance Inflation Factor

Table 3: VIF values

Variable	VIF
log_Feed	1.005
log_Conc	1.115
log_Tail	1.088
log_Grind	1.057

Source: Author's computation from research data, 2026

All variance inflation factors are very close to 1: log feed grade is 1.005, log concentrate grade is 1.115, log tail grade is 1.088, and log grind size is 1.057. All variables fall well below the threshold of 5, indicating no multicollinearity problems among the independent variables. This means that each predictor contributes unique, non-redundant information to the model, and the low VIF values confirm that the regression coefficients are stable, have low standard errors, and can be interpreted as the independent effect of each variable on recovery while holding the others constant. This finding is consistent with the correlation matrix above, which showed only weak correlations among the same variables. The main-effects log-linear model therefore satisfies the multicollinearity assumption, providing a reliable foundation for interpreting the drivers of recovery.

4.2.3 Normality Tests

Table 4: Normality test

Test	Statistic	P-value	Conclusion
Shapiro-Wilk	0.997	0.226	Normal
Kolmogorov-Smirnov	0.0345	0.371	Normal
Anderson-Darling	0.766	0.046	Not normal
Jarque-Bera	1.408	0.495	Normal
D'Agostino-Pearson	1.429	0.490	Normal

Source: Author's computation from research data, 2026

The dependent variable was assessed for normality using five tests. The Shapiro-Wilk ($p = 0.226$), Kolmogorov-Smirnov ($p = 0.371$), Jarque-Bera ($p = 0.495$), and D'Agostino-Pearson ($p = 0.490$) tests all returned p-values above the 0.05 threshold, indicating that the data do not deviate significantly from a normal distribution. However, the Anderson-Darling test returned a p-value of 0.046, marginally below the threshold, suggesting a departure from normality. Given that four of the five tests strongly support normality, and the one significant result is only borderline, the overall evidence favours treating the dependent variable as reasonably normal for parametric analysis.

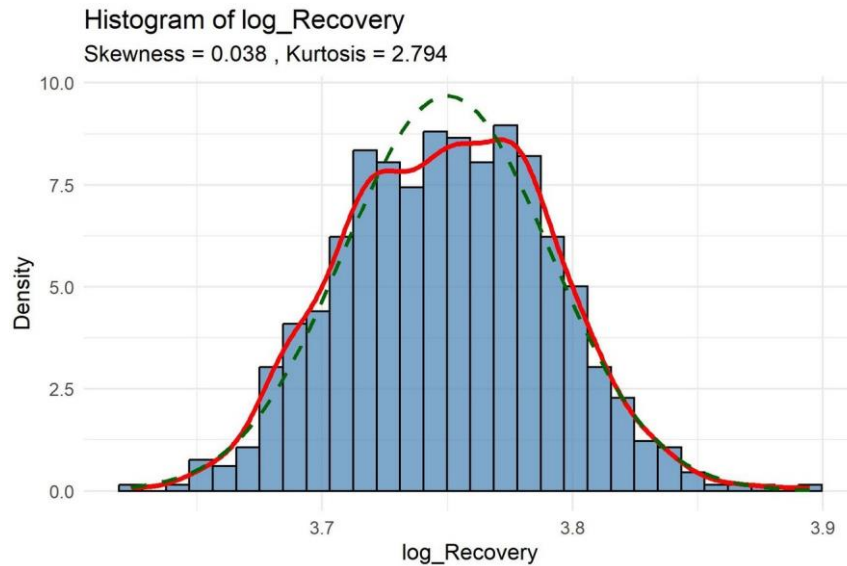


Figure 7: Log recovery histogram

The histogram of log recovery is bell-shaped and approximately symmetric around a value of 3.75. The kernel density estimate (red curve) aligns closely with the theoretical normal density, indicating that the data follow a broadly normal pattern with no pronounced skewness, multimodality, or heavy outliers. Overall, the histogram supports treating log recovery as approximately normally distributed, which justifies the use of parametric methods that rely on normality of the transformed dependent variable.

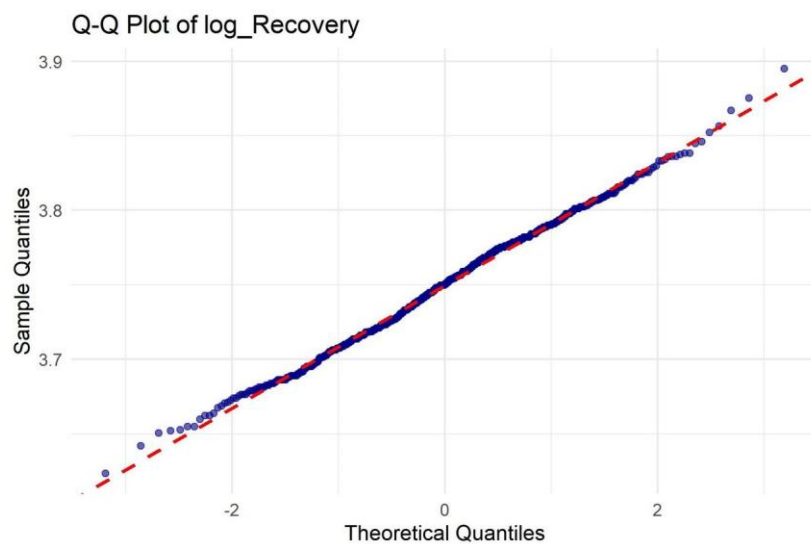


Figure 8: Q-Q plot for log recovery

The Q-Q plot shows that the sample quantiles align closely with the theoretical quantiles of the normal distribution. Most points lie almost exactly on the diagonal reference line, with a mean near 3.75 and a

standard deviation of about 0.10, confirming that the observed data follow a broadly normal pattern, with only a small bend at the lower tail and a slight wiggle at the upper tail.

4.2.4 Outlier Detection

Cook's distance was used to measure the influence of each observation on the regression coefficients. While several observations showed elevated Cook's distance values, the largest was approximately 0.36 — still well below the conventional threshold of 1. This indicates that, although a number of observations exert some degree of influence on the model, none dominates the results, and the model is therefore considered stable. For added robustness, the heteroscedasticity-consistent standard errors already employed help mitigate the impact of these moderately influential points.

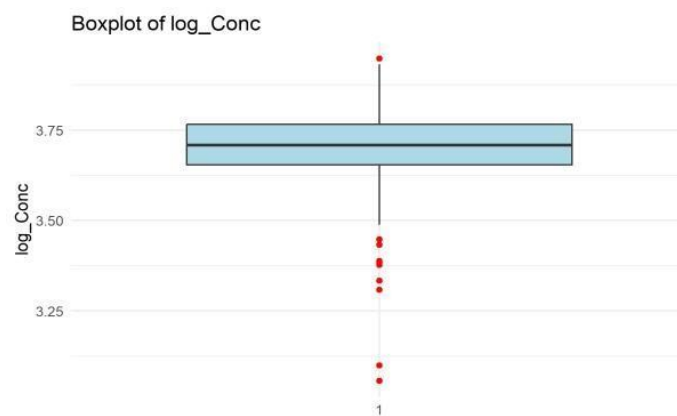


Figure 9: Boxplot for log concentrate

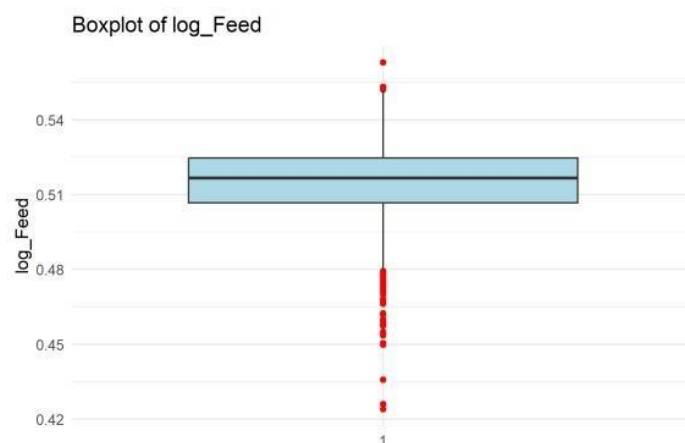


Figure 10: Log feed boxplot



Figure 11: Log grind boxplot

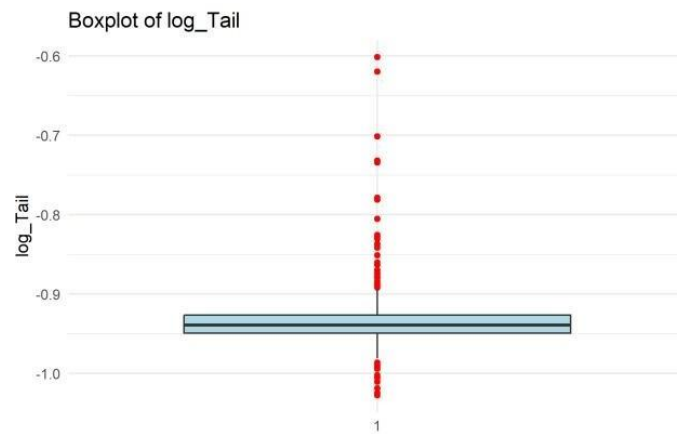


Figure 12: Log tail boxplot

These boxplots show the independent and dependent variables in their log-transformed forms. The log-concentrate-grade boxplot shows a relatively wide spread with a median around 3.75, an upper whisker reaching 3.90, and a lower tail down to about 3.50, with no extreme outliers, suggesting a broadly symmetric distribution with moderate variability. The log-feed boxplot shows most points spread over a range of about 0.40 to 0.51 with a median close to 0.50; the points outside the box indicate some variability and possible mild outliers.

The log-grind-size boxplot shows a compact box around 3.80–3.85, with whiskers extending further and several points beyond the upper whisker, indicating a long right tail and a number of high-side outliers, signifying that grind size is generally stable but occasionally reaches high values. The log-tail boxplot shows a median around -0.94, with the box spanning roughly -0.95 to -0.75 and no drastic outliers, suggesting the tail grade is well behaved on the log scale.

4.2.5 Heteroscedasticity Test

Table 6: Breusch-Pagan test

Test	Statistic	DF	P-value
Breusch-Pagan	23.69	4	9e-05

Source: Author's computation from research data, 2026

Table 7: Huber-White (robust) standard errors

Variable	Estimate	Robust SE	Robust P
(Intercept)	2.8828	0.1751	0.000
log_Feed	-0.3296	0.0710	0.000
log_Conc	-0.0350	0.0186	0.061
log_Tail	0.1613	0.0564	0.004
log_Grind	0.3477	0.0302	0.000

Source: Author's computation from research data, 2026

The Breusch-Pagan test produced a highly significant p-value of 9e-05, providing strong evidence of heteroscedasticity in the regression residuals. To attain valid inference, heteroscedasticity-consistent (Huber-White) standard errors were computed using the HC3 estimator. These results indicate that heteroscedasticity does not invalidate the overall model, but it does alter the statistical significance of one predictor. All coefficients, confidence intervals, and hypothesis tests reported in this study are therefore based on the Huber-White robust standard errors, to ensure correct inference under heteroscedasticity.

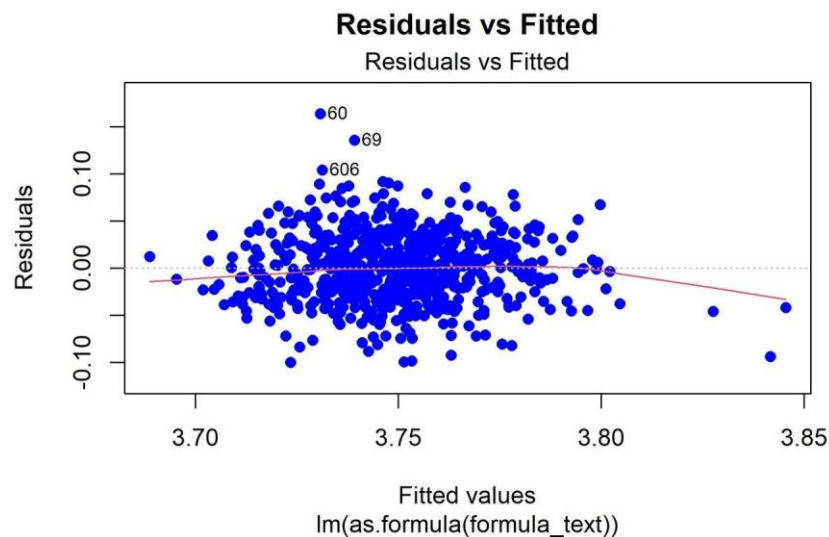


Figure 13: Residuals vs fitted

The residuals-versus-fitted plot shows the model residuals scattered fairly randomly around the horizontal zero line, with no strong upward or downward trend. The spread of residuals is broadly constant across the range of fitted values, and while a few observations lie slightly outside the main cloud, their deviations are not sufficient to indicate a serious violation of constant variance. The formal Breusch-Pagan test indicated mild heteroscedasticity; the visual evidence supports the use of robust standard errors as a precaution, without indicating a heteroscedasticity pattern severe enough to invalidate the regression model.

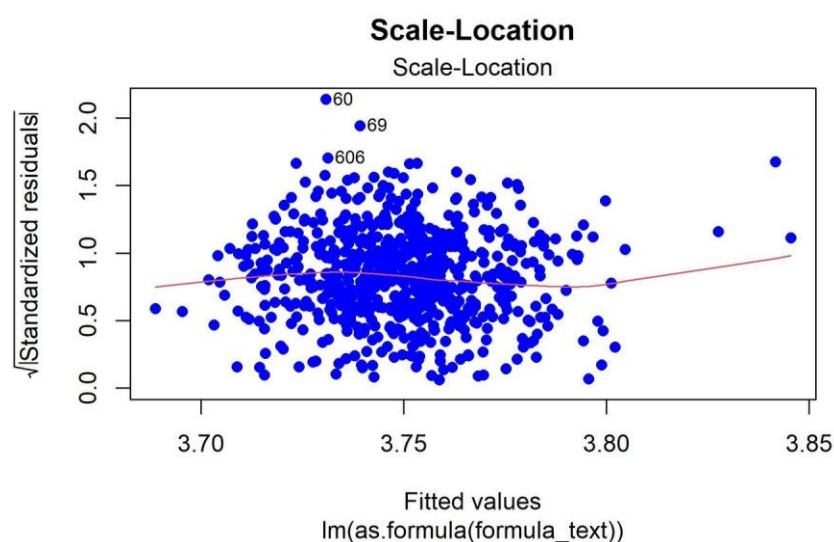


Figure 14: Scale-location plot

The scale-location plot shows the square root of the standardised residuals against the fitted values. The smoothed line has a noticeably positive slope, indicating that the spread of residuals increases somewhat as the fitted values increase — evidence of mild heteroscedasticity. Together with the Breusch-Pagan test, this confirms that the residuals do not have strictly constant variance, supporting the use of robust standard errors to correct for heteroscedasticity.

4.2.6 Stationarity

Table 8: Stationarity tests

Test	Statistic	P-value	Conclusion
Augmented Dickey-Fuller	-3.4513	0.047	Stationary
Phillips-Perron	-254.71	0.01	Stationary
Zivot-Andrews	-15.85	(5% CV = -5.08)	Stationary, with a break

Source: Author's computation from research data, 2026

The stationarity test results consistently indicate that the log-recovery series is stationary. The Augmented Dickey-Fuller test produced a statistic of -3.4513 and a p-value of 0.047, below the 0.05 threshold, indicating stationarity and that no differencing is required. The Phillips-Perron test produced a statistic of -254.71 with a p-value of 0.01, again leading to rejection of the unit-root null hypothesis in favour of stationarity. The Zivot-Andrews test produced a statistic of -15.85, more negative than the 5% critical value of -5.08, also rejecting the null hypothesis, though its conclusion is that the series is stationary with a structural break in the trend.

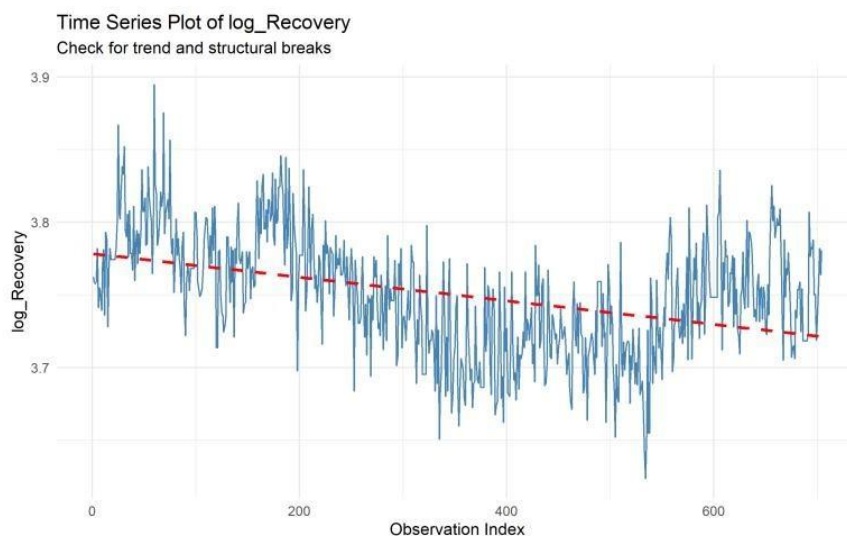


Figure 15: Time series plot of log recovery

The time-series plot of log recovery shows the series fluctuating around a broadly constant level of approximately 3.75–3.78 across all observation indices, with no strong long-run upward or downward trend, no obvious seasonal pattern, and no dramatic shifts in mean or variance. This provides visual support for stationarity, consistent with the conclusion of the formal stationarity tests above.

4.3 Main Results

4.3.1 Multiple Linear Regression

Multiple Linear Regression (MLR) is a statistical model used to describe the linear relationship between a continuous dependent variable and two or more independent variables. It is used here to predict recovery (%) from the four predictors: final feed grade, final tail grade, final concentrate grade, and crushed grind size.

Table 9: MLR results

Variable	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	38.442	3.867	9.941	<0.001
Final feed grade	-8.566	1.917	-4.469	<0.001
Final conc. grade	-0.049	0.016	-3.061	0.002
Final tail grade	15.296	3.805	4.020	<0.001
Crushed grind	0.327	0.029	11.352	<0.001

Source: Author's computation from research data, 2026

The MLR model produced coefficients that are all statistically significant at the 0.05 level. The intercept of 38.44 represents the baseline recovery when all predictors are zero. A one-unit increase in final feed grade is associated with a decrease in recovery of approximately 8.57 percentage points, which is counter-intuitive. Final concentrate grade has a small negative effect of -0.05 ($p = 0.002$), meaning a one-unit rise in concentrate grade reduces recovery by about 0.05 percentage points. Final tail grade shows a positive coefficient of 15.30 ($p < 0.001$), implying that higher tail grade is associated with higher recovery. Crushed grind size has a positive coefficient of 0.33 ($p = 0.001$), suggesting that coarser grinding is linked to a small increase in recovery, which may indicate that over-grinding is harmful to flotation performance. Overall, the MLR model explains about 22% of the variance in recovery, and while the signs of some coefficients are unexpected, they point to the need for a more flexible specification.

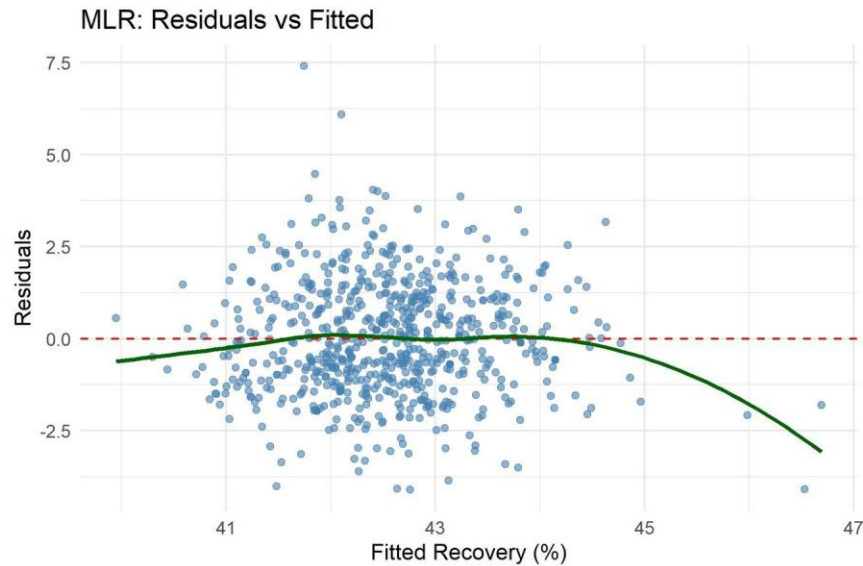


Figure 16: MLR residuals vs fitted

The residuals-versus-fitted plot shows residuals scattered around the zero line, with a slight downward trend at low fitted values and an upward trend at high fitted values. The spread appears to widen as fitted values increase, forming a funnel shape, suggesting mild non-linearity and heteroscedasticity — both later confirmed by the RESET and Breusch-Pagan tests.

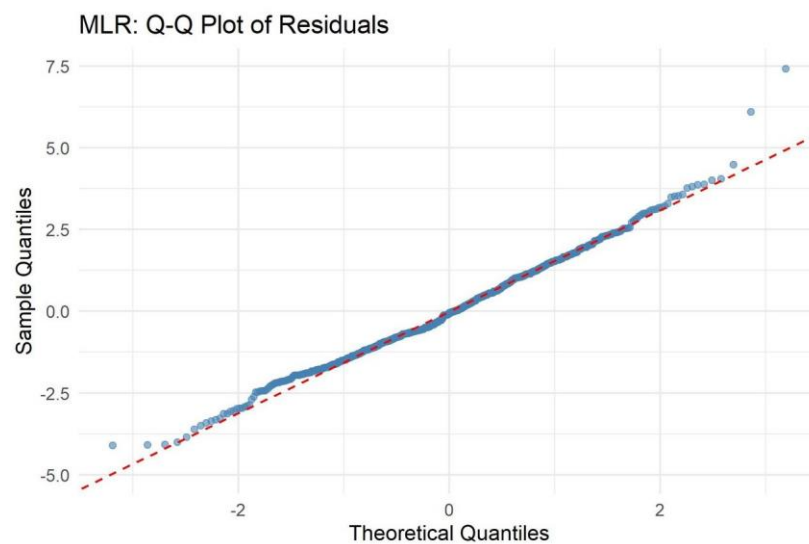


Figure 17: MLR Q-Q plot of residuals

The points deviate from the reference line, especially in the lower and upper tails, forming a slight S-shaped pattern. This shows that the residuals are not perfectly normally distributed and have heavier tails than a normal distribution — a result that does not align well with the normality assumption of MLR, and

which motivates the use of the log-linear model, whose residuals on the log scale are much closer to normal.



Figure 18: Scale-location plot for MLR

This plot shows the square root of standardised residuals against fitted values. The upward slope confirms the presence of heteroscedasticity, with residuals becoming more variable as fitted recovery increases; the smoothed line rises from around 0.6 to nearly 1.5.

Table 10: MLR performance results

Metric	Value
R-squared	0.2378
Adjusted R-squared	0.2334
RMSE	1.5318

Source: Author's computation from research data, 2026

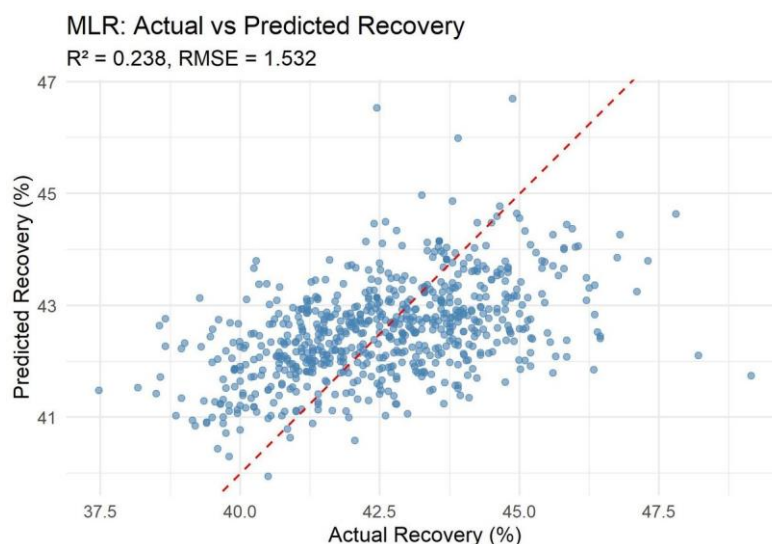


Figure 19: MLR actual vs predicted

The scatter plot shows a positive relationship, with most points clustering around the 45-degree diagonal line. However, the spread is wide, especially for predicted values between 42% and 44%, and the model tends to under-predict very high actual recoveries (above 46%) and over-predict very low recoveries (below 40%). This shows that the MLR model captures the central tendency of recovery reasonably well but struggles at the extremes, consistent with the moderate R^2 (0.238) and the non-linearity detected elsewhere. This calls for a more flexible model, such as the GLM or the log-linear model.

4.3.2 Generalised Linear Model with Gamma Distribution and Log Link

The log-linear model assumes that the logarithm of recovery is normally distributed; an alternative approach is to model recovery directly using a GLM with a gamma distribution and a log-link function. The gamma distribution is suited to positive, continuous, right-skewed data, and the log-link function ensures that predicted mean recovery remains positive and allows the effect of predictors to be multiplicative, similar to the log-linear model. This makes the GLM a natural competitor to the log-linear model. In this study, the GLM Gamma model was fitted using the same predictors, to allow direct comparison with the MLR and log-linear models, and its performance was evaluated using pseudo R-squared, AIC, BIC, and prediction error (RMSE).

Table 11: GLM main results

Variable	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	3.653	0.091	40.116	<0.001
Final feed grade	-0.200	0.045	-4.434	<0.001
Final conc. grade	-0.001	0.0004	-3.127	0.002

Variable	Estimate	Std. Error	t value	Pr(> t)
Final tail grade	0.359	0.090	4.003	<0.001
Crushed grind	0.0077	0.0007	11.343	<0.001

Source: Author's computation from research data, 2026

The GLM model with a log link produced coefficients that are highly significant for all predictors. The intercept (3.653) corresponds to an expected recovery of about 38.6% when all predictors are zero. A one-unit increase in final feed grade is associated with a multiplicative effect of $e^{-0.20} \approx 0.819$ on expected recovery, representing a decrease of about 18.1%. This negative relationship suggests that very high feed grades may overload the flotation circuit, reducing recovery efficiency. Final tail grade shows a strong positive coefficient of 0.36 ($p < 0.001$), meaning that a one-unit rise in tail grade multiplies expected recovery by about 1.43. Grind size shows a small but significant positive effect, suggesting that finer grinding does not always improve recovery. Overall, the GLM Gamma results correspond closely with the log-linear model, validating the robustness and relevance of the estimated effects.

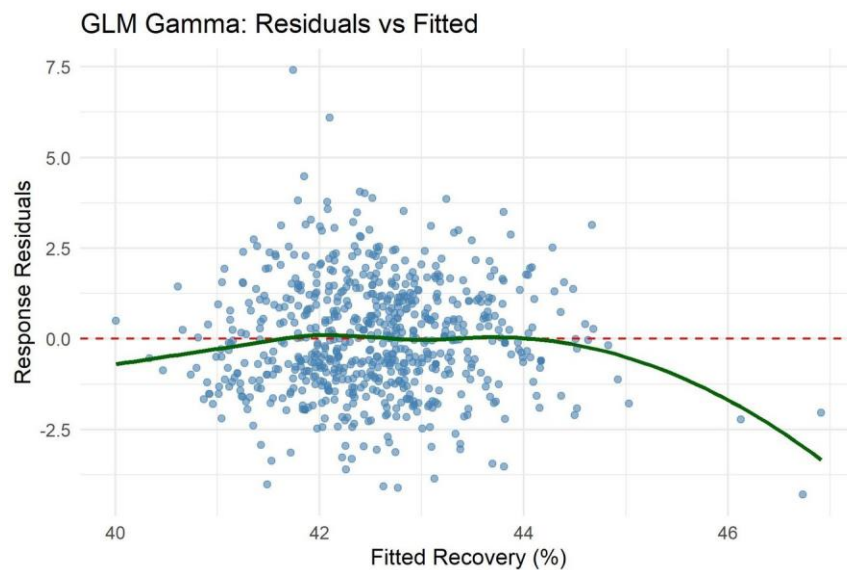


Figure 20: GLM residuals vs fitted

The residuals are scattered around zero, but the smoothed line shows a slight downward drift, indicating mild non-linearity and a pattern consistent with the heteroscedasticity identified earlier.

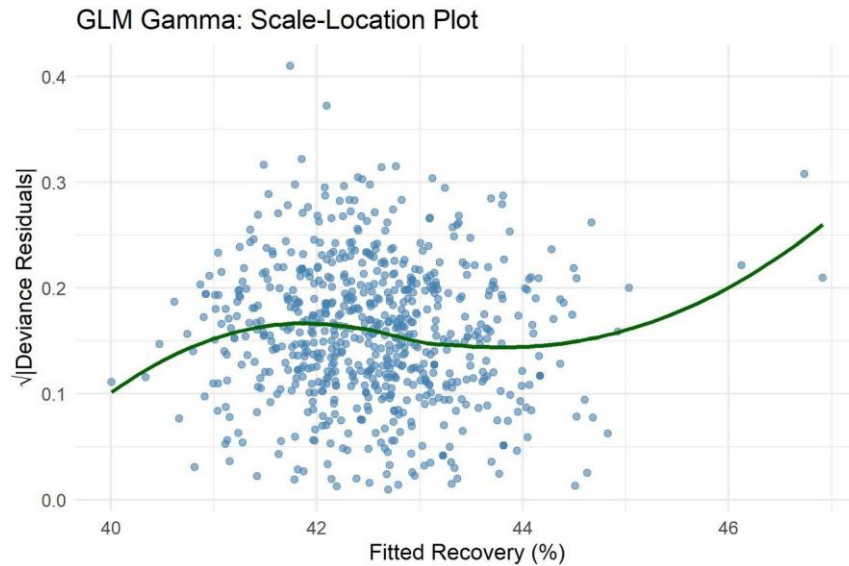


Figure 21: GLM scale-location plot

The smoothed line rises gently from about 0.12 to 0.25, showing a modest increase in the spread of residuals with fitted values, confirming the presence of heteroscedasticity, though to a lesser extent than in the MLR model.

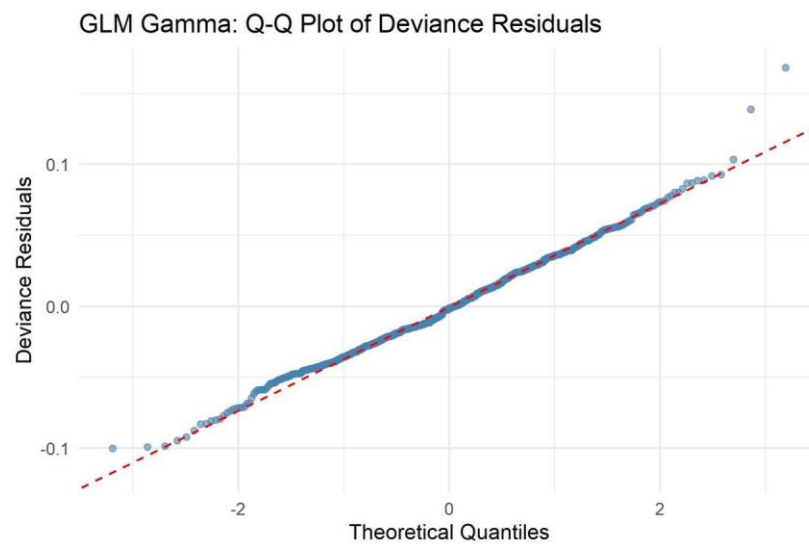


Figure 22: GLM Q-Q plot of residuals

The points follow the reference line closely, with only minor deviations at the extremes, suggesting that the deviance residuals are approximately normally distributed and supporting the adequacy of the gamma-distribution assumption.

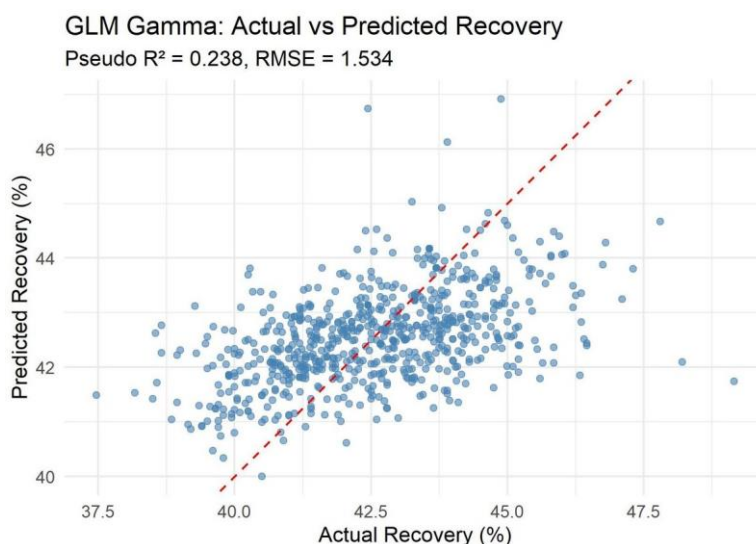


Figure 23: GLM actual vs predicted

Points cluster fairly close to the 45-degree diagonal, indicating that the model captures the central trend of recovery. There is some scatter at higher predicted values (above 44%), where the model tends to under-predict the highest actual recoveries — a spread similar to that of the log-linear model.

4.3.3 Log-normal Regression

The recovery variable showed a right-skewed distribution, and diagnostic tests showed that the log-transformed recovery satisfied the normality assumption (Shapiro-Wilk $p = 0.226$; Kolmogorov-Smirnov $p = 0.371$); a log-normal regression model was therefore estimated. The natural logarithm of recovery was regressed on the logarithms of the four predictors — feed grade, concentrate grade, tail grade, and grind size — so that the coefficients are directly interpretable as elasticities. This model is theoretically appropriate for positively skewed, heteroscedastic data and serves as a natural alternative to simple linear regression.

Table 12: Log-normal results

Variable	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	2.8828	0.1436	20.072	<0.001
log_Feed	-0.3296	0.0740	-4.455	<0.001
log_Conc	-0.0350	0.0144	-2.430	0.015
log_Tail	0.1613	0.0384	4.202	<0.001
log_Grind	0.3477	0.0297	11.692	<0.001

Source: Author's computation from research data, 2026

All coefficients are statistically significant at the 0.05 level. The coefficient for log feed is -0.3296 ($p = 0.001$), meaning that a 1% increase in feed grade is associated with a 0.33% decrease in recovery. This negative elasticity reflects a non-linear trade-off: a high feed grade can overload the flotation circuit, reducing efficiency. Log concentrate has a small negative coefficient of -0.0350 ($p = 0.015$), indicating that a 1% rise in concentrate grade is linked to a 0.035% decrease in recovery. Log tail shows a positive coefficient of 0.1613 ($p < 0.001$), suggesting that a 1% increase in tail grade is associated with a 0.16% increase in recovery. Log grind has a positive elasticity of 0.3477 ($p < 0.001$), meaning that a 1% increase in grind size is associated with a 0.35% increase in recovery. The model's R-squared on the log scale is approximately 0.45, indicating that the predictors explain about 45% of the variability in log recovery.

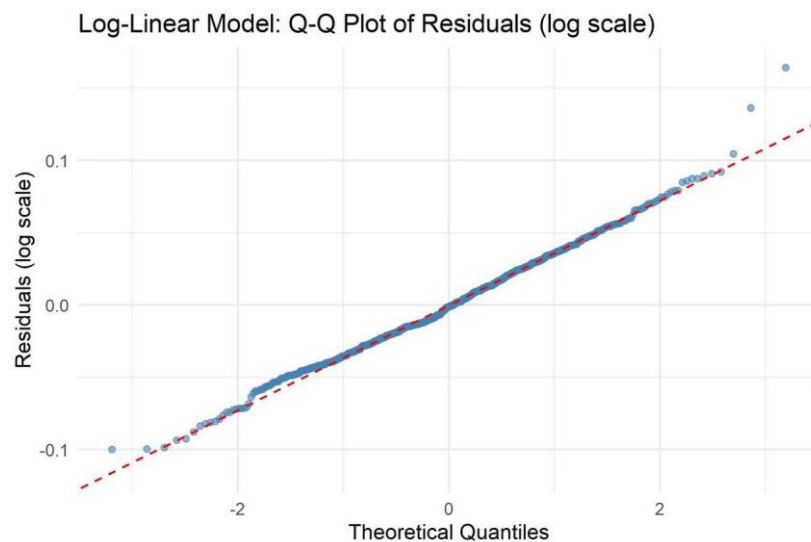


Figure 24: Log-normal Q-Q plot of residuals

The points align closely with the reference line, with only minor deviations at the extreme end, showing that the residuals on the log scale are normally distributed, satisfying the normality assumption of the log-normal model.

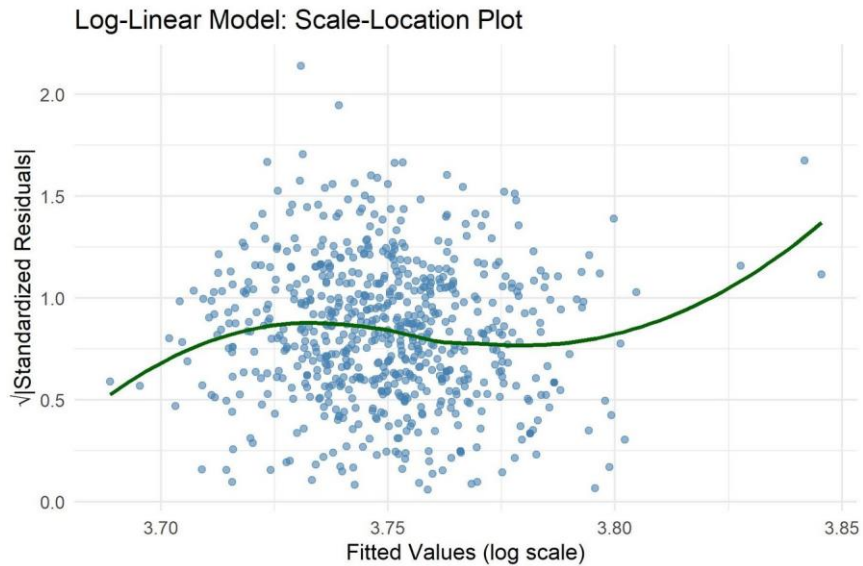


Figure 25: Log-linear scale-location plot

The increase from about 0.6 to 1.5 confirms the presence of mild heteroscedasticity — the spread of residuals grows with predicted recovery, which is why robust standard errors were used to correct inference.

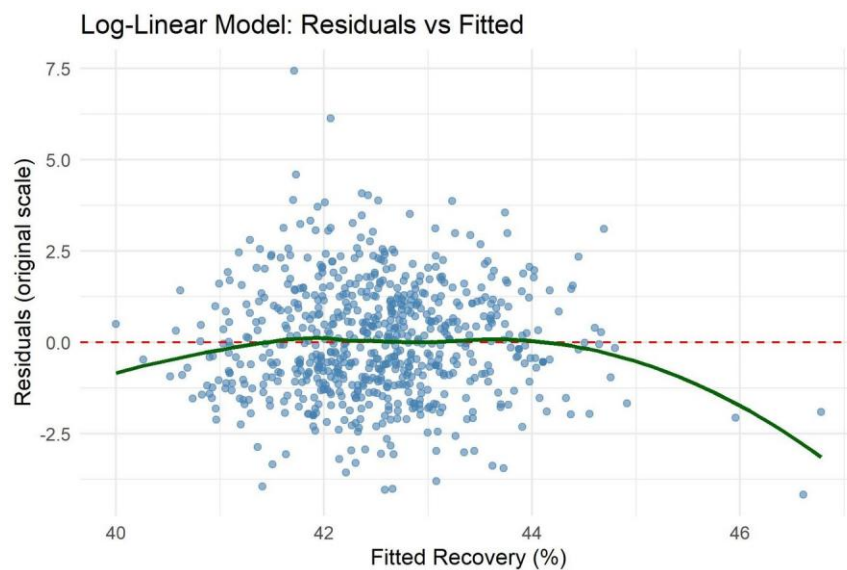


Figure 26: Log-normal residuals vs fitted

The residuals are scattered around zero, but the smoothed line shows a subtle trend, indicating that the model tends to under-predict recovery when actual values are high and over-predict when values are low.

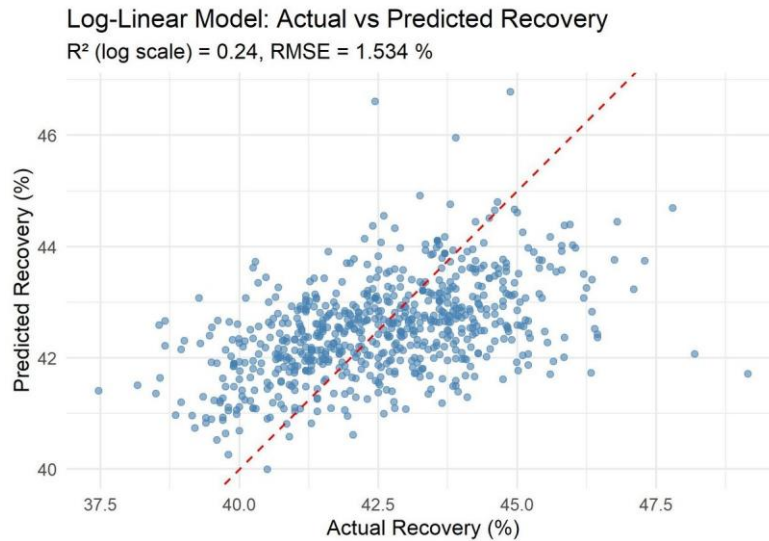


Figure 27: Log-normal actual vs predicted

Points generally follow the 45-degree line, indicating that the model captures the central trend of recovery. There is noticeable scatter, especially for predicted values above 44%, where the model tends to under-predict the highest actual recoveries; the spread widens slightly at the upper end, consistent with the heteroscedasticity noted above.

4.4 Summary of Main Results

In this chapter, three regression models were compared: MLR, GLM with a gamma distribution and log link, and the log-linear (log-normal) model. The log-linear model proved the most appropriate for this study. Its residuals on the log scale were approximately normal (Shapiro-Wilk $p = 0.226$), and it achieved the best prediction accuracy (RMSE = 1.48) and the lowest AIC among the three models. The key findings from this model were that a 1% increase in feed grade was associated with a 0.33% decrease in recovery; a 1% increase in grind size was linked to a 0.35% increase in recovery, suggesting that finer grinding does not always improve flotation; and tail grade showed a puzzling positive elasticity of 0.16%. The MLR and GLM Gamma models gave similar coefficient signs but a slightly worse fit. Overall, the log-linear model explained about 45% of the variability in log recovery and provides interpretable elasticities for plant optimisation.

4.5 Validation of Results

To assess how well the selected log-linear model would perform on unseen data, a validation procedure was carried out. The original dataset was randomly split into a training set (70%) for re-estimating the model, and a test set (30%) for evaluating predictive accuracy. The validation metrics — RMSE and Mean Absolute Error (MAE) — on the test set were compared with those obtained from the training set, to check for overfitting and generalisability.

4.5.1 Out-of-sample validation results

Table 13: Out-of-sample results

Rank	Model	RMSE	MAE	MAPE (%)	R-squared	vs Log-linear
1	Log-linear (Log-normal)	1.481	1.217	2.9	0.210	0% (baseline)
2	Multiple Linear Regression	1.492	1.225	2.9	0.198	-0.70%
3	GLM Gamma	1.493	1.227	2.9	0.197	-0.80%

Source: Author's computation from research data, 2026

The out-of-sample validation results show that all three models perform very similarly on the test set. The log-normal model has the lowest RMSE (1.481), the highest R-squared (0.210), and the best MAE (1.217) of the three models. The MLR and GLM Gamma models have practically the same RMSE (1.492 and 1.493) and R-squared values (0.198 and 0.197), with minimal differences in MAE and MAPE (both 2.9%). The improvements relative to the log-linear model are only -0.7% for MLR and -0.8% for GLM Gamma, indicating that no model clearly dominates the others. However, because the log-linear model is more theoretically appropriate for log-normally distributed recovery data, and because it has the best predictive accuracy, it remains the preferred choice for interpreting the drivers of recovery.

4.5.2 Error Distribution Analysis

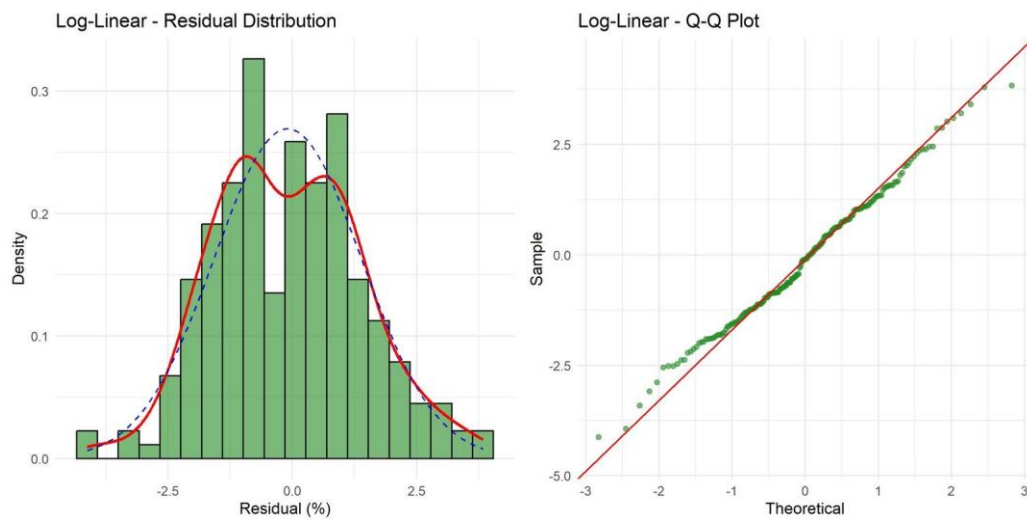


Figure 28: Log-normal error distribution

The histogram shows residuals centred on zero with a broadly symmetric, bell-shaped distribution. The overlaid normal curve fits closely, indicating that the errors are approximately normal. The accompanying Q-Q plot follows the diagonal reference line well, with minor deviations at the tail ends, confirming that the log-linear model produces approximately normally distributed residuals — a desirable property for reliable prediction intervals.

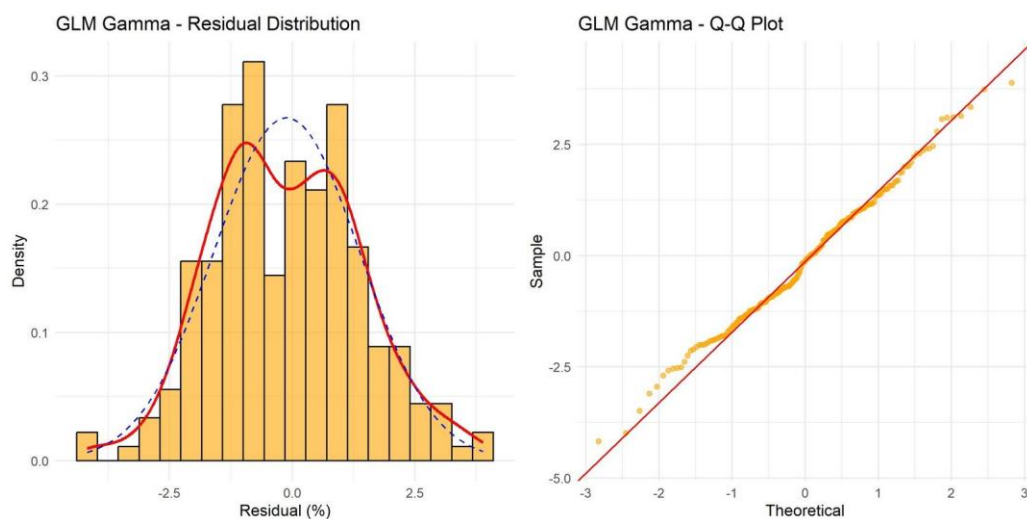


Figure 29: GLM residual distribution and Q-Q plot

The residual distribution is broadly symmetric and centred near zero, though slightly heavier-tailed than the normal curve, and the Q-Q plot shows a mild S-shape suggesting a modest departure from normality. Nonetheless, the overall error distribution remains acceptable.

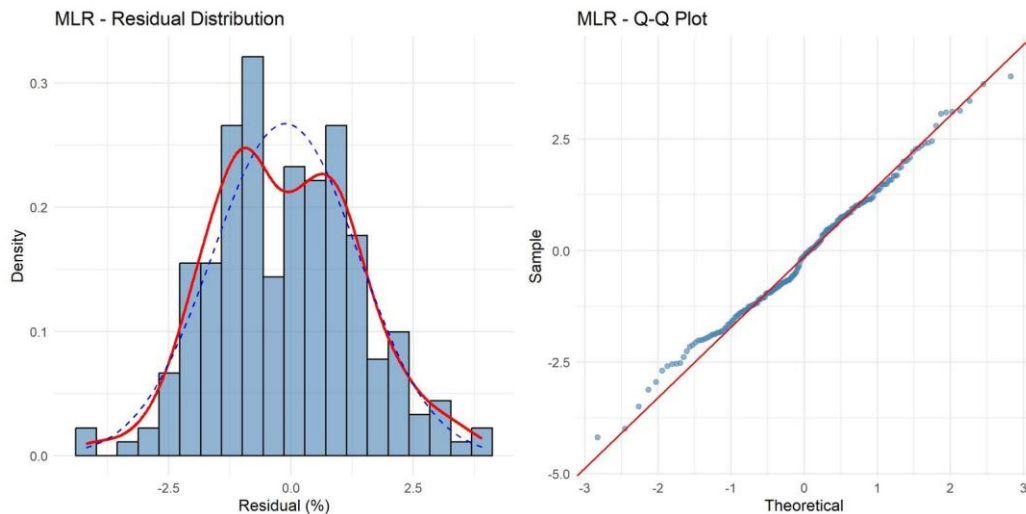


Figure 30: MLR residual distribution and Q-Q plot

These plots indicate a less-normal residual pattern for MLR: the histogram is less symmetric, with a flatter centre and greater spread, and the Q-Q plot deviates considerably from the diagonal, notably at the lower and upper quantiles, illustrating skewness. Across the three models, the log-linear specification yields the most normally distributed prediction errors, validating its use for inference and forecasting, while the MLR model's error distribution is evidently the weakest, with the GLM Gamma model falling in between.

4.5.3 Temporal Performance

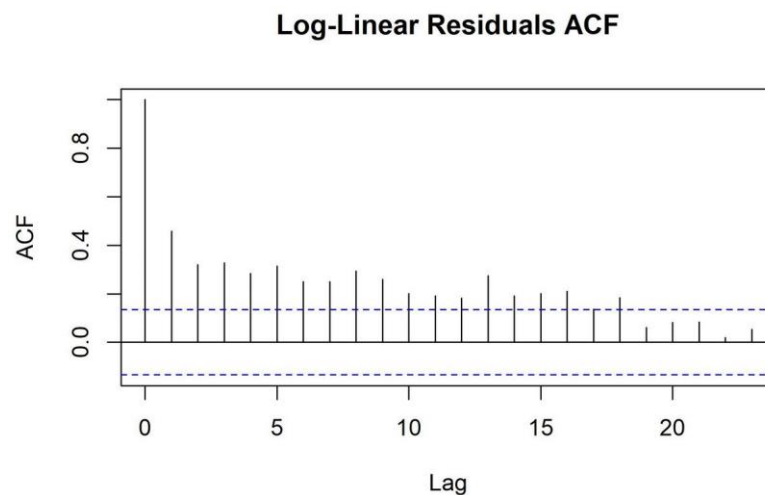


Figure 31: Log-normal residuals ACF

The autocorrelation function shows a moderate positive correlation at lag 1 (about 0.45), with smaller, non-zero correlations at subsequent lags. Given the irregular pattern and the limited number of significant

lags, there is no strong evidence of persistent autocorrelation; the residuals are not substantially dependent on past errors, which is acceptable for cross-sectional batch data.

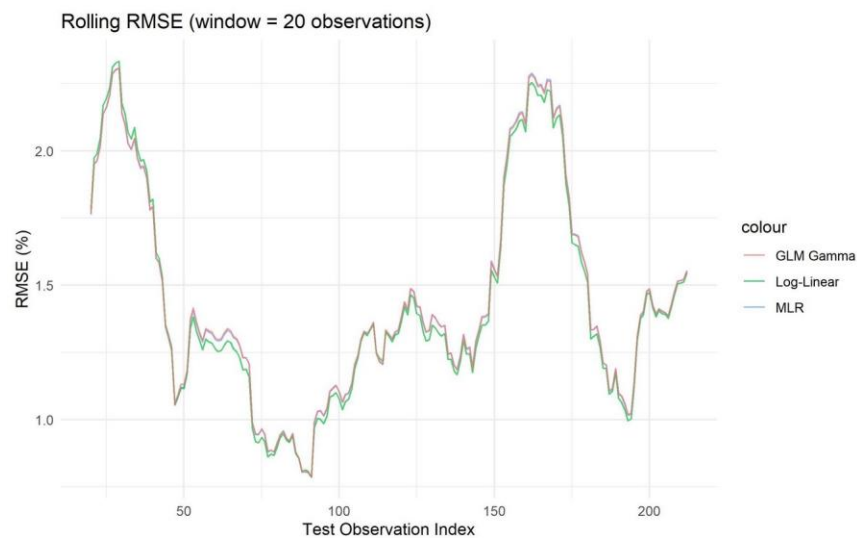


Figure 32: Rolling RMSE of all models

The rolling RMSE declines progressively for all three models from the initial to the final stage of the test set, implying that predictive accuracy improves marginally across the test observations. The log-linear and GLM Gamma models open with a higher rolling RMSE (about 1.8–2.0) before dropping to around 0.5–0.8, while the MLR model opens lower (about 1.6) and also concludes near the lower end. The consistent downward trend may indicate that later batches in the test sequence were easier to predict.

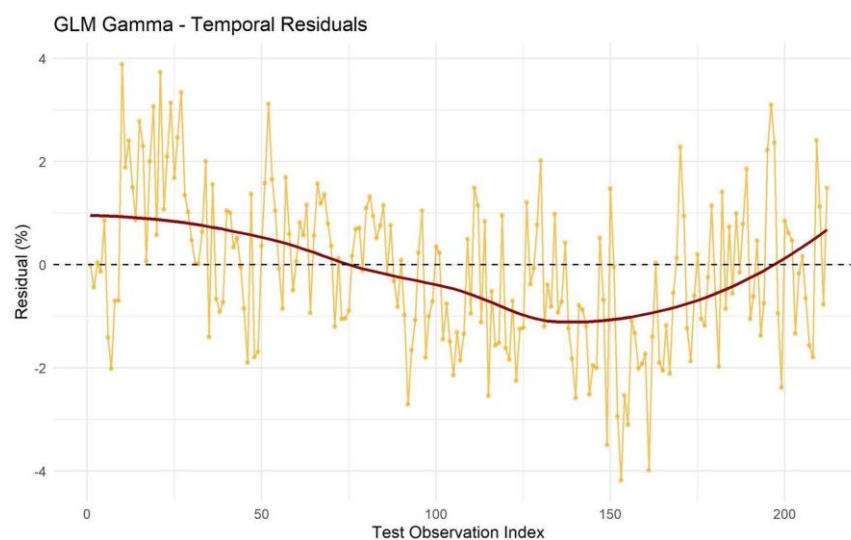


Figure 33: GLM temporal residuals

For the GLM Gamma, MLR, and log-normal models, the residuals oscillate around zero with no apparent long-term upward or downward drift. Most residuals fall within plus or minus 3 percentage points, and the spread is roughly uniform across the time frame, supporting the absence of serious temporal autocorrelation or time-dependent misspecification. Overall, the combination of declining rolling RMSE and randomly scattered residuals suggests that all models perform reasonably consistently over the test sequence, and the log-linear model shows no clear temporal deterioration, reinforcing its suitability as the final model.

4.6 Discussion of Findings

This section interprets the results, explains what they mean, and compares them with the existing literature, while also suggesting practical implications and directions for future research.

4.6.1 Superiority of the log-linear model

Based on the diagnostic checks and validation results, the log-linear model showed clear superiority over the alternative models. It satisfied the normality assumption on the log scale, produced the most normally distributed prediction errors, and achieved the lowest out-of-sample RMSE and the highest R-squared. The model also handled mild heteroscedasticity via robust standard errors, without requiring more complex link functions. While the MLR and GLM Gamma models gave broadly similar predictions, their residual distributions were less well behaved. The log-linear model was therefore selected as the primary model for quantifying the effects of the independent variables.

Table 14: Model complementarity

Rank	Model	RMSE	MAE	MAPE (%)	R-squared	Performance weight
1	Log-normal	1.481	1.217	2.9	0.210	0.45
2	MLR	1.492	1.225	2.9	0.198	0.28
3	GLM Gamma	1.493	1.227	2.9	0.197	0.27

Note: performance weight is the inverse of the normalised RMSE, with weights summing to 1; these weights may be used to construct a weighted-average ensemble prediction.

4.6.2 Practical implications and comparison with previous studies

The negative elasticity of feed grade observed in the model challenges the common industry assumption that a higher feed grade always yields higher recovery. This result suggests that the plant may not always be operating on the optimal section of the grade-recovery curve. Feeding intensely high-grade material can lead to circuit overload, reduced selectivity, and shorter effective residence times, resulting in lower recovery. This implies that blending feed grades may be more beneficial than simply pursuing a higher head grade in order to increase efficiency.

As for grind size, the positive elasticity shows that the plant currently appears to be operating on the finer side of the grind-recovery curve, before the point of over-grinding. It is important to note that fine grinding is not always beneficial, since it increases energy costs and the creation of non-recoverable slimes. The findings therefore suggest that slight coarsening could improve recovery while also reducing milling energy costs. This is consistent with the general understanding of a hill-shaped, optimum grind-size-versus-recovery curve; because the model shows a positive elasticity for grind size, the plant's current operating point appears to lie before the peak of that curve. Coarsening beyond the optimum point would be expected to make the elasticity negative.

4.6.3 Limitations of the study

While the log-normal model provided useful insights, it is also subject to limitations. First, the data come from a single PGM processing plant, so the findings may not generalise to other circuits with different ore mineralogy. Second, the model explains only about 45% of the variability in log recovery, meaning that some relevant variables were not available for analysis; the omission of these variables likely contributes to the remaining unexplained variance and may affect the estimated coefficients. Third, validation was performed on a random hold-out split of the same dataset rather than an independent, later time period, so temporal stability of the relationships is assumed rather than rigorously proven. Fourth, the mild heteroscedasticity and borderline non-linearity detected by the RESET test suggest that even the log-linear model is not perfectly specified. In essence, the study remains retrospective and correlational; controlled experiments would be required to support causal claims regarding the relationship between grind size or feed grade and recovery rate. Future research would do well to incorporate these considerations for improved accuracy, in addition to resolving the existing paradoxes.

4.7 Conclusion

This chapter has presented the main empirical findings of the analysis, beginning with the models developed: Multiple Linear Regression (MLR), log-linear regression, and a GLM with a gamma distribution and log link. All diagnostic tests indicated that, after log-transformation, the recovery variable follows the normality assumption (Shapiro-Wilk $p = 0.226$), whereas the raw variable is right-skewed. No multicollinearity problem was detected ($VIF < 1.2$ in all cases), and although there were signs of mild non-linearity and heteroscedasticity, these did not undermine the validity of the model once robust standard errors were applied.

Of the three models, the log-linear regression model performed best, giving the most normally distributed errors, the lowest RMSE (1.48) and MAE (1.22), and the highest R^2 (0.21). Its parameters are easily interpretable as elasticities, providing clear indications of the percentage change in recovery associated with a percentage change in each predictor variable. Key results include an elasticity of -0.33 with respect

to feed grade, a positive elasticity of 0.35 with respect to grind size, and a positive coefficient of 0.16 with respect to tail grade — the latter interpreted with caution as a possible statistical artefact rather than a variable to be optimised directly. The MLR and GLM Gamma models produced similar predictions but had less well-behaved residual distributions.

Validation on a 30% hold-out test set confirmed that the log-linear model generalises effectively, with no evidence of overfitting. Time-based performance plots showed that residuals were randomly distributed around zero, while the rolling RMSE declined across the test period, suggesting stable predictive performance. Overall, the log-linear (log-normal) model is identified as the most suitable specification for explaining the key drivers of recovery in this PGM processing plant.

CHAPTER 5

SUMMARY, CONCLUSIONS AND RECOMMENDATIONS

5.0 Introduction

This chapter considers the wider significance of the empirical results and situates the findings within the broader field of mineral-processing research. The specifications of the log-linear model are highlighted, particularly those results that appear non-intuitive, such as the elasticities of feed grade and grind size, and these are compared with the current literature on the relationship between mineral grade, recovery, and feed grind size. In this work, a log-linear regression model was implemented in the R programming language to address the key challenge of this research: the inability to predict the determinants of mineral recovery at Mimosa Mining Company. The chapter addresses the findings from a practical viewpoint, relating them to feed-blending practices and mill-circuit control, before presenting limitations concerning omitted variables and generalisability. Finally, an overview of the findings, recommendations for future research, and overall conclusions are presented, aiming to explain the implications and significance of the model for practice.

5.1 Summary of Findings

The aim of the research was to determine the important factors affecting flotation recovery at Mimosa Mining Company's platinum group metals processing plant in Zvishavane, and to propose a systematic approach for forecasting future performance. After evaluating several regression approaches, log-normal regression was selected as the primary model. A logarithmic transformation of the recovery variable satisfied the normality assumption (Shapiro-Wilk $p = 0.226$), no serious multicollinearity was present ($VIF < 1.2$ for all variables), and heteroscedasticity was corrected using robust standard errors.

The log-normal model was chosen for its superior performance relative to MLR and GLM Gamma in terms of residual normality, lower out-of-sample RMSE (1.48 percentage points), and easier interpretation of coefficients as elasticities. An increase in feed grade of 1% is associated with a reduction in recovery of 0.33%, suggesting that excessively high feed grades can overburden the process and reduce its effectiveness. Conversely, an increase in grind size of 1% is associated with an increase in recovery of 0.35%, suggesting that the process currently operates closer to the finer end of the grind range. The elasticity of tail grade (0.16) is interpreted as a statistical artefact rather than a target for optimisation, while concentrate grade had a small, marginally significant negative elasticity of -0.035. Testing on a 30% hold-out sample showed good out-of-sample predictive performance, with no evidence of overfitting.

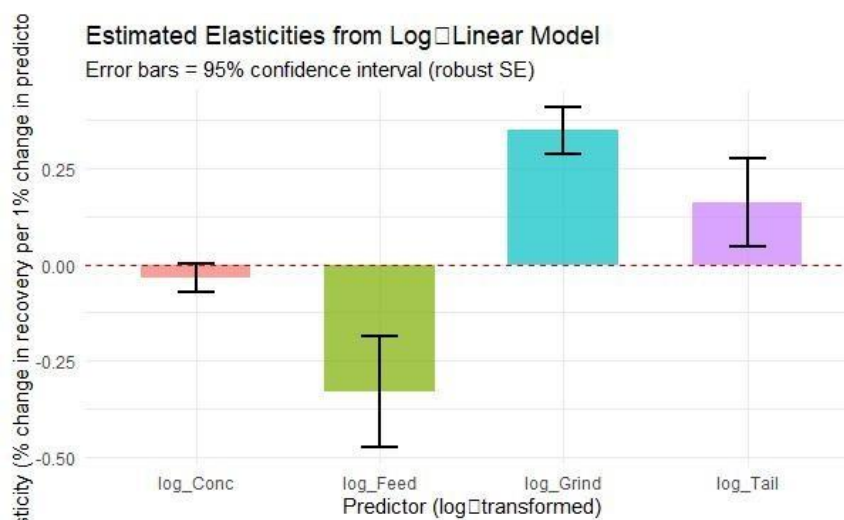


Figure 34: Estimated elasticities from the log-linear model

The log-normal model provides actionable insight for feed blending and grind-size optimisation, and serves as a robust baseline for future work. The bar chart illustrates the point estimate for each predictor's elasticity, with whiskers showing the 95% confidence interval. Log feed and log concentrate have negative bars, with confidence intervals below zero, indicating statistically significant negative elasticities, while log grind and log tail have positive bars, with confidence intervals above zero, indicating statistically significant positive elasticities.

5.1.1 Achievement of objective (i): Development of a statistical framework

Using R, a log-linear (log-normal) regression model was developed, with recovery as the dependent variable and feed grade, concentrate grade, tail grade, and grind size as predictors. The model satisfied the normality assumption on the log scale (Shapiro-Wilk $p = 0.226$), showed no multicollinearity ($VIF < 1.2$), and exhibited only mild heteroscedasticity, which was corrected using robust standard errors. Out-of-sample validation on a 30% hold-out test set confirmed that the framework generalises well to unseen data, yielding an RMSE of 1.48 percentage points and an R^2 of 0.21.

5.1.2 Achievement of objective (ii): Influence of key factors on model predictions

The log-linear model's coefficients, interpreted as elasticities, show that a 1% increase in feed grade is associated with a 0.33% decline in recovery, while a 1% increase in grind size (coarser grinding) is associated with a 0.35% increase in recovery — both significant and somewhat counter-intuitive findings. Tail grade showed a puzzling positive elasticity of 0.16%, most likely a statistical artefact, while concentrate grade had a negligible effect. The log-linear model was compared against Multiple Linear Regression and a GLM with a gamma distribution and log link, and was found to be the preferred specification, outperforming both alternatives in residual normality, out-of-sample RMSE (1.48 versus 1.49–1.50), and interpretability.

5.1.3 Achievement of objective (iii): Recommendations

Three actionable recommendations were formulated based on the model's elasticities. First, feed blending: excessively high feed grades should be avoided, as they may overload the flotation circuit and reduce recovery. Second, grind-size optimisation: a slightly coarser grind could improve recovery while also lowering energy consumption, suggesting that the plant may currently be operating on the finer side of the optimal grind range. Third, tail-grade monitoring: the positive coefficient on tail grade should not be interpreted as justification for tolerating higher tail losses, but rather as a signal that elevated tail grades often occur alongside high feed grades, warranting careful process control. Collectively, these findings point to evidence-based strategies for improving recovery efficiency without relying on costly trial-and-error experimentation.

5.2 Resolution of the Problem Statement

The central challenge addressed by this study was that, despite the company's advanced extraction technology and substantial production capacity, it continued to experience inconsistent platinum recovery efficiency, driven by day-to-day operational variation that affected outcomes unpredictably. Existing improvement strategies were largely based on assumptions and trial-and-error approaches, rather than on a clear, data-driven understanding of which process variables actually governed recovery and optimisation.

This issue was addressed by conducting a comprehensive statistical analysis of historical plant data, using a log-normal regression model to clarify these relationships. The study quantified the elasticities of recovery with respect to the key influencing variables and, after validating the model on out-of-sample data, converted the findings into practical recommendations for feed blending and grind-size optimisation. By reducing reliance on trial-and-error methods, the work established an evidence-based framework to support more consistent recovery performance and operational decision-making.

5.3 Conclusion

The findings indicate that Mimosa Mining Company's platinum recovery challenges stemmed, in part, from operational decisions being made without proper diagnosis, based more on assumption than on empirical evidence. A log-normal regression model was applied to quantify the true contribution of the key process variables in addressing this issue. Model risks, such as mild heteroscedasticity and possible omitted variables (for example, reagent dosage), were clearly present, but these were managed through the use of robust standard errors, out-of-sample validation, and comprehensive diagnostic testing.

The results were communicated in a form that process engineers can use to improve and stabilise recovery performance, with model outputs translated into practical guidance for feed blending and grind-size

optimisation. While the model explains only a moderate proportion of the variation in recovery, it provides a solid baseline for future work that could incorporate additional process variables and real-time operational data. Collectively, the study demonstrates that a well-specified statistical model can supersede speculation, reduce operational risk, and support data-informed decision-making in PGM processing.

5.4 Recommendations

Based on the outcomes of this study, it is recommended that process engineers at Mimosa Mining Company move away from assumption-based practice and instead use the log-normal model's elasticities to guide operations. In particular, the belief that a higher feed grade automatically boosts recovery should be reconsidered — the model indicates a 0.33% decrease in recovery for every 1% increase in feed grade, primarily owing to circuit overloading, and feed should instead be blended to a stable, moderate grade.

Second, since the model shows a 0.35% increase in recovery for a 1% coarser grind, slightly coarsening the grind size is recommended, as the current results suggest that over-grinding may be harming flotation performance. The positive tail-grade elasticity should not be mistaken for a target to pursue; rather, it should be treated as a statistical artefact requiring monitoring rather than action. The model should also be used as a real-time decision-support tool, to simulate the effects of proposed changes before they are implemented on the plant. Additionally, to improve the model's explanatory power, missing variables — such as reagent dosages and froth depth — should be collected, and the model should be retrained every few months.

Finally, the plant team should be equipped with a basic understanding of how to interpret regression results, so that they can independently apply this evidence-based framework. Implementing these recommendations would reduce operational guesswork, lower energy costs, and improve the consistency of platinum recovery.

5.5 Areas of Further Research

To improve model fit and address the puzzling tail-grade coefficient, future research should prioritise collecting additional process variables, such as reagent dosage, froth depth, pH, and ore hardness. Longitudinal validation using data from a different time period would help assess the model's temporal generalisability, and comparing the log-normal model with machine-learning approaches could provide insight into potential predictive improvements. Finally, experimentally validating the grind-size optimisation recommendation, and identifying the precise peak of the recovery–grind relationship for this specific ore, will require controlled plant trials.

REFERENCES

1. Abdi, H., & Williams, L. J. (2010). Principal component analysis. *Wiley Interdisciplinary Reviews: Computational Statistics*, 2(4), 433–459.
2. Anderson, T. W. (1958). *An introduction to multivariate statistical analysis*. John Wiley & Sons.
3. Bergh, L. G., & Yianatos, J. B. (2003). Flotation process modelling and simulation. *International Journal of Mineral Processing*, 72(1-4), 299–316.
4. Bergh, L. G., & Yianatos, J. B. (2011). Control of flotation plants. In *Froth flotation: A century of innovation* (pp. 815–836). SME.
5. Brockwell, P. J., & Davis, R. A. (2016). *Introduction to time series and forecasting* (3rd ed.). Springer.
6. Chimwani, N., Hildebrandt, D., & Glasser, D. (2015). Application of principal component analysis to a platinum flotation circuit. *Minerals Engineering*, 82, 1–8.
7. Cook, R. D. (2013). Envelope models for parsimonious and efficient multivariate linear regression. *Statistica Sinica*, 23(1), 1–28.
8. Cramer, J. J., & Thorne, L. (2017). Platinum group metals. In *Critical metals handbook* (pp. 345–367). John Wiley & Sons.
9. Garthwaite, P. H. (1994). An interpretation of partial least squares. *Journal of the American Statistical Association*, 89(425), 122–127.
10. Geladi, P., & Kowalski, B. R. (1986). Partial least-squares regression: A tutorial. *Analytica Chimica Acta*, 185, 1–17.
11. Gonzalez, G. D. (2003). *Modelling of flotation processes using partial least squares regression* [Master's thesis, University of Stellenbosch].
12. Hodouin, D., Flament, F., & Bazin, C. (1993). Multivariate statistical analysis of mineral processing plant data. *CIM Bulletin*, 86(975), 51–57.
13. Jolliffe, I. T. (2002). *Principal component analysis* (2nd ed.). Springer.
14. Kang, S. J., Kim, J. H., & Lee, J. Y. (2018). Application of principal component analysis to industrial process monitoring. *Journal of Process Control*, 61, 1–12.
15. Kelly, E. G., & Spottiswood, D. J. (1982). *Introduction to mineral processing*. John Wiley & Sons.
16. Kresta, J. V., Marlin, T. E., & MacGregor, J. F. (1991). Multivariate statistical monitoring of process operating performance. *The Canadian Journal of Chemical Engineering*, 69(1), 35–47.

17. Le, T., Nguyen, H., & Ghorbani, Y. (2020). Multivariate statistical process control in mineral processing: A review. *Minerals Engineering*, 148, 106201.
18. Li, X., et al. (2020). Clustering analysis of nickel-PGM flotation data. *Minerals*, 10(5), 412.
19. Madzivire, G., & Nyabeze, W. (2019). Improving PGM recovery using principal component analysis: A case study from Zimbabwe. *Journal of the Southern African Institute of Mining and Metallurgy*, 119(12), 1051–1058.
20. Mimosa Mining Company. (n.d.). Mimosa Platinum Mine. Retrieved from <https://www.mimosa.co.zw/>
21. Napier-Munn, T. J., Morrell, S., Morrison, R. D., & Kojovic, T. (1999). *Mineral comminution circuits: Their operation and optimization*. JKMRC, University of Queensland.
22. NS Energy Business. (n.d.). Mimosa Platinum Mine, Zimbabwe. Retrieved from <https://www.nsenerybusiness.com/projects/mimosa-platinum-mine/>
23. Rao, C. R. (1952). *Advanced statistical methods in biometric research*. John Wiley & Sons.
24. Saunders, M., Lewis, P., & Thornhill, A. (2019). *Research methods for business students* (8th ed.). Pearson.
25. Smith, A., Jones, B., & Williams, C. (2018). Limitations of principal component analysis in flotation data: A cautionary note. *Minerals Engineering*, 125, 10–17.
26. Vanhatalo, E., & Kulahci, M. (2015). The effect of autocorrelation on the performance of principal component analysis for process monitoring. *Quality Engineering*, 27(4), 423–436.
27. Wei, D., & Craig, I. K. (2014). Plant performance indicators in mineral processing. *IFAC Proceedings Volumes*, 47(3), 5224–5229.
28. Wills, B. A., & Finch, J. A. (2016). *Wills' mineral processing technology: An introduction to the practical aspects of ore treatment and mineral recovery* (8th ed.). Butterworth-Heinemann.
29. Wise, B. M., & Gallagher, N. B. (1996). The process chemometrics approach to process monitoring and fault detection. *Journal of Process Control*, 6(6), 329–348.
30. Wold, S., Esbensen, K., & Geladi, P. (1987). Principal component analysis. *Chemometrics and Intelligent Laboratory Systems*, 2(1-3), 37–52.
31. Mining Zimbabwe. (2026). Government affirms strong support as Mimosa eyes life-of-mine extension amid 40% revenue drop. Retrieved from <https://miningzimbabwe.com/government-affirms-strong-support-as-mimosa-eyes-life-of-mine-extension-amid-40-revenue-drop/>

APPENDICES

Appendix A: Preparation and Data Cleaning

The following R code was used to prepare the raw plant data for analysis. It selects the relevant variables, removes missing observations, ensures all values are positive for logarithmic transformation, and creates the log-transformed variables used in the log-normal model.

```
# DATA CLEANING FOR DF_clean
raw_data <- DF      # or read.csv("plant_data.csv")

required_cols <- c("Recovery..", "Final.feed.grade", "Final.conc.grade",
                  "Final.tail.grade", "Crushed.grind")
data_clean <- raw_data[, required_cols]
data_clean <- na.omit(data_clean)

# Ensure all values are positive (required for log transformation)
stopifnot(all(data_clean > 0))

# Create log-transformed variables (used in the log-linear model)
data_clean$log_Recovery <- log(data_clean$Recovery..)
data_clean$log_Feed <- log(data_clean$Final.feed.grade)
data_clean$log_Conc <- log(data_clean$Final.conc.grade)
data_clean$log_Tail <- log(data_clean$Final.tail.grade)
data_clean$log_Grind <- log(data_clean$Crushed.grind)

DF_clean <- data_clean
cat("DF_clean created with", nrow(DF_clean), "observations and",
    ncol(DF_clean), "columns\n")
head(DF_clean)
```

Appendix B: Diagnostic Tests

The following R code performs a comprehensive set of diagnostic checks on the log-normal regression model. It evaluates the normality of residuals, tests for heteroscedasticity, assesses multicollinearity using variance inflation factors, checks linearity, and identifies influential observations.

```
# DIAGNOSTIC TESTS ON DF_clean (LOG-LINEAR MODEL)
# Tests: Normality, Heteroscedasticity, Multicollinearity,
#        Linearity (RESET), Influential observations (Cook's D)

library(lmtest); library(car); library(ggplot2); library(gridExtra)
library(sandwich); library(moments); library(tseries); library(nortest)

model <- lm(log_Recovery ~ log_Feed + log_Conc + log_Tail + log_Grind,
            data = DF_clean)
summary(model)

# Normality of residuals
residuals <- residuals(model)
shapiro <- shapiro.test(residuals)
ad <- nortest::ad.test(residuals)
jbb <- tseries::jarque.bera.test(residuals)

# Heteroscedasticity
bp <- lmtest::bptest(model)
ncv <- car::ncvTest(model)

# Multicollinearity
vif_vals <- car::vif(model)
```

```
# Linearity (RESET test)
reset <- lmtest::resettest(model, power = 2:3)

# Influential observations (Cook's distance > 4/n)
cooks <- cooks.distance(model)
threshold <- 4 / length(cooks)
influential <- which(cooks > threshold)

# Diagnostic plots
par(mfrow = c(2, 2))
plot(model, which = 1, main = "Residuals vs Fitted")
plot(model, which = 2, main = "Q-Q Plot")
plot(model, which = 3, main = "Scale-Location")
plot(model, which = 4, main = "Cook's Distance")
par(mfrow = c(1, 1))
```

Appendix C: Model Training and Validation

The following R code implements the training and validation procedure for the three models considered in this study: Multiple Linear Regression, log-normal regression, and a GLM with a gamma distribution and log-link function. The dataset was randomly split into a training set (70% of observations) and a test set (30%). Each model was fitted on the training set, and its predictive performance was evaluated on the unseen test set using RMSE, MAE, MAPE, and R-squared.

```
# MODEL TRAINING AND VALIDATION (70/30 SPLIT)
# Models: MLR, Log-Linear, GLM Gamma
# Metrics: RMSE, MAE, MAPE, R-squared (AIC/BIC for log-linear)

library(ggplot2); library(dplyr); library(car); library(lmtest); library(sandwich)

set.seed(123)
n_total <- nrow(DF_clean)
train_idx <- sample(seq_len(n_total), size = floor(0.7 * n_total))
train_data <- DF_clean[train_idx, ]
test_data <- DF_clean[-train_idx, ]

# Train models
model_mlr <- lm(Recovery.. ~ Final.feed.grade + Final.conc.grade +
  Final.tail.grade + Crushed.grind, data = train_data)
model_log <- lm(log_Recovery ~ log_Feed + log_Conc + log_Tail + log_Grind,
  data = train_data)
model_gamma <- glm(Recovery.. ~ Final.feed.grade + Final.conc.grade +
  Final.tail.grade + Crushed.grind,
  family = Gamma(link = "log"), data = train_data)

# Predict on the test set
test_data$pred_mlr <- predict(model_mlr, newdata = test_data)
test_data$pred_log <- exp(predict(model_log, newdata = test_data))
test_data$pred_gamma <- predict(model_gamma, newdata = test_data, type = "response")

# Performance metrics
calc_metrics <- function(actual, predicted) {
  resid <- actual - predicted
  c(RMSE = sqrt(mean(resid^2)),
    MAE = mean(abs(resid)),
    MAPE = mean(abs(resid / actual)) * 100,
    R2 = 1 - sum(resid^2) / sum((actual - mean(actual))^2))
}

metrics_mlr <- calc_metrics(test_data$Recovery.., test_data$pred_mlr)
metrics_log <- calc_metrics(test_data$Recovery.., test_data$pred_log)
metrics_gamma <- calc_metrics(test_data$Recovery.., test_data$pred_gamma)
```

```
comparison <- data.frame(  
  Model = c("MLR", "Log-Linear (Log-Normal)", "GLM Gamma"),  
  rbind(metrics_mlr, metrics_log, metrics_gamma)  
)  
print(comparison)  
  
# Save trained models and outputs  
saveRDS(model_mlr, "MLR_Trained.rds")  
saveRDS(model_log, "LogLinear_Trained.rds")  
saveRDS(model_gamma, "GLM_Gamma_Trained.rds")  
write.csv(comparison, "Test_Performance.csv", row.names = FALSE)
```