University of Helsinki
Faculty of Science
Master's Programme in Mathematics and Statistics

Master's thesis

Theory and Application of Linear Regression Models with Box-Cox Transformations

Pabasari Amarasinghe

Supervisor:
Leena Kalliovirta

May 2025

Title: Theory and Application of Linear Regression Models with Box-Cox Transformations
Author: Pabasari Amarasinghe
Month and year: May 2025
Page count: 49 pp.

Abstract:

This thesis investigates the theoretical framework and applied methodology of linear regression models incorporating Box-Cox transformations, with a focus on improving model performance and interpretability in real-world data analysis. It is motivated by the challenge of modeling agricultural outcomes, particularly the relationship between soil nutrient concentrations and corn yield. Traditional linear regression models often struggle with data that violate assumptions of normality, linearity, and homoscedasticity. To address these limitations, Box-Cox transformations are applied to both the response and explanatory variables to stabilize variance, normalize residuals, and improve the linear structure of the data.

The empirical analysis of the thesis is based on a dataset comprising 215 field observations from a geographically distributed agricultural field. The analysis develops and compares several regression models, starting with a baseline model using untransformed variables and progressively incorporating Box-Cox transformed response and predictor variables. Variable selection is guided by stepwise procedures aimed at identifying parsimonious models that balance goodness-of-fit with model simplicity. Model evaluation relies on diagnostic plots and residual analysis to assess assumptions and performance.

This thesis emphasizes the practical value of transformation techniques, particularly the Box-Cox method in enhancing linear regression models for agricultural data. The approach improves adherence to statistical assumptions, enables more reliable inference, and increases predictive accuracy. These findings have broader implications for researchers and practitioners working with complex, non-normal data in applied fields such as agronomy, environmental science, and resource management.

Keywords: Box-Cox transformation, linear regression, agriculture

Contents

Chapter 1

Introduction

Linear regression models are fundamental tools in statistical analysis, widely used to describe the relationship between a continuous response variable and one or more explanatory variables. These models provide interpretable coefficients and allow for hypothesis testing and prediction in diverse fields such as agriculture, economics, medicine, and engineering. However, linear regression relies on assumptions including linearity, normality, and homoscedasticity (constant variance) of residuals, which are often violated in practical applications. When these assumptions do not hold, model estimates can be biased or inefficient, and inference may become unreliable. To address these challenges, transformation techniques are employed to improve the validity of linear regression model assumptions. Among these, the Box-Cox transformation stands out as a flexible and systematic approach to identify appropriate power transformations for both the response and explanatory variables. This family of transformations can stabilize variance, normalize residuals, and linearize relationships, thereby enhancing the accuracy and interpretability of regression models.

The aim of this thesis is to investigate how Box-Cox transformations can be used to improve the performance and interpretability of linear regression models when assumptions are violated. This includes both a theoretical review and an empirical application using agricultural data. The theoretical component discusses the statistical foundations of linear regression, the formulation and estimation of Box-Cox transformations, and diagnostic tools used to assess model adequacy. It also considers the implications of transformation choices for statistical inference and the role of stepwise model selection procedures in balancing model complexity and fit. The empirical part applies these methods to an agricultural dataset, analyzing the relationship between chemical elements and corn yield across a spatially structured 16-hectare field. The dataset contains measurements of multiple soil nutrients and yield observations from 215 sampling locations. Based on the preliminary analysis of the dataset, the Box-Cox transformations are applied to the

response and the explanatory variables to improve model performance and uncover underlying patterns. A sequence of regression models is developed, starting from simple additive forms and progressing to more complex models including interaction terms among chemical elements, with the aim of capturing the key factors influencing yield variation.

In Chapter 2, the theoretical background of linear regression models is presented, including the underlying assumptions and the limitations that arise when these assumptions are violated. The Box-Cox transformation is introduced as a method to address such issues, with emphasis on its formulation, estimation, and role in improving model validity. Chapter 3 presents the dataset, details the preliminary analysis conducted to assess distributional characteristics and guide the transformation strategy. A series of regression models are developed and evaluated further. Finally, Chapter 4 concludes the thesis by summarizing the main findings, discussing limitations, and suggesting directions for future research.

Chapter 2

Methodology

This chapter presents the methodological framework used to analyze the relationship between soil nutrient concentrations and corn yield. It focuses on linear regression models enhanced by Box-Cox transformations, which improve model performance and help satisfy key statistical assumptions. The chapter begins by defining essential statistical concepts, including random variables, the statistical model, and the likelihood function. It then details the model-building process, covering data transformation, variable selection strategies, and the evaluation criteria used to compare alternative models. Diagnostic tools for assessing model fit and the validity of assumptions are discussed further. Together, these methods provide the foundation for the analysis presented in the data analysis chapter.

2.1 Multiple Linear Regression

2.1.1 Model Specification

Multiple linear regression attempts to model the relationship between two or more explanatory variables and a response variable by fitting a linear equation to observed data. Here we use multiple linear regression to model corn yield as a function of soil quality measurements. A clear understanding of the statistical foundation of this model is essential to ensure that the modeling choices and subsequent interpretations are valid. As described by Saikkonen (2017), in multiple linear regression, the relationship between the dependent variable and the explanatory variables is assumed to be linear. Let $Y_1, \ldots, Y_n$ represent random variables corresponding to each observational unit, and $y_1, \ldots, y_n$ their observed values. Then the form of a linear model can be expressed as

$$Y_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \varepsilon_i, \quad i = 1, \ldots, n, \tag{2.1}$$

where x_{ij} are observed values of the explanatory variables, $\beta_1, \ldots, \beta_p$ are unknown parameters to be estimated, and ε_i is the random error term as-

sociated with observation i which accounts for the variability not explained by the predictors. The explanatory variables x_{ij} are either fixed values or treated as nonrandom quantities. The combination $\beta_1 x_{i1} + \cdots + \beta_p x_{ip}$ is known as the systematic part or the structure of the model.
Saikkonen (2017) mentioned, the term linearity in the context of regression means that the systematic part is a linear function of the parameters $\beta_1, \ldots, \beta_p$ and that the error term is added separately. This implies that transformations applied to explanatory variables, such as $x_{i2} = x_{i1}^2$ or $x_{i2} = \ln(x_{i1})$, introduce nonlinearity with respect to the predictors but do not affect the linearity of the model concerning the parameters $\beta_1, \ldots, \beta_p$. The regression model (2.1) must be supplemented by specifying the joint probability distribution of the observations. This step ensures that the model is well-defined in a statistical sense. If we assume that the error terms follow a normal distribution with constant variance, we can express the linear model as

$$Y_1, \ldots, Y_n \perp\!\!\!\perp, \quad Y_i \sim \mathcal{N}(\mathbf{x}_i'\boldsymbol{\beta}, \sigma^2), \quad \boldsymbol{\beta} \in \mathbb{R}^p, \quad \sigma^2 > 0. \tag{2.2}$$

Here, the vector $\mathbf{x}_i = [x_{i1}, \ldots, x_{ip}]'$ represents the explanatory variables, while $\boldsymbol{\beta} = [\beta_1, \ldots, \beta_p]'$ contains the unknown parameters to be estimated. $Y_1, \ldots, Y_n$ are mutually independent random variables, each following a normal distribution with mean $\mathbf{x}_i'\boldsymbol{\beta}$ and variance σ^2. The assumption $\sigma^2 > 0$ ensures that the variance is positive, with σ^2 being a nuisance parameter rather than a parameter of primary interest. Therefore, the formulation (2.2) of the linear model corresponds to its statistical definition in statistical inference, where we are primarily concerned with estimating and making inferences about the parameter vector $\boldsymbol{\beta}$. However, in practice, the term linear model is sometimes used more loosely, without explicitly specifying the joint probability distribution of the observations or errors. In some cases, the assumption of uncorrelated errors is used instead of assuming full independence. As noted by Rencher and Christensen (2012), the linear model (2.1) can be represented in the Matrix form as

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{2.3}$$

where

$$\mathbf{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}, \quad \mathbf{X} = \begin{bmatrix} x_{11} & \ldots & x_{1p} \\ x_{21} & \ldots & x_{2p} \\ \vdots & \vdots & \vdots \\ x_{n1} & \ldots & x_{np} \end{bmatrix}, \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{bmatrix} \quad \text{and} \quad \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}.$$

Since the errors are assumed to be independent and normally distributed with the expectation of zero and the covariance matrix of $\sigma^2\mathbf{I}_n$, we know that $\boldsymbol{\varepsilon} \sim \mathcal{N}\left(\mathbf{0}, \sigma^2\mathbf{I}_n\right)$, where $\mathbf{I}_n$ is the $(n \times n)$ identity matrix. Then the response variables follow a multivariate normal distribution.

$$\mathbf{Y} \sim \mathcal{N}(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}_n), \quad \boldsymbol{\beta} \in \mathbb{R}^p, \quad \sigma^2 > 0. \tag{2.4}$$

Following Saikkonen (2017), in statistical inference, the parameters $\boldsymbol{\beta}$ and σ^2 of a linear model are usually analyzed using the normal distribution representation (2.4) without explicitly considering the error terms. In some cases, this form can be derived naturally by starting with the assumption that observations are independent and follow a normal distribution

$$Y_i \sim \mathcal{N}(\mu_i, \sigma^2).$$

Here, the expected value μ_i can be assumed to have a linear representation, $\mu_i = \mathbf{x}_i'\boldsymbol{\beta}, \quad (i = 1, \ldots, n)$. This assumption implies that the response variable is normally distributed around a mean that depends linearly on the predictor variables. Error terms in the model represent deviations from the expected values and can be interpreted as measurement errors or unmodeled influences. Understanding these error terms is crucial for evaluating model assumptions. Two key concepts related to errors are fitted values and residuals. The fitted value for an observation is given by $\hat{\mu}_i = \mathbf{x}_i'\hat{\boldsymbol{\beta}}$, while the residual is defined as $\hat{\varepsilon}_i = y_i - \mathbf{x}_i'\hat{\boldsymbol{\beta}}$. These residuals provide information about the variance parameter σ^2 and can help assess the validity of model assumptions.

2.1.2 Residual Diagnostics

As noted by Weisberg (2005), under the assumptions of the linear regression model (2.2), the residuals, $\hat{\boldsymbol{\varepsilon}} = \mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}$, inherit specific properties. The expectation of the residuals is zero, $\mathbb{E}[\hat{\boldsymbol{\varepsilon}}] = \mathbf{0}$, provided that the model is correctly specified. The covariance of the residuals is given by

$$\mathrm{Cov}(\hat{\boldsymbol{\varepsilon}}) = \sigma^2(\mathbf{I}_n - \mathbf{H}),$$

where $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ is the hat matrix ($\mathbf{H}$ is called the hat matrix because it transforms the vector of observed responses $\mathbf{Y}$ into the vector of fitted responses $\hat{\mathbf{Y}}$). The diagonal elements, h_{ii}, of $\mathbf{H}$ measure the leverage of each observation, with high-leverage points exerting disproportionate influence on the fitted values. The residuals are not independent and exhibit heteroscedasticity, even when the true errors ε_i are homoscedastic. This is because

$$\mathrm{Var}(\hat{\varepsilon}_i) = \sigma^2(1 - h_{ii}).$$

Consequently, careful interpretation of diagnostic plots and tests is necessary to ensure accurate model evaluation. As emphasized by Weisberg (2005) key plots include residuals and fitted values, which help detect non-linearity, heteroscedasticity, or outliers. The ideal pattern is random scatter around zero with constant variance, while violations such as curvature suggest model misspecification and funnel shapes indicate heteroscedasticity. A normal Q-Q plot is used to assess the normality of residuals, with the ideal pattern being points aligning closely with the 45 degree line. Deviations from normality,

such as heavy tails or skewness, indicate violations, and are visible in the Q-Q plot. These graphical methods are rooted in foundational work by Wilk and Gnanadesikan (1968), who formalized probability plotting techniques for data analysis. While Q-Q plots are most common for normality assessment emphasizing tail behavior through quantile comparisons their P-P plots offer complementary insights by comparing cumulative distribution functions, making them more sensitive to deviations in the distribution's center. As demonstrated in their work, Q-Q plots better detect tail anomalies, whereas P-P plots excel at identifying shifts of location or kurtosis. This theoretical framework underscores why modern diagnostics often use Q-Q plots for normality checks while maintaining P-P plots as specialized tools for specific distributional assessments. The scale-location plot evaluates homoscedasticity, where the ideal pattern is a horizontal band of residuals without a trend. The standardized residuals and leverage plot is another important graphical diagnostic tool that helps identify influential observations and potential issues with model fit. Building on the foundational framework for residuals developed by Cox and Snell (1968), a particularly useful measure in this context is Cook's Distance, introduced by R. D. Cook (1977), which combines both leverage and the residuals' magnitude to assess the influence of each data point on the estimated regression coefficients. Cook's Distance is calculated as

$$D_i = \frac{(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}^{(i)})'(\mathbf{X}'\mathbf{X})(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}^{(i)})}{p\hat{\sigma}^2},$$

where $\hat{\boldsymbol{\beta}}^{(i)}$ represents the estimated regression coefficients when the i-th observation is omitted from the dataset and p is the number of parameters. As demonstrated by R. D. Cook (1977), Cook's Distance provides a measure of the change in the estimated coefficients when a particular observation is removed. Observations with Cook's Distance greater than 1 ($D_i > 1$) are considered influential and should be closely examined, as their removal could lead to a significant change in the model's results. Additionally, as mentioned by Saikkonen (2017), residual analysis is useful for detecting variations in variance, dependence structures, or deviations from normality in errors. Such diagnostics can be more informative than directly examining the original observations. However, in many cases, minor deviations from normality are not problematic, as theoretical results can still hold under asymptotic approximations. Nonetheless, applying the linear model to datasets exhibiting strong deviations from normality may lead to incorrect inferences and should be avoided. In this thesis, residual analysis was used to check if the assumptions of the linear regression models were met. By examining residual plots, we could identify any issues like non-linearity, heteroscedasticity, or influential outliers that could affect the model's validity. It turns out that the residual diagnostics refine the models in the empirical application.
The bootstrap is a resampling technique introduced by Efron and Tibshi-

rani (1994) that enables estimation of the sampling distribution of a statistic using only the observed data. It is especially valuable when analytical expressions for standard errors or confidence intervals are difficult to derive, or when classical inference assumptions, such as normality of residuals, may not hold. As a nonparametric and data-driven method, the bootstrap makes minimal assumptions about the underlying distribution of the data, making it broadly applicable to a wide range of statistical problems. In the bootstrap procedure, multiple resamples of the same size are drawn with replacement from the empirical distribution. For each resample, the statistic of interest is recomputed, resulting in a distribution of bootstrap replicates. This empirical distribution serves as an approximation to the true sampling distribution of the estimator. Using this, one can construct confidence intervals for unknown parameters without relying on asymptotic normality or explicit variance formulas.

Theoretically, let F be the unknown true distribution and $\hat{F}$ be the empirical distribution from which the bootstrap samples are drawn. As demonstrated by Efron et al. (1994), the estimator $\hat{\theta}$ calculated from the original data can be viewed as an estimate of a parameter $\theta = t(F)$. The bootstrap approximates the distribution of $\hat{\theta}$ under F by computing the distribution of $\hat{\theta}^*$ under $\hat{F}$. As the sample size increases, and under certain regularity conditions, the distribution of $\sqrt{n}(\hat{\theta} - \theta)$ converges to the same limit as that of $\sqrt{n}(\hat{\theta}^* - \hat{\theta})$, thereby justifying the bootstrap's validity. In the thesis, bootstrap confidence intervals are applied to the standardized residuals in the Q-Q plots to provide a more rigorous assessment of the residuals' conformity to the theoretical distribution, typically normality. The bootstrap approach captures the sampling variability of the residual quantiles without relying on parametric assumptions, which is especially important when model residuals may not perfectly meet classical normality conditions.

2.1.3 Model Fitting and Maximum Likelihood Estimation

By applying the probability density function of the multivariate normal distribution, the joint density function for vector $\mathbf{Y}$ is

$$f_{\mathbf{Y}}(\mathbf{y}; \boldsymbol{\beta}, \sigma^2) = (2\pi\sigma^2)^{-n/2} \exp\left\{ -\frac{1}{2\sigma^2} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \right\}. \tag{2.5}$$

This function expresses the likelihood of observing the data given the parameters $\boldsymbol{\beta}$ and σ^2. Then the log-likelihood function for the parameters $\boldsymbol{\beta}$ and σ^2 is

$$\ell(\boldsymbol{\beta}, \sigma^2) = -\frac{n}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}). \tag{2.6}$$

Differentiating $\ell(\boldsymbol{\beta}, \sigma^2)$ with respect to $\boldsymbol{\beta}$ and setting the gradient to zero gives

$$\frac{\partial \ell}{\partial \boldsymbol{\beta}} = \frac{1}{\sigma^2} \mathbf{X}'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) = \mathbf{0}. \tag{2.7}$$

Solving for $\boldsymbol{\beta}$, we obtain,

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}. \tag{2.8}$$

Similarly, differentiating the log-likelihood with respect to σ^2 and setting the derivative equal to zero yields

$$\hat{\sigma}^2 = \frac{1}{n}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})'(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}). \tag{2.9}$$

However, satisfying the first-order condition alone is not sufficient to guarantee that $\hat{\boldsymbol{\beta}}$ and $\hat{\sigma}^2$ are maximum points. It is necessary to verify the second-order condition. The Hessian is the matrix of second derivatives of the likelihood with respect to the parameters. To verify the maximum in the full parameter vector $(\boldsymbol{\beta}, \sigma^2)$, we consider the full Hessian matrix of the log-likelihood

$$H(\boldsymbol{\beta}, \sigma^2) = \begin{bmatrix} \dfrac{\partial^2 \ell}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}'} & \dfrac{\partial^2 \ell}{\partial \boldsymbol{\beta} \partial \sigma^2} \\ \dfrac{\partial^2 \ell}{\partial \sigma^2 \partial \boldsymbol{\beta}'} & \dfrac{\partial^2 \ell}{\partial (\sigma^2)^2} \end{bmatrix}.$$

Therefore we get

$$H(\boldsymbol{\beta}, \sigma^2) = \begin{bmatrix} -\dfrac{1}{\sigma^2}\mathbf{X}'\mathbf{X} & \dfrac{1}{(\sigma^2)^2}\mathbf{X}'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \\ \left[\dfrac{1}{(\sigma^2)^2}\mathbf{X}'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})\right]' & \dfrac{n}{2(\sigma^2)^2} - \dfrac{(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})}{(\sigma^2)^3} \end{bmatrix}.$$

Evaluating this at the MLE values $(\hat{\boldsymbol{\beta}}, \hat{\sigma}^2)$, we get

$$H(\hat{\boldsymbol{\beta}}, \hat{\sigma}^2) = \begin{bmatrix} -\frac{1}{\hat{\sigma}^2}\mathbf{X}'\mathbf{X} & \mathbf{0} \\ \mathbf{0} & -\frac{n}{2(\hat{\sigma}^2)^2} \end{bmatrix}.$$

Since both blocks on the diagonal are negative definite, the full Hessian is negative definite. This confirms that the log-likelihood function is strictly concave in the full parameter vector, and the MLE $(\hat{\boldsymbol{\beta}}, \hat{\sigma}^2)$ is indeed a global maximum.

After fitting the model, we need to determine how well the model fits the data. According to Saikkonen (2017) the coefficient of determination or the R^2 value is a statistical measure of how close the data are to the fitted regression line. It represents the proportion of the variance in the dependent variable that is predictable from the independent variables. Also, it provides a measure of how well observed outcomes are replicated by the model, based on the proportion of total variation of outcomes explained by the model. Considering the regression model (2.1), one fundamental quantity is the total sum of squares (SST), defined as

$$SST = \sum_{i=1}^{n}(y_i - \bar{y})^2,$$

where y_i represents the observed values, and

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i.$$

The regression sum of squares (SSR) is defined by

$$SSR = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2,$$

where $\hat{y}_i$ denotes the fitted values obtained from the regression model, and n is the total number of observations. Then the residual sum of squares (SSE) is given by

$$SSE = \sum_{i=1}^{n} \hat{\varepsilon}_i^2.$$

In models that include an intercept term, the total variation can be decomposed as

$$SST = SSR + SSE.$$

Then the coefficient of determination (R^2) is derived by

$$R^2 = 1 - \frac{SSE}{SST} = \frac{SSR}{SST}.$$

The R^2 value varies between 0 and 1 and closer to 1 suggests a better fit, meaning the model explains a larger fraction of the variability in the response variable. However, model selection should not rely solely on R^2, as adding more predictors always increases its value, potentially leading to overfitting. To address this, information criteria such as the Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) are used. As introduced by Akaike (1974), the AIC is a widely used metric that balances the goodness of fit of a model with its complexity and it is calculated as follows,

$$\text{AIC} = 2k - 2\ln(\hat{L}),$$

where k represents the number of parameters in the model (including the intercept and the error variance if applicable), and $\hat{L}$ is the maximized value of the likelihood function of the model. The AIC aims to penalize models with excessive parameters while rewarding models that fit the data well. A lower AIC value indicates a better model, as it suggests that the model achieves a good balance between fit and complexity. In practice, the model with the lowest AIC is usually selected as the optimal model. Similar to AIC, the Bayesian Information Criterion (BIC) also penalizes models for including too many parameters, but it does so more strongly, especially when

the sample size is large. The BIC was originally proposed by Schwarz (1978) and is defined as

$$\text{BIC} = k \ln(n) - 2 \ln(\hat{L}),$$

where k is the number of estimated parameters, n is the number of observations, and $\hat{L}$ is the maximized value of the likelihood function for the model. As stated by Kass and Raftery (1995), the BIC imposes a heavier penalty on the number of parameters compared to AIC, making it more conservative in selecting complex models. A lower BIC value indicates a better model, and the model with the lowest BIC is typically preferred when comparing multiple candidate models. BIC is particularly useful when the dataset is large, as it discourages over-fitting more strongly than AIC. These tools ensure that the model fits the data well while maintaining an appropriate level of complexity, thus guiding the selection of the most appropriate model for analyzing the relationship between dependent variable and the independent variables.

2.2 Box-Cox Transformation

The Box-Cox method, introduced by Box and Cox (1964), is a technique for selecting a suitable transformation of the response variable and the predictor variables in both simple and multiple regression models. This transformation is particularly useful when dealing with variables that exhibit skewness or heteroscedasticity, as it allows for a flexible adjustment of the scale of the data through a power parameter, denoted as λ. Heteroscedasticity violates the standard regression assumption of homoscedasticity (constant error variance) and can lead to inefficient estimates and invalid inference. The Box-Cox method addresses this by identifying an optimal λ to stabilize variance and improve normality, with common transformations. While the Box-Cox method is widely used for variance stabilization, Spitzer (1982) highlights that its effectiveness depends on the underlying distribution of the data. It is mentioned that the transformation is most reliable when the data exhibits monotonic heteroscedasticity (variance changes systematically with the mean). For cases where heteroscedasticity is non-monotonic or multimodal, alternative approaches (such as weighted least squares) may be more appropriate.

2.2.1 Transforming the Response variable

Definition and Normalization

As described by Weisberg (2005), the Box-Cox method can be applied to the response variable using a modified power transformation. In the context of the linear model introduced in Section 2.1, the response variable is denoted by Y_i, representing the outcome for the i-th observational unit. To

distinguish notation in this section and maintain consistency with the transformation framework, we refer to the response variable as W. For strictly positive values of W, the Box-Cox transformation is defined as

$$W^{(\lambda_w)} = \begin{cases} \frac{W^{\lambda_w}-1}{\lambda_w}, & \lambda_w \neq 0, \\ \log W, & \lambda_w = 0. \end{cases} \tag{2.10}$$

As stated by Atkinson (1985), to estimate λ_w, the likelihood function must account for the change in scale induced by the transformation. The likelihood of the transformed observations relative to the original observations includes the Jacobian (J)

$$J = \prod_{i=1}^{n} \left| \frac{\partial W_i^{(\lambda_w)}}{\partial W_i} \right|. \tag{2.11}$$

For the power transformation in (2.10), $\frac{\partial W_i^{(\lambda_w)}}{\partial W_i} = W_i^{\lambda_w - 1}$, so that taking the logarithm yields

$$\log J = (\lambda_w - 1) \sum_{i=1}^{n} \log W_i = n(\lambda_w - 1) \log \dot{W}, \tag{2.12}$$

where $\dot{W}$ is the geometric mean of the observations. The geometric mean of the response variable is given by

$$\dot{W} = \left(\prod_{i=1}^{n} W_i \right)^{1/n}$$

where W_i represents the i-th observation of the untransformed dependent variable. The geometric mean $\dot{W}$ is computed as a constant at the initialization stage using

$$\ln \dot{W} = \frac{1}{n} \sum_{i=1}^{n} \ln W_i.$$

Then the normalized Box-Cox transformation is defined as

$$W^{*(\lambda_w)} = \frac{W^{(\lambda_w)}}{J^{1/n}} = \frac{W^{\lambda_w} - 1}{\lambda_w \dot{W}^{\lambda_w - 1}}, \tag{2.13}$$

where J is the Jacobian determinant from Equation (2.11). This alternative transformation is preferred when applying Box-Cox transformation to the response variable and for several λ_w values it can be defined as

$$W^{(\lambda_w)} = \begin{cases} \frac{W^{\lambda_w} - 1}{\lambda_w \dot{W}^{\lambda_w - 1}}, & \lambda_w \neq 0, \\ \dot{W} \log W, & \lambda_w = 0. \end{cases} \tag{2.14}$$

As emphasized by Box et al. (1964), unlike other transformations that aim for linearity, the Box-Cox method is primarily used to improve the normality of the regression residuals. The idea is to find a transformation that makes the residuals as close to normally distributed as possible. In the context of improving the model, the Box-Cox transformation was used to address issues of skewness and heteroscedasticity in the data. As discussed by Fox (2016), the transformation allows for more flexible modeling by adjusting the scale of the data through a transformation parameter, λ_w. Since heteroscedasticity violates the assumption of constant error variance in linear regression models, applying this transformation helps stabilize variance and potentially improves the normality of the residuals, which enhances the accuracy of parameter estimates and inference. By applying the Box-Cox transformation to both the response and predictor variables, it would be easy to achieve a better fit and mitigate the impact of non-normality, leading to more reliable results in the final model.

Maximum Likelihood Estimation of λ_w

After applying the Box-Cox transformation to each observation we create a vector $\mathbf{W}^{(\lambda_w)}$

$$\mathbf{W}^{(\lambda_w)} = \begin{pmatrix} W_1^{(\lambda_w)} \\ W_2^{(\lambda_w)} \\ \vdots \\ W_n^{(\lambda_w)} \end{pmatrix},$$

Then for any fixed λ_w, we estimate the linear model

$$\mathbf{W}^{(\lambda_w)} = \mathbf{X}\beta + \boldsymbol{\varepsilon} \tag{2.15}$$

where $\mathbf{X}$ is a matrix containing the untransformed predictors and the coefficient vector $\boldsymbol{\beta}$ contains the regression parameters including the intercept. The error terms follow $\boldsymbol{\varepsilon} \sim \mathcal{N}(0, \sigma^2\mathbf{I}_n)$, indicating independent and identically distributed normal random variables with mean zero and constant variance σ^2. Draper and Smith (1998) states that the basic idea is that, if an appropriate λ_w could be found, an additive model with normally distributed, independent, and homogeneous error structure could be fitted by the maximum likelihood method. The maximum likelihood estimation of the transformation parameter λ_w proceeds through the following steps. First, we select a range of candidate values for λ_w, typically within the interval $(-1, 1)$, though sometimes extended to $(-2, 2)$ for initial exploration. This range is covered using 11 to 21 equally spaced values, with the option to refine promising sub-intervals later if necessary. For each candidate λ_w value, we compute the transformed response vector $\mathbf{W}^{(\lambda_w)}$ according to the Box-Cox transformation formula. Special care is taken at $\lambda_w = 0$, where we use the logarithmic transformation. The transformed data is then fitted using

ordinary least squares regression to the model $\mathbf{W}^{(\lambda_w)} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, and the corresponding residual sum of squares $SSE(\lambda_w)$ is recorded. Following Box et al. (1964), we consider the linear model (2.15) where the transformed response variable follows a normal distribution with mean $\mathbf{X}\boldsymbol{\beta}$ and variance σ^2. The likelihood function for the model (2.15) is given by

$$L(\boldsymbol{\beta}, \lambda_w, \sigma^2) = (2\pi\sigma^2)^{-n/2} \exp\left(-\frac{1}{2\sigma^2}(\mathbf{W}^{(\lambda_w)} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{W}^{(\lambda_w)} - \mathbf{X}\boldsymbol{\beta})\right) J(\lambda_w, \mathbf{W}) \tag{2.16}$$

where $J(\lambda_w, \mathbf{W}) = \prod_{i=1}^{n} |\partial W_i^{(\lambda_w)}/\partial W_i| = \prod_{i=1}^{n} W_i^{\lambda_w - 1}$ is the Jacobian of the transformation. The estimation procedure for the Box-Cox transformation parameter λ_w based on maximum likelihood is proposed. For a fixed value of λ_w, the maximum likelihood estimate of the regression coefficients is given by

$$\hat{\boldsymbol{\beta}}(\lambda_w) = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}^{(\lambda_w)}.$$

The corresponding estimate of the error variance is

$$\hat{\sigma}^2(\lambda_w) = \frac{SSE(\lambda_w)}{n} = \frac{(\mathbf{W}^{(\lambda_w)} - \mathbf{X}\hat{\boldsymbol{\beta}})'(\mathbf{W}^{(\lambda_w)} - \mathbf{X}\hat{\boldsymbol{\beta}})}{n},$$

where $SSE(\lambda_w)$ denotes the sum of squared errors for the model fitted with transformation parameter λ_w. Maximizing the likelihood (2.16) over λ_w is equivalent to maximizing the corresponding log-likelihood (up to an additive constant), which simplifies to

$$\ell(\lambda_w) = -\frac{n}{2}\log(SSE(\lambda_w)/n) + (\lambda_w - 1)\sum_{i=1}^{n} \ln W_i, \tag{2.17}$$

where $SSE(\lambda_w) = n\hat{\sigma}^2(\lambda_w)$ The estimation proceeds by plotting $SSE(\lambda_w)$ versus λ_w (or alternatively $\ln SSE(\lambda_w)$ versus λ_w when the residual sums vary widely in magnitude). A smooth curve is fitted to these points, and the value $\hat{\lambda}_w$ that minimizes this curve is identified as the maximum likelihood estimate. In practice, we often round $\hat{\lambda}_w$ to the nearest interpretable value from the sequence $-2, -1, -0.5, 0, 0.5, 1, 2$, though some applications may warrant finer gradations or retention of the exact estimate. To aid visualization, we often normalize the log-likelihood as

$$\ell^*(\lambda_w) = \ell(\lambda_w) - \ell(\hat{\lambda}_w),$$

so that the maximum value of ℓ^* is zero. To construct an approximate $100(1-\alpha)\%$ confidence interval for the Box-Cox transformation parameter λ_w, we use the likelihood ratio approach. We consider those values of λ_w that satisfy

$$\ell(\hat{\lambda}_w) - \ell(\lambda_w) \leq \frac{1}{2}\chi_1^2(1-\alpha), \tag{2.18}$$

where $\ell(\lambda_w)$ is the log-likelihood defined in equation (2.17), and $\chi^2_1(1-\alpha)$ is the $(1-\alpha)$-quantile of the chi-squared distribution with 1 degree of freedom. This condition defines an interval around $\hat{\lambda}_w$ where the log-likelihood is not significantly smaller than its maximum. For instance, for a 95% confidence level, we use $\chi^2_1(0.95) = 3.84$, yielding a cutoff of approximately

$$\ell(\hat{\lambda}_w) - \ell(\lambda_w) \leq 1.92.$$

In practice, this means drawing a horizontal line at height $\ell(\hat{\lambda}_w) - \frac{1}{2}\chi^2_1(1-\alpha)$ on the plot of $\ell(\lambda_w)$ versus λ_w, and finding the two intersection points with the curve. These two values of λ_w form the endpoints of the approximate confidence interval. Equivalently, this can be expressed in terms of the profile sum of squares $SSE(\lambda_w)$, defined as the residual sum of squares from the model with transformation parameter λ_w. In this case, we cut the curve at

$$SSE(\lambda_w) \leq SSE(\hat{\lambda}_w) \exp\left\{\frac{\chi^2_1(1-\alpha)}{n}\right\}, \tag{2.19}$$

or, on the log scale,

$$\ln S(\lambda_w) \leq \ln S(\hat{\lambda}_w) + \frac{\chi^2_1(1-\alpha)}{n}. \tag{2.20}$$

Both forms provide an equivalent rule for constructing a confidence interval, the endpoints are those values of λ_w at which the residual sum of squares increases by no more than a factor determined by the chi-squared distribution. The final choice of λ_w considers both statistical optimality and practical interpretability, with the selected value then used to transform the data for all subsequent analyses.

2.2.2 Transforming the Predictor variable

For predictor variables, transformations are often applied to achieve a more linear relationship with the response variable or to stabilize variance. In the framework established in Section 2.1, the explanatory variables are denoted by x_{ij}, representing the value of the j-th predictor for the i-th observational unit. To avoid notational confusion with earlier sections, we denote the predictor variables in this context as v_j, which play the same role as x_{ij}, but are used here for transformation purposes as introduced in Section 2.2. The transformation is defined as

$$v_j^{(\lambda_j)} = \begin{cases} \frac{v_j^{\lambda_j}-1}{\lambda_j}, & \text{if } \lambda_j \neq 0, \\ \log(v_j), & \text{if } \lambda_j = 0. \end{cases}$$

As Weisberg (2005) demonstrates, this scaled power transformation ensures continuity across all values of λ, including the logarithmic case as λ ap-

proaches zero, and preserves the direction of association between the predictor and response variables, unlike basic power transformations where negative λ values invert the relationship. The scaled transformation is particularly advantageous, as it maintains interpretability while allowing flexibility in selecting the optimal λ. In practice, the transformation parameter λ is chosen to minimize the residual sum of squares in the regression model.

2.2.3 Deriving the Likelihood Function and Covariance Matrix for the fully transformed model

Velilla (1993) discussed an extension in the context of transforming multivariate data to normality, which is beneficial for multivariate regression or other multivariate analyses. Spitzer (1982) discusses the estimation of transformation parameters in the context of regression models where both the response and predictor variables may require transformation. The transformed regression model can be expressed as

$$W_i^{(\lambda_w)} = \gamma_0 + \gamma_1 v_{i1}^{(\lambda_1)} + \gamma_2 v_{i2}^{(\lambda_2)} + \cdots + \gamma_p v_{ip}^{(\lambda_p)} + \varepsilon_i^{(\lambda)}, \tag{2.21}$$

where $W_i^{(\lambda_w)}$ represents the Box-Cox transformed response variable for the i^{th} observation, γ_0 denotes the intercept term, γ_j $(j = 1, \ldots, p)$ are the regression coefficients corresponding to each transformed predictor $v_{ij}^{(\lambda_j)}$, and $\varepsilon_i^{(\lambda)} \sim \mathcal{N}(0, \sigma^2)$ represents independent normally distributed error terms specific to the transformed model. In matrix form, the model becomes

$$\mathbf{W}^{(\lambda_w)} = \mathbf{V}^{(\boldsymbol{\lambda})}\boldsymbol{\gamma} + \boldsymbol{\varepsilon}^{(\lambda)}, \tag{2.22}$$

where $\mathbf{W}^{(\lambda_w)}$ represents the $n \times 1$ vector of transformed response variables, $\mathbf{V}^{(\boldsymbol{\lambda})}$ denotes the $n \times (p+1)$ design matrix (including a column of ones for the intercept), $\boldsymbol{\lambda} = (\lambda_w, \lambda_1, \ldots, \lambda_p)'$ is the vector of Box-Cox transformation parameters for both the response and predictors, and $\boldsymbol{\gamma} = (\gamma_0, \gamma_1, \ldots, \gamma_p)'$ represents the vector of regression coefficients corresponding to each transformed predictor. Under the assumption that there exists $\boldsymbol{\lambda}$, the vector of transformation parameters such that the errors $\varepsilon_i^{(\lambda)}$ are assumed to be normally distributed with mean zero and constant variance σ^2, the density function for the$i$-th error term is given by

$$f(\varepsilon_i^{(\lambda)}) = (2\pi\sigma^2)^{-1/2} \exp\left(-\frac{(\varepsilon_i^{(\lambda)})^2}{2\sigma^2}\right). \tag{2.23}$$

The likelihood function for the transformed model then takes the form

$$L^*(\gamma, \lambda, \sigma^2; v, W) = \prod_{i=1}^{n} f(\varepsilon_i^{(\lambda)}) \cdot W_i^{\lambda_w - 1} \tag{2.24}$$

where the last term represents the Jacobian of the transformation from $\varepsilon_i^{(\lambda)}$ to W. Taking the logarithm of (2.24) yields the log-likelihood function

$$\begin{aligned}\ell((\boldsymbol{\gamma}, \boldsymbol{\lambda}, \sigma^2) = \ln L^* = k - \frac{n}{2}\ln\sigma^2 \\ - \frac{1}{2\sigma^2}\sum_{i=1}^{n}\Big(W_i^{*(\lambda_w)} - \gamma_0 - \gamma_1 v_{i1}^{(\lambda_1)} - \gamma_2 v_{i2}^{(\lambda_2)} \\ - \cdots - \gamma_p v_{ip}^{(\lambda_p)}\Big)^2 \\ + (\lambda_w - 1)\sum_{i=1}^{n}\ln W_i,\end{aligned} \tag{2.25}$$

where k is a constant term and n denotes the number of observations. Let the complete parameter vector be $\boldsymbol{\theta} = (\boldsymbol{\gamma}', \boldsymbol{\lambda}', \sigma^2)'$ where $\boldsymbol{\gamma} = (\gamma_0, \gamma_1, \ldots, \gamma_p)'$ are the regression coefficients and $\boldsymbol{\lambda} = (\lambda_w, \lambda_1, \ldots, \lambda_p)'$ contains the transformation parameters of the predictors. The log-likelihood function $\ell(\boldsymbol{\gamma}, \boldsymbol{\lambda}, \sigma^2)$ is a monotonic transformation of the original likelihood (2.24), ensuring both yield identical parameter estimates when maximized. Under the assumption of normality of the $\boldsymbol{\varepsilon}^{(\lambda)}$ obtains estimators $\hat{\boldsymbol{\gamma}}$, $\hat{\boldsymbol{\lambda}}$ and $\hat{\sigma}^2$ which are Best Asymptotically Normal (BAN) under general regularity conditions. The asymptotic covariance matrix of the parameter estimates is given by the inverse expected negative Hessian

$$\left(-E\left[\frac{\partial^2\ell(\boldsymbol{\gamma}, \boldsymbol{\lambda}, \sigma^2)}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\right]\right)^{-1}, \tag{2.26}$$

which attains the Cramér-Rao lower bound. Regardless of the estimation method used for the Box-Cox transformed model in (2.21), we can consistently estimate the parameter covariances using the inverse Hessian matrix (2.26). In practice, the expected value of the second derivative matrix cannot be obtained for functions which contain the Box-Cox transformation, the expectations in (2.26) are too complex to evaluate. Therefore, we instead compute the observed inverse Hessian matrix

$$(-\frac{\partial^2\ell(\boldsymbol{\gamma}, \boldsymbol{\lambda}, \sigma^2)}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'})^{-1}. \tag{2.27}$$

As established by Goldfeld and Quandt (1972), (2.26) is consistently estimated by (2.27) if the estimators are sufficient. Therefore, the use of (2.27) for statistical inference is justified. The Box-Cox transformation can be extended to multivariate settings where the model has several response variables. Velilla (1993) and Gnanadesikan (1977) discussed this extension in the context of transforming multivariate data to normality, which is beneficial for multivariate regression or other multivariate analyses. This extension allows for greater flexibility when analyzing complex datasets with multiple interdependent variables.

2.3 Moran's I test

In this thesis, Moran's I test (introduced by Moran (1950)) was used to assess the spatial autocorrelation of corn yield and predictor variables to investigate potential spatial dependencies across the 215 locations in the dataset. Anselin (1995) stated that the spatial autocorrelation measures the degree to which the values of a variable at one location are similar to the values of the same variable at neighboring locations. If spatial autocorrelation exists, observations located near each other may exhibit similar or dissimilar values, violating the assumption of independence that is required for traditional regression models. Moran's I is defined as,

$$I = \frac{n}{G} \cdot \frac{\sum_{k=1}^{n} \sum_{l=1}^{n} g_{kl}(y_k - \overline{y})(y_l - \overline{y})}{\sum_{k=1}^{n}(y_k - \overline{y})^2} \tag{2.28}$$

where y_k and y_l represent the observed values at locations k and l, respectively, $\overline{y}$ denotes the mean of the observed values, g_{kl} is the spatial weight between locations k and l, n is the total number of observations, and G represents the sum of all spatial weights. As described by Li, Calder, and Cressie (2007), the test statistic ranges from -1 to +1. A value close to +1 indicates strong positive spatial autocorrelation, where nearby observations are similar. A value close to 0 suggests no spatial autocorrelation (i.e., a random spatial pattern), and a value close to -1 indicates negative spatial autocorrelation, where nearby observations are dissimilar. The statistical significance of Moran's I is evaluated by comparing the observed value of the statistic to a distribution generated under the null hypothesis, which assumes no spatial autocorrelation. The null hypothesis (H_0) posits that the values of the variable are randomly distributed across space. The alternative hypothesis (H_1) suggests that there is significant spatial autocorrelation either positive (clustering) or negative (dispersion). According to Getis and Ord (1992), a p-value below the significance threshold (typically 0.05) would indicate that the null hypothesis is rejected, suggesting that the spatial distribution of the variable is non-random.

Chapter 3

Data Analysis

3.1 Introduction

This thesis utilizes a dataset extracted from Chapter 3 in the book Methods of Multivariate Analysis by Rencher et al. (2012). The corresponding SAS command files and access to the official FTP server were provided in the book. The dataset originates from Baker Field in 1997, a 16-hectare experimental site extensively described by Colvin, Jaynes, Karlen, Laird, and Ambuel (1997). Situated within the Clarion-Nicollet-Webster soil association in central Iowa USA, the field is characterized by low-relief, swell-and-swale topography.

As documented by Colvin et al. (1997), data collection followed established precision agriculture research protocols, ensuring comprehensive spatial coverage. The field was divided into 8 transects, each containing 28 equally spaced sampling points. There were 9 missing observations, resulting in a total of 215 data points available for analysis. The data cleaning and preparation were carried out by Colvin et al. (1997). The transects are run in an east-west direction and were designed to capture the spatial variability of soil quality and crop yield throughout the field. The corn yield (in bushels per acre) and ten soil nutrient concentrations of chemical elements Boron (B), Calcium (Ca), Copper (Cu), Iron (Fe), Potassium (K), Magnesium (Mg), Manganese (Mn), Sodium (Na), Phosphorus (P), and Zinc (Zn) were recorded and concentrations were likely measured in parts per million (ppm), which is standard for micronutrient concentrations in soil analysis. Yield measurements were obtained using combine mounted yield monitors, integrated with positional data collected using GPS technology. Soil samples were analyzed to determine nutrient concentrations at each sampling location.

This thesis employs multiple linear regression models to analyze the relationship between soil nutrients and corn yield. To address issues such as non-normality and heteroscedasticity, Box-Cox transformations were applied

where necessary. By incorporating transformation techniques and variable selection criteria, this thesis aims to develop statistical models that enhance understanding of the soil-yield relationship while applying the theory of linear regression models and the Box-Cox Transformations. Five regression models were developed for this analysis. These models provide a comprehensive framework for evaluating the influence of soil nutrient concentrations on corn yield.

Analysis was conducted using *R* version (`v4.5.0`; R Core Team 2025), a widely used language for statistical computing and graphics. Several R packages were employed to support the data analysis. The *readxl* package (`v1.4.5`; Wickham and Bryan 2023) was used to import the Excel dataset into *R*. The *ggplot2* package (`v3.5.2`; Wickham 2016) was used to create high quality visualizations for preliminary data analysis and residual diagnostics. To assess correlations between variables, scatterplot matrices were generated using the *GGally* package (`v2.2.1`; Schloerke, D. Cook, Wickham, Crowley, Hofmann, Marbach, Anderson, Grolemund, and Wang 2023). The *MASS* (`v7.3-65`; Venables and B. D. Ripley 2002) and *car* (`v3.1-3`; Fox and Weisberg 2019) packages were used for applying Box-Cox transformations and for model selection via stepwise AIC. The *boot* package (`v1.3-31`; B. Ripley 2021) was used to perform bootstrap resampling, specifically for generating confidence intervals on residual Q-Q plots. Together, these tools enabled rigorous data analysis in line with the objectives of the study.

The Model 1 is the full regression model and is a baseline multiple linear regression model using the observed values of both the response variable and the predictor variables. Model 2 incorporates Box-Cox transformations for both the response and predictor variables to stabilize variance, improve normality in the response and to linearize the dependence between the response and the predictors. Model 2A is a refined version of Model 2, where the Akaike Information Criterion (AIC) is applied to Model 2 to select the most influential predictors while maintaining model simplicity. Model 2i is an extension of Model 2 that includes all the possible two-term interactions between the predictor variables, allowing more complex dependence structure between the response variables and the predictor variables. Model 2iB is a simplified version of Model 2i, where the Bayesian Information Criterion (BIC) is used to retain only the most significant interaction based predictors, balancing interpretability and predictive accuracy. The decision to use BIC for selecting model 2iB is based on its more substantial penalty for the number of parameters, making it a better criterion when dealing with many predictors and the risk of overfitting. AIC works well when the model complexity is not too high, so it was applied when choosing the model 2A.

To visualize the spatial relationship between the variability in corn yield and topography, the Figure 3.1 presents a topographic map, derived from the study by Colvin et al. (1997), and illustrates the variation in corn yield acc-

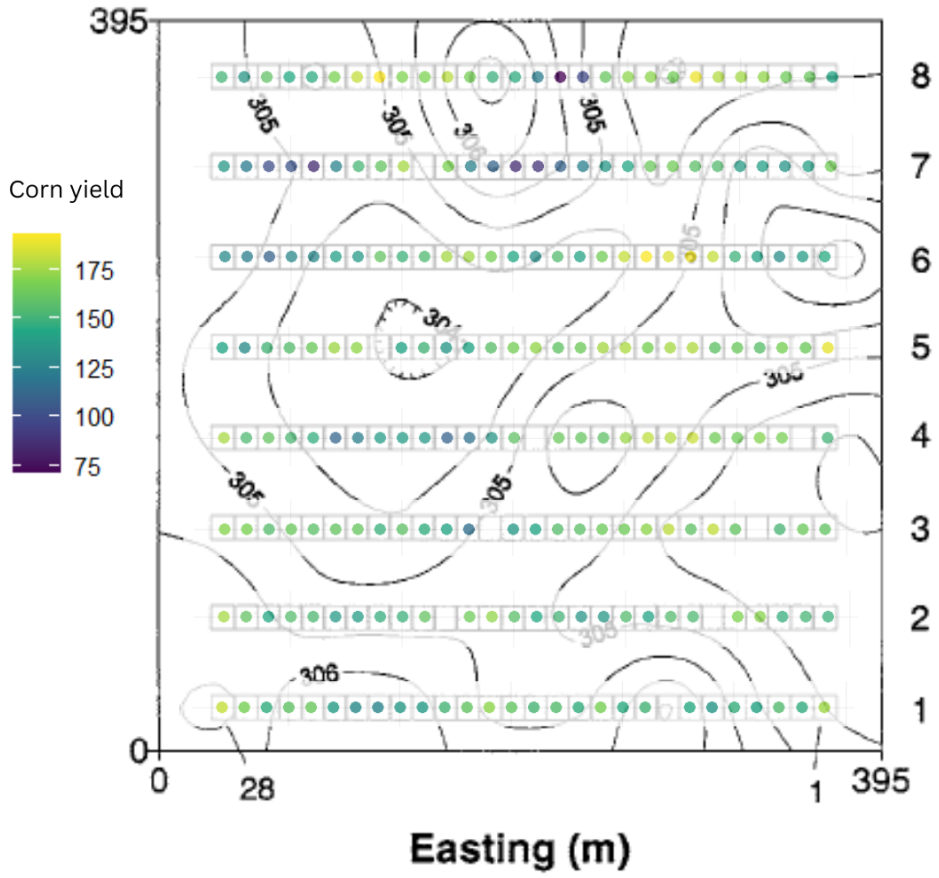

Figure 3.1: Illustration of corn yield and elevation across the study area. Topographic map is from the article by Colvin et al. (1997). The color gradient represents yield variation, with the color range with blue indicating lower yield and yellow indicating higher yield, while the contour lines indicate elevation changes. The numbers with the contour line are in meters.

ross the study area relative to the elevation. It highlights how field topography might influence yield variability, providing a clearer view of areas with higher and lower productivity. The color gradient in the Figure 3.1 is a visual aid to quickly identify regions of high and low yield. Blue shades represent areas with lower yield, while yellow indicates regions with higher yield. The overlaying contour lines show the field's elevation changes, providing insight into how topographical features may influence yield variability. The visual depiction of yield in relation to elevation suggests how field topography, such as slopes and depressions, might contribute to differences in productivity across the area. Elevation levels are known to influence water retention

and drainage patterns. Areas at lower elevations tend to retain more water, which could be beneficial for crop growth, particularly in regions with adequate rainfall. However, there is always the possibility that the lower areas suffer from excessive water retention, potentially leading to lower yields in certain areas. Conversely, higher-elevation areas tend to drain water more quickly. This could lead to drier soil conditions, potentially limiting crop productivity. Drier soil might hinder nutrient uptake and stress the plants, leading to reduced yields in these areas.
Soil properties, including nutrient availability, are often correlated with elevation. Areas with favorable soil conditions, such as better nutrient retention or higher organic matter content, might lead to higher corn yields. In contrast, regions with less fertile soils, often found at higher or lower elevations, might experience suboptimal growth. The variability in soil quality, influenced by elevation, might explain why certain areas show higher yields despite being at a slightly higher elevation. It suggests that local factors, like the availability of essential nutrients, can mitigate the negative effects of higher elevation. Furthermore, topographical features—such as slopes, depressions, and ridges create localized microclimates within the field. These microclimates can have a significant impact on corn yield by influencing water retention, nutrient availability, and overall plant growth. Although the inclusion of additional variables, such as rainfall data, more precise location variables, or further soil properties, could have potentially improved the model's performance, these data were not available for analysis.

3.2 Properties of the dependent variable

A Preliminary analysis on the dataset was performed to better understand the characteristics of the dataset and assess the distributions of the response variable. The histogram in Figure 3.2 and the Q-Q plot in Figure 3.3 reveal that significant left skewness, indicating non-normality in the distribution of the response variable. This suggests the need for transformation techniques to stabilize the variance and improve the normality of the response variable in subsequent modeling steps.
Figure 3.4 depicts the scatterplot of all the variables in the dataset. It allows evaluation of pairwise relationships and correlations between soil nutrients and corn yield. The matrix provides a comprehensive view of the relationships between corn yield and various soil nutrients, including B, Ca, Cu, Fe, K, Mg, Mn, Na, P, and Zn. The upper triangle of the matrix presents Pearson correlation coefficients (first developed by Pearson (1895)), while the lower triangle contains scatterplots illustrating pairwise relationships. The diagonal elements display univariate density plots, which help assess the distribution of each variable. This figure shows that corn yield exhibits strong positive correlations with several soil nutrients, particularly Zn ($r = 0.61$),

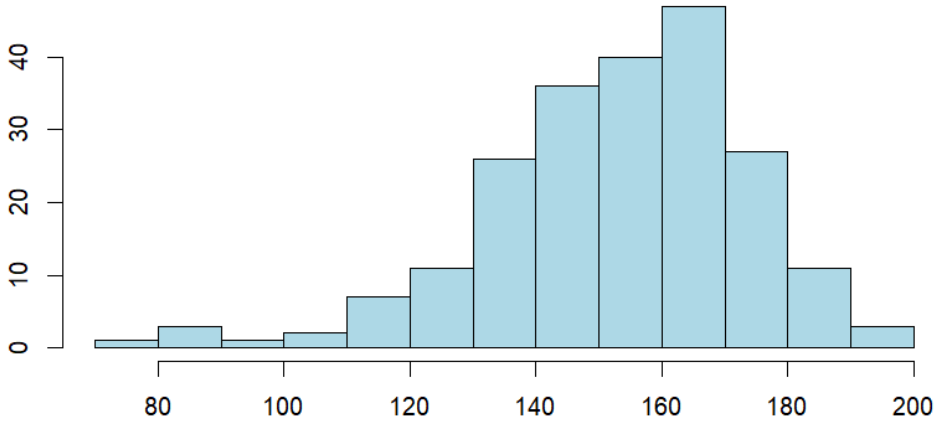

Figure 3.2: The distribution of corn yield in bushels per acre in 1997, measured across 215 locations in the 16-hectare Baker field.

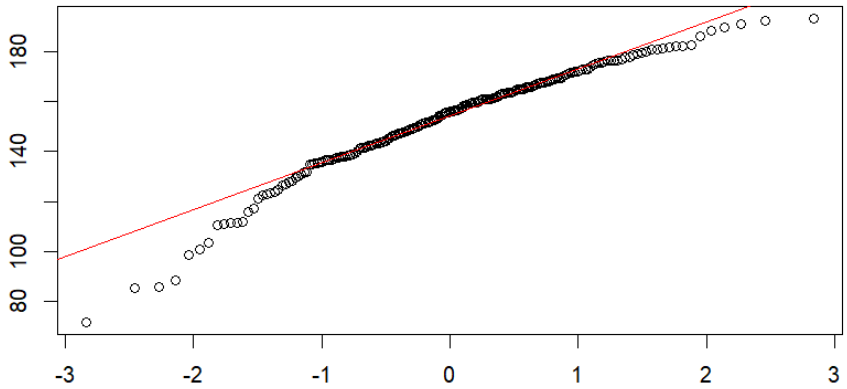

Figure 3.3: Q-Q plot of corn yield comparing the sample quantiles (vertical line) of corn yield data to theoretical quantiles (horizontal line) of a normal distribution. The red line represents the expected quantiles under the normal distribution and serves as a reference.

Cu ($r = 0.62$), and K ($r = 0.43$), suggesting that these elements play a significant role in influencing higher yields. Moderate correlations are observed with Mn ($r = 0.46$) and Mg ($r = 0.49$), while weaker correlations with Na ($r = 0.16$) and B ($r = 0.09$) indicate a more limited or no impact on yield. These observations highlight the importance of Zn, Cu, and K in influencing corn yield, while Mn and Mg show a somewhat weaker association. Additionally, Figure 3.4 shows that strong positive inter-nutrient correlations are evident, such as the relationship between Mg and Cu ($r = 0.70$), as well as Mg and Zn ($r = 0.68$), likely reflecting soil chemistry interactions. Notable relationships are also observed between Cu and Zn ($r = 0.77$) and K and P

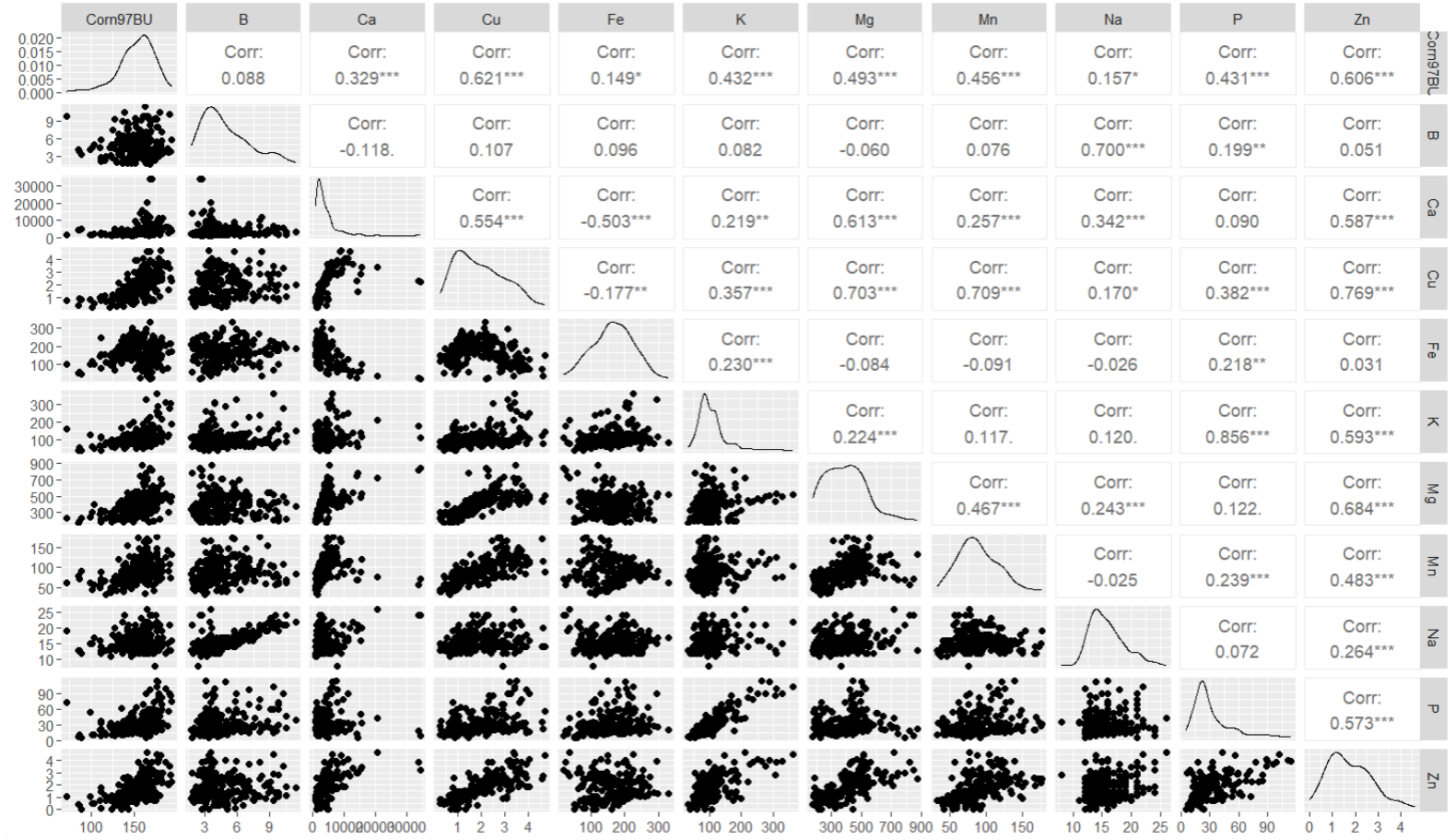

Figure 3.4: Scatterplot matrix for the soil nutrient variables and corn yield. The matrix includes pairwise scatterplots, density plots, and correlation coefficients.

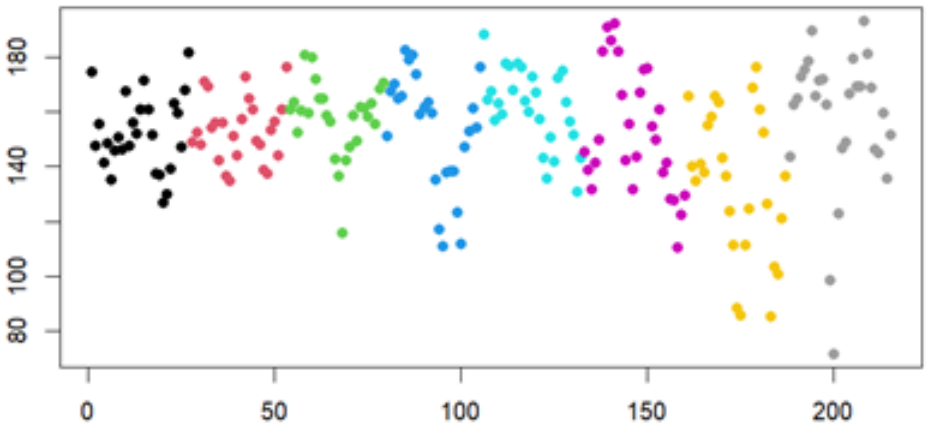

Figure 3.5: Corn yield plotted against index values for all locations in the Baker Field dataset. Each color represents a transect as black for transect 1, red for transect 2, green for transect 3, blue for transect 4, cyan for transect 5, magenta for transect 6, yellow for transect 7, and gray for transect 8.

($r = 0.86$), suggesting possible co-dependencies between these nutrients. Conversely, some nutrient pairs, such as Na and Mn ($r = -0.03$), exhibit weak or negligible correlations, indicating minimal direct association between these variables. Such high correlations among predictor variables often raise concerns about multicollinearity, which can lead to inflated standard errors and reduce the precision of coefficient estimates in regression models. However, as discussed by Wooldridge (2009), multicollinearity does not bias ordinary least squares (OLS) estimators, meaning the estimated coefficients remain valid. It is noted that multicollinearity only increases the variance of the estimates, which can reduce statistical significance but does not invalidate the model. Since the primary objective of this analysis is to estimate conditional mean effects rather than interpret individual coefficients precisely, multicollinearity is not a major concern. Furthermore, as long as the regression includes a sufficiently large sample size and all the relevant explanatory variables, multicollinearity does not pose a severe issue in practice. Figure 3.5 highlights spatial variation in corn yield across the field. While some transects exhibit higher yield concentrations, others display more dispersed or lower values. Notably, Transect 8 (shown in gray) stands out with more scattered and generally lower yields, suggesting potential spatial heterogeneity in soil conditions or other environmental factors.

3.3 Statistical Data Analysis

To further investigate the relationships between soil nutrients and corn yield, a series of statistical modeling techniques were employed. Multiple linear

regression models were constructed, with the goal of identifying key predictors influencing the yield as described in the introduction 3.1. We look at each model individually to examine their specific characteristics, performance metrics, and the significance of their predictor variables. Furthermore, we aim to search for the best possible explanation on the variation in the corn yield using the explanatory variables.

3.3.1 Model 1

Model 1 was formulated to investigate the direct relationship between corn yield and the 10 soil nutrient variables. While Colvin et al. (1997) focused on spatiotemporal yield variability using descriptive statistics and geospatial techniques, they did not develop regression models linking soil nutrients to yield. However, their findings on soil drainage, topography, and nutrient effects directly motivate our regression analysis. Model 1 extends their work by quantitatively assessing nutrient-yield relationships, revealing an R^2 of 0.46, indicating that the explanatory variables explain approximately 46% of the dependent variable. Model 1 is

$$\begin{aligned} y_i = \beta_0 + \beta_1 \cdot \mathrm{Zn}_i + \beta_2 \cdot \mathrm{K}_i + \beta_3 \cdot \mathrm{Mn}_i + \beta_4 \cdot \mathrm{P}_i + \beta_5 \cdot \mathrm{Na}_i + \beta_6 \cdot \mathrm{Mg}_i \\ + \beta_7 \cdot \mathrm{B}_i + \beta_8 \cdot \mathrm{Ca}_i + \beta_9 \cdot \mathrm{Fe}_i + \beta_{10} \cdot \mathrm{Cu}_i + \varepsilon_i, \end{aligned} \tag{3.1}$$

where y_i represents the raw corn yield data for observations i, where $i = 1, \ldots, 215$. The summary of the estimated Model 1 is given in the Appendix A. To evaluate the normality assumption for Model 1, a Q-Q plot with 99% bootstrap confidence intervals was used (Figure 3.6). This indicates that the standardized residuals generally follow a normal distribution, as all points fall within the expected range of variation. While this suggests no significant departures from normality at the chosen confidence level, there are slight deviations at the tails, particularly for transects 7 and 8. A 99% confidence level is more appropriate here because it provides a stricter threshold for detecting deviations from normality, reducing the risk of false positives. Since Q-Q plots are sensitive to minor variations, especially in the tails, a higher confidence level ensures that only substantial departures from normality are considered statistically meaningful.

The scale-location plot (Figure 3.7) for Model 1 residuals helps assess the assumption of constant variance (homoscedasticity). This assumption means that the variance of the residuals should remain roughly constant across all levels of fitted values. The increase in scatter would indicate heteroscedasticity, which is a condition where the variability of the residuals changes across levels of the predictor or fitted values. The loess (locally estimated scatterplot smoothing) curve developed by Cleveland (1979) is a nonparametric method for smoothing a series of data in which no assumptions are made about the underlying structure of the data. As discussed by Sharma, Swayne, and Obimbo (2015), loess and lowess (locally weighted scatterplot

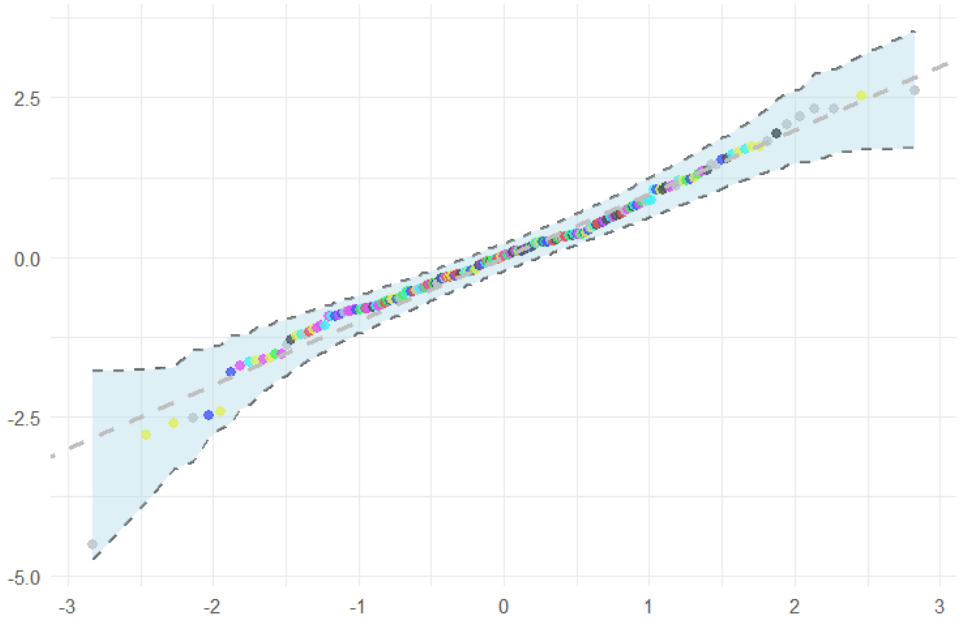

Figure 3.6: Q–Q plot of standardized residuals from Model 1 with 99% bootstrap confidence intervals. The x-axis shows the theoretical quantiles from the standard normal distribution, and the y-axis shows the standardized residuals. The 45 degree line represents the theoretical quantiles under a standard normal distribution, while shaded region indicates the 99% bootstrap confidence envelope. Each color represents a transect as black for transect 1, red for transect 2, green for transect 3, blue for transect 4, cyan for transect 5, magenta for transect 6, yellow for transect 7, and gray for transect 8.

smoothing developed by Cleveland (1979)) methods perform localized polynomial regression (typically linear in lowess and quadratic in loess) on the data. The flexibility of these approaches allows them to capture nonlinear trends without requiring a model. This helps to reflect the constant variance of the theoretical errors and adds information in the graph to help interpretation. In Figure 3.7 the slightly downward-sloping loess curve suggests a minor reduction in residual spread as fitted values increase, indicating a potential departure from constant variance.
Figure 3.8 is used to identify influential observations that may have an impact on the regression model. The red line represents a loess smoothing curve that helps visualize the general trend in residuals across different leverage values. Notably, two observations from the transect 4 in the top-right corner of the plot exhibit high leverage but low standardized residuals. These points are not outliers in terms of the response variable, but their unusual predictor values give them significant influence on the model. Their proxim-

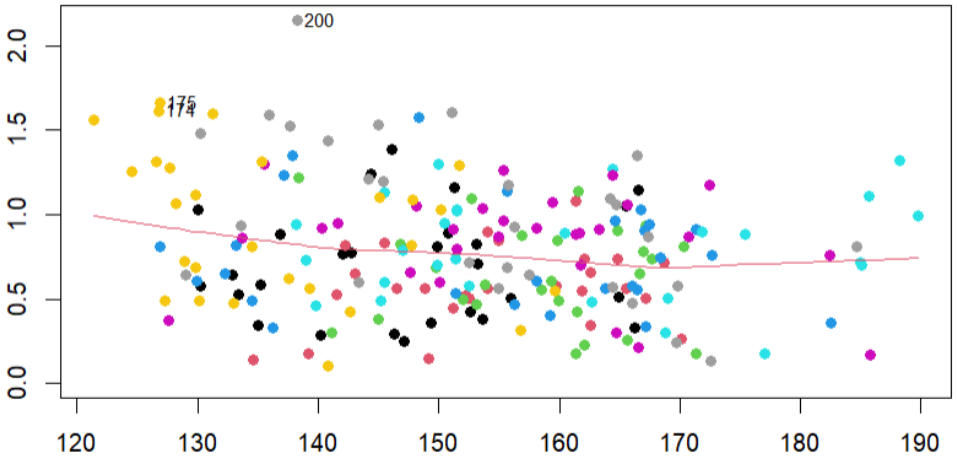

Figure 3.7: Scale-location plot for Model 1 standardized residuals. The horizontal axis represents the fitted values, and the vertical axis shows the square root of the standardized residuals. The red line is a loess smoothing curve developed by Cleveland (1979). Each color represents a transect as black for transect 1, red for transect 2, green for transect 3, blue for transect 4, cyan for transect 5, magenta for transect 6, yellow for transect 7, and gray for transect 8.

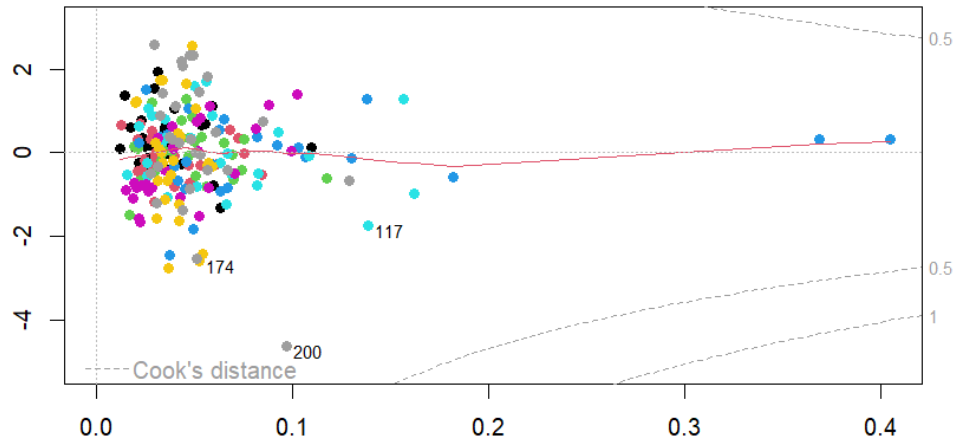

Figure 3.8: Standardized residuals (vertical axis) and leverage (horizontal axis) plot for Model 1 residuals. The red line is a loess smoothing curve. The dashed lines represent Cook's distance contours developed by R. D. Cook (1977). Each color represents a transect as black for transect 1, red for transect 2, green for transect 3, blue for transect 4, cyan for transect 5, magenta for transect 6, yellow for transect 7, and gray for transect 8.

ity to or location beyond the Cook's distance (introduced by R. D. Cook (1977)) thresholds reinforces their potential as influential observations. The

loess curve remains relatively flat, suggesting that residuals do not show a strong systematic pattern with increasing leverage.

3.3.2 Model 2

Since the corn yield is based on non-normal, to improve the assumptions of normality and linearity, Box-Cox transformations were applied to the dependent variable and each of the explanatory variables. For strictly positive values y_i, the Box-Cox transformation is applied as

$$z_i = c(y_i) = \begin{cases} \frac{y_i^\lambda - 1}{\lambda} & \text{if } \lambda \neq 0, \\ \log(y_i) & \text{if } \lambda = 0, \end{cases}$$

where y_i is the i-th observation of the original variable, z_i is the corresponding transformed observation and λ is the transformation parameter, estimated to optimize normality and linearity. The optimal λ values were determined using the maximum likelihood method proposed by Box et al. (1964) and extended by Velilla (1993), which simultaneously estimates transformations for both response and predictor variables.
The Table 3.1 presents the estimated and rounded Box-Cox transformation parameters for each variable, with their corresponding Wald 95% confidence intervals (developed by Wald (1943)). These estimates aim to improve the normality of the response variable and enhance the linearity between the explanatory and dependent variables. For the response variable (corn yield), the estimated transformation parameter was approximately $\hat{\lambda} \approx 3$, indicating a substantial deviation from normality in the raw data. Among the explanatory variables, most chemical elements (K, Mn, P, Na, Mg, Ca) required logarithmic transformations ($\hat{\lambda} \approx 0$), while others (Zn, B, Fe, Cu) were transformed with square root transformations ($\hat{\lambda} \approx 0.5$). While these transformation choices were informed by estimated Box–Cox parameters, alternative transformations could have been explored. However, such extensions were not pursued in this analysis. After applying the Box-Cox transformations and improving the model with transformed variables, the updated regression model explained approximately 50% of the variability in the transformed dependent variable, as indicated by the R^2 value. The summary of the estimated Model 2 is given in the Appendix A. The formulation for Model 2 is

$$\begin{aligned} z_i = {} & \beta_0 + \beta_1 \cdot \sqrt{\mathrm{Zn}_i} + \beta_2 \cdot \log(\mathrm{K}_i) + \beta_3 \cdot \log(\mathrm{Mn}_i) + \beta_4 \cdot \log(\mathrm{P}_i) \\ & + \beta_5 \cdot \log(\mathrm{Na}_i) + \beta_6 \cdot \log(\mathrm{Mg}_i) + \beta_7 \cdot \sqrt{\mathrm{B}_i} + \beta_8 \cdot \log(\mathrm{Ca}_i) \\ & + \beta_9 \cdot \sqrt{\mathrm{Fe}_i} + \beta_{10} \cdot \sqrt{\mathrm{Cu}_i} + \varepsilon_i, \quad i = 1, \ldots, 215, \end{aligned} \tag{3.2}$$

where z_i represents the corn yield after applying the Box-Cox transformation $c(y_i)$ for the i^{th} observation y_i.

Table 3.1: Estimated Box-Cox power transformations for the response and predictor variables, including rounded powers and their corresponding 95% Wald confidence intervals.

Variable	**Estimated Power**	**Rounded Power**	**Lower bound**	**Upper bound**
Corn yield	2.89	3	2.18	3.60
Zn	0.73	0.5	0.48	0.99
K	-0.24	0	-0.49	0.01
Mn	0.23	0	-0.07	0.53
P	-0.03	0	-0.19	0.12
Na	0.12	0	-0.34	0.58
Mg	-0.04	0	-0.31	0.22
B	0.25	0.5	0.00	0.51
Ca	-0.13	0	-0.26	0.01
Fe	0.39	0.5	0.17	0.61
Cu	0.43	0.5	0.30	0.56

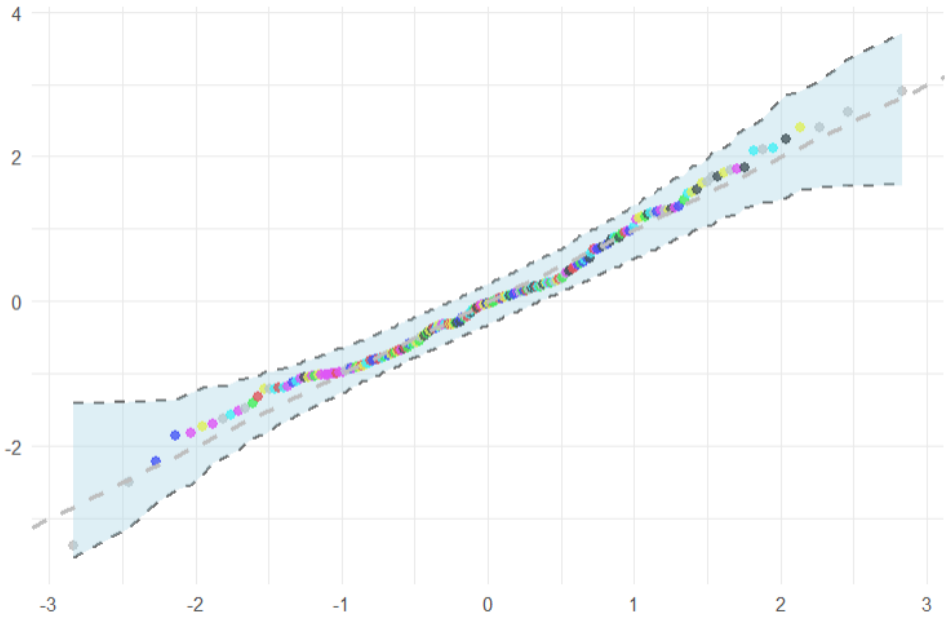

Figure 3.9: Q–Q plot of standardized residuals from Model 2 with 99% bootstrap confidence intervals. The x-axis shows the theoretical quantiles from the standard normal distribution, and the y-axis shows the standardized residuals. The 45 degree line represents the theoretical quantiles under a standard normal distribution, while shaded region indicates the 99% bootstrap confidence envelope. Each color represents a transect as black for transect 1, red for transect 2, green for transect 3, blue for transect 4, cyan for transect 5, magenta for transect 6, yellow for transect 7, and gray for transect 8.

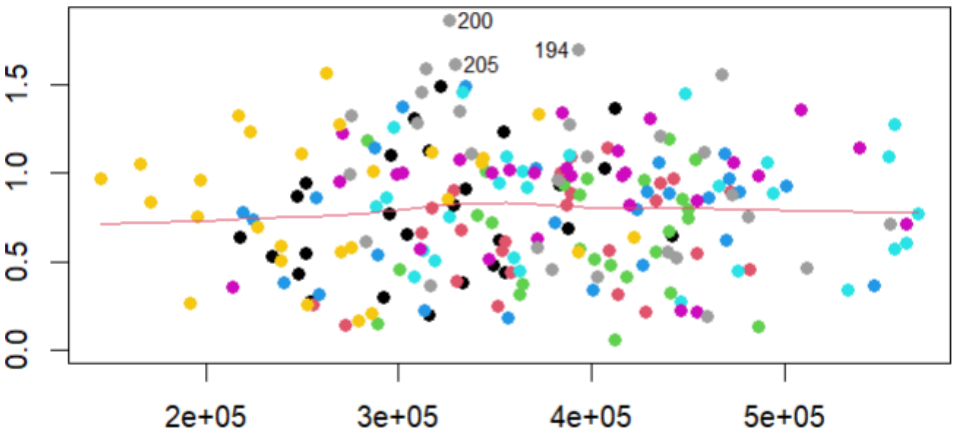

Figure 3.10: Scale-location plot for Model 2. The horizontal axis represents the fitted values, and the vertical axis shows the square root of the standardized residuals. The red line is a loess smoothing curve. Each color represents a transect as black for transect 1, red for transect 2, green for transect 3, blue for transect 4, cyan for transect 5, magenta for transect 6, yellow for transect 7, and gray for transect 8.

Moving on to the residual diagnostics, Figure 3.9 for Model 2 shows that the standardized residuals closely follow the theoretical quantiles of the standard normal distribution, with nearly all points lying along the dashed reference line and within the 99% bootstrap confidence envelope. This indicates that the normality assumption is well met for Model 2, with minimal deviation in the tails and no clear outliers across transects. Compared to Model 1, Model 2 appears to exhibit slightly improved normality of residuals.
The scale-location plot (Figure 3.10) for Model 2 helps assess the assumption of constant variance (homoscedasticity). The fairly uniform spread of standardized residuals across fitted values indicates that the variance remains consistent, supporting the assumption. Additionally, the loess smoothing curve remains relatively flat, showing no clear trend, which suggests that the Box-Cox transformation has effectively stabilized variance and improved the model's fit. Figure 3.11 shows that, compared to Model 1, Model 2 handles outliers better and is less affected by them, as seen from the lower Cook's distance values. The Cook's distance values in Model 2 are significantly lower (ranging from 0 to 0.15) compared to the much higher values in Model 1. This indicates that Model 2 is less sensitive to individual data points. The reduction in Cook's distance suggests that Model 2 has a better fit to the data, likely due to improved model specification. Therefore, Model 2 appears to be a more reliable and stable regression model compared to Model 1.

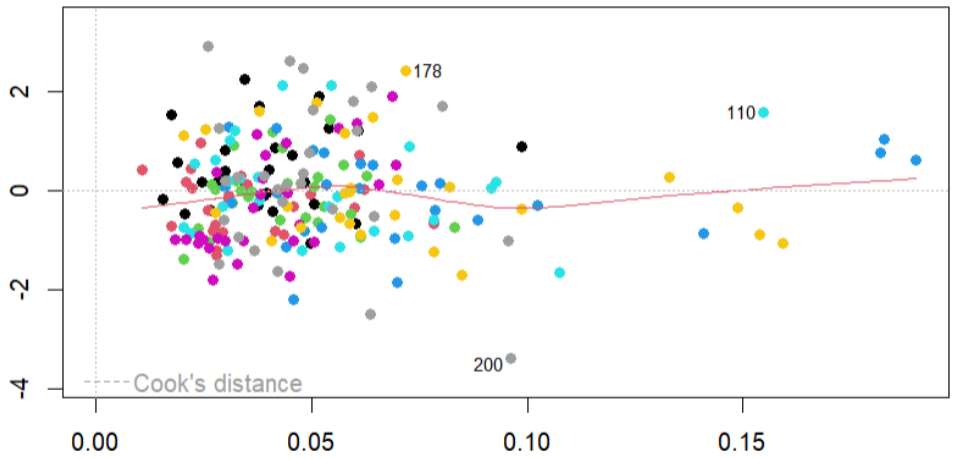

Figure 3.11: Standardized residuals (vertical axis) and leverage (horizontal axis) plot for Model 2 residuals. The red line is a loess smoothing curve. Each color represents a transect as black for transect 1, red for transect 2, green for transect 3, blue for transect 4, cyan for transect 5, magenta for transect 6, yellow for transect 7, and gray for transect 8.

3.3.3 Model 2A

Next regression model was selected after applying the Akaike Information Criterion (AIC) to Model 2, to achieve a balance between goodness-of-fit and model complexity. The formulation for Model 2A is

$$\begin{aligned} z_i = \beta_0 + \beta_1 \cdot \log(\mathrm{K}_i) + \beta_2 \cdot \log(\mathrm{P}_i) + \beta_3 \cdot \sqrt{\mathrm{Zn}_i} \\ + \beta_4 \cdot \sqrt{\mathrm{Cu}_i} + \varepsilon_i, \quad i = 1, \ldots, 215. \end{aligned} \tag{3.3}$$

The selected predictors K, P, Zn and Cu were consistent with the findings from the preliminary and statistical analysis, which highlighted their moderate relationships with corn yield. The R^2 value of the Model 2A was 0.51, slightly higher than that value of Model 2, but the model demonstrated improved stability. The summary of the estimated Model 2A is given in the Appendix A. The residual diagnostic plots of Model 2A closely resemble those of Model 2, exhibiting minimal heteroscedasticity and improved adherence to normality assumptions compared to Model 1. This suggests that the AIC-selected Box-Cox transformed model effectively stabilizes variance and enhances the overall model fit.
The Table 3.2 presents the regression coefficients for Model 2A. This model includes a subset of transformed predictor variables chosen to balance explanatory power and model simplicity. The high F-statistic ($F_{4,210} = 57$, $p < 2.2 \times 10^{-16}$) provides strong evidence against the null hypothesis that all regression coefficients are simultaneously zero. This confirms that the model, as a whole, significantly explains variation in the transformed corn

Table 3.2: Estimated regression coefficients, and p-values for the predictors in Model 2A.

Variable	**Estimate**	p**-value**
$\sqrt{\text{Zn}}$	39622	0.12
log(K)	47372	0.06
log(P)	30105	0.03
$\sqrt{\text{Cu}}$	142108	0.00

yield. While the table also reports individual p-values, caution is warranted in interpreting them. These values are based on t-tests that do not account for potential correlations among the estimated predictors. As such, a high p-value for a particular variable does not necessarily imply that the variable lacks relevance. In fact, some variables with less significant individual tests may still contribute substantially to the model jointly with other predictors. For instance, log(P) and $\sqrt{\text{Cu}}$ appear individually significant at the 5% level, but this should not lead to the exclusion of log(K) or $\sqrt{\text{Zn}}$, which have higher p-values. These predictors may still play an important role in explaining the variation of the corn yield when considered jointly with the others.

3.3.4 Model 2i

Model 2i represents an expanded version of Model 2, where two-term interactions were added to better capture the relationships between the predictors and the transformed response variable. The formulation for Model 2i is

$$\begin{aligned} z_i = {} & \beta_0 + \beta_1 \cdot \sqrt{\text{Zn}_i} + \beta_2 \cdot \log(\text{K}_i) + \beta_3 \cdot \log(\text{Mn}_i) \\ & + \beta_4 \cdot \log(\text{P}_i) + \beta_5 \cdot \log(\text{Na}_i) + \beta_6 \cdot \log(\text{Mg}_i) + \beta_7 \cdot \sqrt{\text{B}_i} \\ & + \beta_8 \cdot \log(\text{Ca}_i) + \beta_9 \cdot \sqrt{\text{Fe}_i} + \beta_{10} \cdot \sqrt{\text{Cu}_i} \\ & + \sum_{j<k} \beta_{jk} \cdot (\text{Predictor}_{j,i} \cdot \text{Predictor}_{k,i}) + \varepsilon_i, \quad i = 1, \ldots, 215. \end{aligned} \tag{3.4}$$

This model includes all main effects and second-order interactions between pairs of predictors. The R^2 value for Model 2i was found to be 0.53, indicating a moderate proportion of variance in the transformed yield is explained by the model. However, the model's complexity with numerous interaction terms relative to the sample size raises concerns about estimation efficiency. The limited dataset size may lead to inefficient parameter estimates, making the results potentially unreliable for interpretation. The summary of the estimated Model 2i is given in the Appendix A. In Figure 3.12, the loess smoothing curve suggests that the key assumptions of linear regression are met. There is no significant deviation from the zero line which indicates linearity in the relationship between the dependent and explanatory variables.

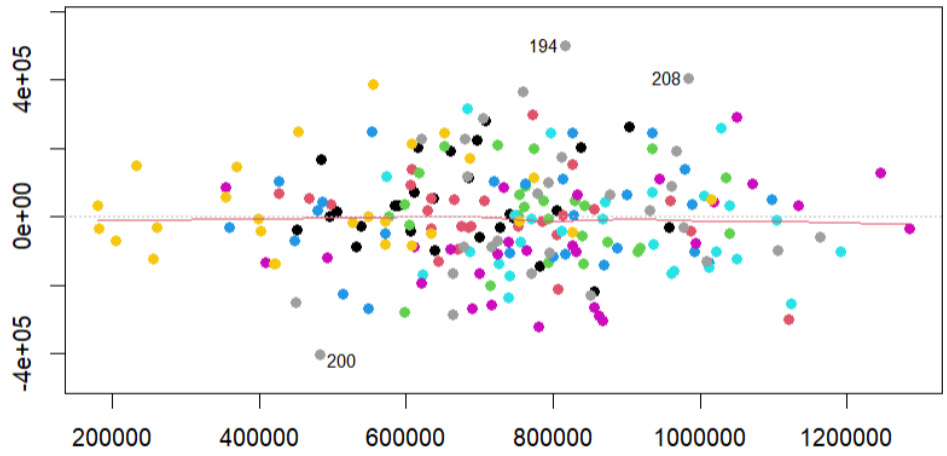

Figure 3.12: Residuals (vertical) and fitted values (horizontal) plot for Model 2i. The red line shows a loess smooth curve fitted to the residuals. Each color represents a transect as black for transect 1, red for transect 2, green for transect 3, blue for transect 4, cyan for transect 5, magenta for transect 6, yellow for transect 7, and gray for transect 8.

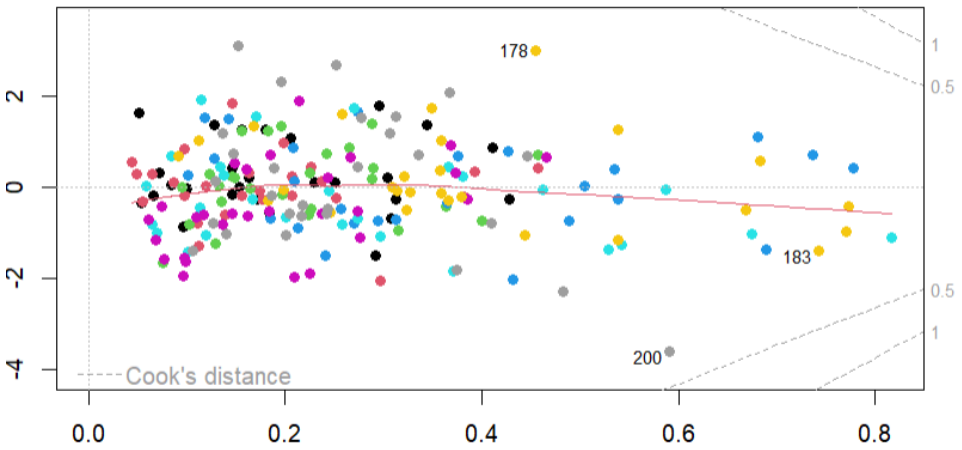

Figure 3.13: Standardized residuals (vertical axis) and leverage (horizontal axis) plot for Model 2i residuals. The red line is a loess smoothing curve. Each color represents a transect as black for transect 1, red for transect 2, green for transect 3, blue for transect 4, cyan for transect 5, magenta for transect 6, yellow for transect 7, and gray for transect 8.

The relatively even spread of residuals around the zero line suggests homoscedasticity, meaning the variance of the residuals appears constant across the range of fitted values. Since the Q-Q plot for Model 2i residuals shows the same pattern as the Q-Q plot for Model 2 (Figure 3.9), it suggests that both models have similar residual distributions. This implies that the resid-

uals for both models are approximately normally distributed.
The standardized residuals and leverage plot (Figure 3.13) for Model 2i shows higher Cook's distance values (up to 0.8) compared to the previous models. This indicates that the inclusion of interaction terms has increased the model's sensitivity to influential observations.

3.3.5 Model 2iB

Model 2iB was selected based on the Bayesian Information Criterion (BIC), which aims to balance model fit and complexity. This model was obtained by reducing the number of predictors and the interaction terms from Model 2i and the formulation for Model 2iB is

$$\begin{aligned} z_i = \beta_0 + \beta_1 \cdot \log(\mathrm{P}_i) + \beta_2 \cdot \log(\mathrm{Na}_i) + \beta_3 \cdot \sqrt{\mathrm{Cu}_i} \\ + \beta_4 \cdot \log(\mathrm{P}_i) \cdot \log(\mathrm{Na}_i) + \varepsilon_i, \quad i = 1, \ldots, 215. \end{aligned} \tag{3.5}$$

The R^2 value for Model 2iB was 0.52, which is slightly lower than that of Model 2i. While this suggests a small reduction in explanatory power, the model is notably simpler, with only three main terms and one interaction term, making it more interpretable. The summary of the estimated Model 2iB is given in the Appendix A. Model 2i is relatively complex due to the inclusion of many two-term interactions between the predictors, which increases the number of parameters to be estimated. This added complexity can lead to overfitting, where the model captures not only the underlying relationship between predictors and response, but also random variation or noise specific to the sample data. In such cases, the model may perform well on the training data (while showing a high R^2), but generalize poorly to new data because the estimated effects of certain interactions may not hold outside the sample. The large number of interaction terms in Model 2i increases the risk of overfitting, as some of these interactions may reflect invalid associations rather than genuine, meaningful effects. Therefore, despite a slightly lower R^2, the more parsimonious structure of Model 2iB may provide a more reliable explanation of yield variation across the field.
The residuals and fitted values plot for Model 2iB, being similar to that for Model 2i (Figure 3.12) with a flat loess smoothing line, suggest that both models exhibit similar characteristics in terms of their residuals. The Q-Q plot of standardized residuals for Model 2iB closely resembles that of Model 2 (Figure 3.9), with nearly all points aligning well with the reference line. This agreement suggests that the residuals follow a normal distribution, supporting the model's assumption of normality. The standardized residuals and leverage plot (Figure 3.14) for Model 2iB shows leverage values ranging only from 0 to 0.15, indicating that no observations have extremely high leverage. This is a positive sign, as it suggests that no single observation has predictor values far from the mean that could disproportionately influence the

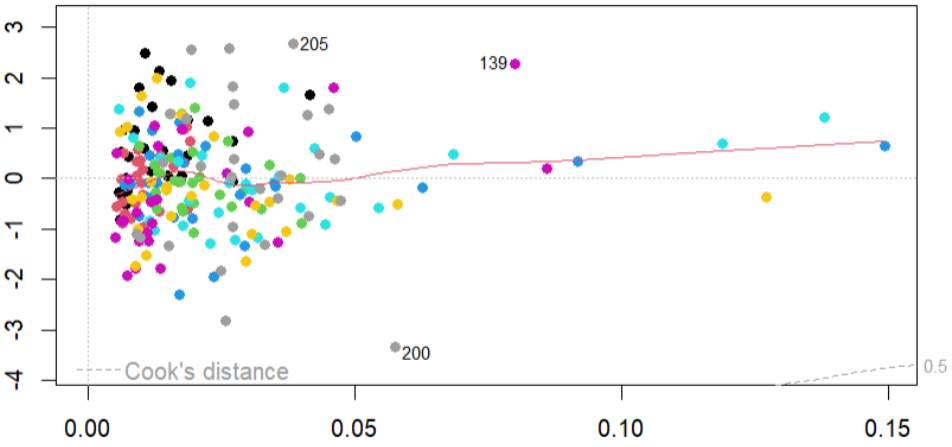

Figure 3.14: Standardized residuals (vertical axis) and leverage (horizontal axis) plot for Model 2iB. The red line is a loess smoothing curve. Each color represents a transect as black for transect 1, red for transect 2, green for transect 3, blue for transect 4, cyan for transect 5, magenta for transect 6, yellow for transect 7, and gray for transect 8.

Table 3.3: Estimated regression coefficients and p-values for the Model 2iB.

Predictor	**Estimate**	p**-value**
log(P)	531878	0.00
log(Na)	569808	0.00
$\sqrt{\text{Cu}}$	167617	0.00
$\log(\text{P}) \cdot \log(\text{Na})$	-169860	0.00

model. The smoother loess curve indicates a slight upward trend in residuals at higher leverage values. However, this trend should be interpreted cautiously. The right tail of the plot contains relatively few observations, and the loess curve in this region is thus less reliable since loess is a non-parametric smoothing method that can be sensitive to data sparsity. The majority of the data points lie on the left side of the plot, where the loess curve more accurately reflects the local structure of the residuals. Therefore, interpretation should focus primarily on this densely populated region, where the curve suggests no clear pattern, supporting assumptions of linearity and homoscedasticity in this domain. In contrast, Model 2i includes a large number of interaction terms, resulting in many parameters to estimate. This increases the likelihood that some estimated coefficients are inaccurate and inefficient, especially if the true values of those coefficients are close to zero. Such estimation can reduce model reliability. The simpler structure of Model 2iB avoids this issue by focusing on a smaller set of well-supported terms. The Table 3.3 presents the regression results for Model 2iB. This model yields

a highly significant F-statistic ($F_{4,210} = 58.41$, $p < 2.2 \times 10^{-16}$), confirming that the combination of these predictors offers substantial explanatory power for the transformed corn yield. It can be noted that all included predictors have relatively low p-values, suggesting they contribute meaningfully to explaining the variation in transformed corn yield. The variables $\sqrt{\text{Cu}}$, log(P), and log(Na) show positive associations with yield, implying that increases in P, Cu and Na are generally associated with higher productivity. However, the interaction term $\log(\text{P}) \cdot \log(\text{Na})$ is negative, indicating that the joint influence of P and Na may be less beneficial than their individual contributions. This could reflect a biological interaction where high Na levels diminish the effectiveness of P uptake. However rather than interpreting the significance of individual predictors in isolation, emphasis should be placed on the joint explanatory value of the model, consistent with the considerations discussed under Model 2A.

Model 2A and Model 2iB represent two parsimonious approaches to modeling transformed corn yield, selected using different information criteria, AIC and BIC, respectively. While both models aim to balance predictive accuracy with model simplicity, they differ in their selected predictors. Model 2A includes log(K), log(P), $\sqrt{\text{Zn}}$, and $\sqrt{\text{Cu}}$, and Model 2iB includes only three main effects log(P), log(Na), and $\sqrt{\text{Cu}}$ as well as their interaction term $\log(\text{P}) \cdot \log(\text{Na})$. Both models identify log(P) and $\sqrt{\text{Cu}}$ as significant predictors, suggesting a consensus that P and Cu play a crucial role in explaining the variation in corn yield across the field. However, the remaining predictors differ. Model 2A includes log(K) and $\sqrt{\text{Zn}}$, nutrients that emerged from earlier preliminary analysis but are not statistically significant at conventional levels in the final model. This inclusion reflects AIC's tendency to accept slightly weaker variables for better overall model fit. On the other hand, Model 2iB includes log(Na) and its interaction with log(P), both highly significant. This highlights how interaction effects can reveal conditional relationships that might not be captured by main effects alone. The interaction term, in particular, implies that the positive effect of phosphorus on yield diminishes when sodium levels are high, an insight not detectable in Model 2A's additive formulation. Model 2A, although simpler than its parent model, Model 2, still includes more predictors and transformations than Model 2iB. Meanwhile, Model 2iB, with only four terms (including one interaction), achieves a slightly higher R^2, suggesting a more efficient use of information. The use of BIC in Model 2iB is important when dealing with many potential interaction terms, as seen in its parent model (Model 2i). Therefore, while both models agree on the importance of P and Cu, they diverge in the inclusion of other variables, with Model 2A favoring K and Zn and Model 2iB emphasizing Na and its interaction with P. Together, the two models offer complementary perspectives on the soil nutrient factors most relevant to corn yield prediction. A comparison of BIC values as presented in the Table 3.4 reinforces the strength of the BIC-selected Model 2iB. Among

Table 3.4: BIC values for candidate models

Model	**BIC**
Model 2	5547.04
Model 2i	5720.57
Model 2A	5517.51
Model 2iB	5514.77

the models considered, Model 2iB achieves the lowest BIC, narrowly outperforming Model 2A. Both perform substantially better than their more complex parent models, Model 2 and Model 2i which include more predictors and interactions. These results suggest that reducing model complexity through principled variable selection, especially when accounting for interaction effects, can lead to more efficient models without sacrificing explanatory power. Thus, while both simplified models offer valuable insights, Model 2iB provides the most parsimonious and statistically supported explanation of yield variation.

3.4 Spatial Distribution and Yield Variability

To investigate spatial patterns in corn yield across the 215 sampled locations, Moran's I test was employed to assess spatial autocorrelation. The resulting Moran's I statistic was -0.004, indicating a negligible negative spatial autocorrelation. This suggests that adjacent locations do not exhibit systematically similar or dissimilar yield values. Furthermore, the associated p-value of 0.5 is considerably higher than the conventional significance threshold of 0.05. As a result, we fail to reject the null hypothesis of spatial independence. These findings support the conclusion that there is no statistically significant evidence of spatial clustering or dispersion in corn yield across the field. Therefore, it is reasonable to proceed under the assumption that spatial autocorrelation does not substantially influence the distribution of yield in this dataset. Beyond this general finding, specific anomalies in the spatial distribution were noted through residual diagnostics. Notably, two prominent deviations were identified within the 8th transect at locations 7 and 13 where the observed yields were 189.5 and 71.6 bushels per acre, respectively. These values represent substantial deviations from model based expectations and may indicate the presence of localized environmental or agronomic factors not captured by the measured variables. Their inclusion underscores the importance of site-specific investigation and may motivate future targeted sampling or qualitative field assessment.

An examination of terrain variation across the transects, particularly using contour plots, reveals that transects 3 and 5 exhibit relatively uniform spacing between contour lines. This uniformity suggests flatter topogra-

phy, which likely contributes to more consistent growing conditions and thus more stable yield outcomes. In contrast, transects 1 and 8 show irregular or closely spaced contours, indicating steeper or more uneven terrain. Such topographical variability may expose these areas to microclimatic extremes such as enhanced water runoff or wind exposure resulting in more variable crop performance. It is also important to acknowledge the possible influence of unmeasured environmental factors. Variables such as rainfall measurements, soil compaction, moisture content, and additional soil chemical or biological properties were not available in the current dataset but could play a significant role in explaining yield variability. Incorporating such data in future studies would likely enhance the explanatory power of the spatial models and provide a more complete understanding of the agronomic system.

Chapter 4

Conclusion

The thesis explored the application of linear regression models with Box-Cox transformations to investigate the relationship between soil nutrient concentrations and corn yield across a 16-hectare agricultural field. The dataset included measurements from 215 locations on 10 different soil nutrients and corn yield. Given the presence of skewed distributions and nonlinear relationships, Box-Cox transformations were applied to both the response and explanatory variables to stabilize variance and improve model fit. A series of regression models were developed to capture the relationship between the transformed yield and various nutrient predictors. Initial preliminary analysis and diagnostic tools supported the evaluation of model fit and the selection of variables. The modeling process began with simpler models and gradually incorporated additional complexity through interaction terms, culminating in Model 2i, which included all main effects and second-order interactions. While Model 2i achieved the highest R^2 value, indicating moderate explanatory power, it also introduced a large number of parameters, increasing the risk of overfitting and reducing model interpretability.

To address concerns about model complexity, model selection criteria were employed. Model 2A was selected based on the Akaike Information Criterion (AIC), and Model 2iB was selected based on the Bayesian Information Criterion (BIC). These models represent two parsimonious yet effective approaches to modeling yield variation. Model 2A retained four predictors, log(K), log(P), $\sqrt{\text{Zn}}$, and $\sqrt{\text{Cu}}$, while Model 2iB included three main effects such as log(P), log(Na), and $\sqrt{\text{Cu}}$ along with their interaction term $\log(\text{P}) \cdot \log(\text{Na})$. Despite their simpler structures, both Model 2A and Model 2iB maintained good model performance with R^2 values of 0.51 and 0.52, respectively. Moreover, residual diagnostics for these models indicated that key assumptions of linear regression were reasonably met, including linearity, homoscedasticity, and normality of residuals. Importantly, these parsimonious models offered better interpretability and reduced sensitivity to influential observations compared to more complex alternatives.

The analysis was based on data collected from a one field during a single growing season, which may restrict the generalizability of the results to different locations or time periods. Even though the spatial autocorrelation was not detected, the sampling resolution may have been too coarse to reveal subtle spatial dependencies that could influence yield. Additionally, although Box-Cox transformations helped improve model fit, they are limited to monotonic transformations and may not fully capture complex nonlinear patterns in the data.

Future research could expand upon this work in several directions. Studies with finer spatial resolution or multiple seasons could provide deeper insight into spatial or temporal dynamics of soil-crop relationships. Nonlinear modeling approaches, such as generalized additive models, may offer greater flexibility in capturing complex relationships and interactions among variables. Additionally, exploring alternative transformations of the explanatory variables could further improve model fit or uncover underlying nonlinear effects. Incorporating additional environmental covariates such as rainfall, temperature, or soil moisture could enhance the model's reliability and provide stronger support for data-driven agricultural decision-making.

Bibliography

Akaike, H (1974). "A new look at the statistical model identification". In: *IEEE Transactions on Automatic Control* 19.6, pp. 716–723.

Anselin, L (1995). "Local indicators of spatial association—LISA". In: *Geographical Analysis* 27.2, pp. 93–115.

Atkinson, A C (1985). *Plots, Transformations, and Regression: An Introduction to Graphical Methods of Diagnostic Regression Analysis*. Oxford University Press.

Box, G E P and D R Cox (1964). "An analysis of transformations". In: *Journal of the Royal Statistical Society*, pp. 211–252.

Cleveland, W S (1979). "Robust locally weighted regression and smoothing scatterplots". In: *Journal of the American Statistical Association* 74.368, pp. 829–836.

Colvin, T S, D B Jaynes, D L Karlen, D A Laird, and J R Ambuel (1997). "Yield variability within a central Iowa field". In: *Transactions of the ASAE*.

Cook, R D (1977). "Detection of Influential Observations in Linear Regression". In: *Technometrics* 19.1, pp. 15–18.

Cox, D R and E J Snell (1968). "A General Definition of Residuals". In: *Journal of the Royal Statistical Society. Series B (Methodological)* 30.2. 28 pages, pp. 248–275.

Draper, N R and H Smith (1998). *Applied Regression Analysis*. 3rd. Wiley. Chap. 13, pp. 280–285.

Efron, B and R Tibshirani (1994). *An introduction to the bootstrap*. Chapman & Hall/CRC.

Fox, J (2016). *Applied Regression Analysis and Generalized Linear Models*. Third edition. SAGE Publications. Chap. 4.

Fox, J and S Weisberg (2019). *car: Companion to Applied Regression*. R package version 3.1-0. URL: `https://CRAN.R-project.org/package=car`.

Getis, A and J K Ord (1992). "The analysis of spatial association by use of distance statistics". In: *Geographical Analysis* 24.3, pp. 189–206.

Gnanadesikan, R (1977). *Methods for Statistical Data Analysis of Multivariate Observations*. John Wiley & Sons.

Goldfeld, S M and R E Quandt (1972). *Nonlinear Methods in Econometrics*. Vol. 77. Contributions to Economic Analysis. Amsterdam: North-Holland Publishing Company.

Kass, R E and A E Raftery (1995). "Bayes Factors". In: *Journal of the American Statistical Association* 90.430, pp. 773–795.

Li, H, C A Calder, and N Cressie (2007). "Beyond Moran's I: Testing for Spatial Dependence Based on the Spatial Autoregressive Model". In: *Geographical Analysis* 39.4, pp. 357–375.

Moran, P A P (1950). "Notes on continuous stochastic phenomena". In: *Biometrika* 37.1-2, pp. 17–23.

Pearson, K (1895). "Notes on regression and inheritance in the case of two parents". In: *Proceedings of the Royal Society of London* 58, pp. 240–242.

R Core Team (2025). *R: A Language and Environment for Statistical Computing*. Version 4.5.0 (2025-04-11). R Foundation for Statistical Computing. Vienna, Austria. URL: `https://www.R-project.org/`.

Rencher, A C and W F Christensen (2012). *Methods of Multivariate Analysis*. 3rd. Wiley.

Ripley, B (2021). *boot: Bootstrap Functions (Originally by Angelo Canty for S)*. R package version 1.3-28. URL: `https://CRAN.R-project.org/package=boot`.

Saikkonen, P (2017). "*Linear models*". Lecture notes, University of Helsinki.

Schloerke, B, D Cook, H Wickham, J Crowley, D Hofmann, M Marbach, E Anderson, G Grolemund, and S Wang (2023). *GGally: Extension to 'ggplot2'*. R package version 2.1.2. URL: `https://CRAN.R-project.org/package=GGally`.

Schwarz, G E (1978). "Estimating the dimension of a model". In: *The Annals of Statistics* 6.2, pp. 461–464.

Sharma, S, D A Swayne, and C Obimbo (2015). "Automating the Smoothing of Time Series Data". In: *Journal of Environmental & Analytical Toxicology* 5.5.

Spitzer, J J (1982). "A Primer on Box-Cox Estimation". In: *Review of Economics and Statistics* 64.2, pp. 307–313.

Velilla, S (1993). "A note on the multivariate Box-Cox transformation to normality". In: *Statistics & Probability Letters*.

Venables, W N and B D Ripley (2002). *Modern Applied Statistics with S*. Fourth. R package version 7.3-60. New York: Springer.

Wald, A (1943). "Tests of statistical hypotheses concerning several parameters when the number of observations is large". In: *Transactions of the American Mathematical Society* 54.3, pp. 426–482.

Weisberg, S (2005). *Applied Linear Regression*. 3rd ed. Wiley Series in Probability and Statistics. John Wiley & Sons, Inc.

Wickham, H (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. URL: `https://ggplot2.tidyverse.org`.

Wickham, H and J Bryan (2023). *readxl: Read Excel Files*. R package version 1.4.2. URL: https://CRAN.R-project.org/package=readxl.

Wilk, M B and R Gnanadesikan (1968). "Probability plotting methods for the analysis of data". In: *Biometrika* 55.1, pp. 1–17.

Wooldridge, J M (2009). *Introductory Econometrics: A Modern Approach*. 4th. Boston, MA: Cengage Learning.

Appendix A

Summary of the estimated models

Summary of the Model 1

```
Call:
lm(formula = Corn97BU ~ (Zn + K + Mn + P + Na + Mg + B + Ca +
    Fe + Cu), data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-66.647  -8.132   0.289   8.051  38.485

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 99.9477940  9.0365559  11.060  < 2e-16 ***
Zn           0.0605328  2.4457687   0.025 0.980279
K            0.0250436  0.0442193   0.566 0.571779
Mn           0.0616807  0.0537394   1.148 0.252406
P            0.1231888  0.1181029   1.043 0.298154
Na           0.2425047  0.7193104   0.337 0.736362
Mg           0.0092341  0.0129092   0.715 0.475237
B           -0.2203443  0.9570173  -0.230 0.818136
Ca           0.0006444  0.0005269   1.223 0.222758
Fe           0.0908001  0.0251081   3.616 0.000377 ***
Cu           8.4303066  2.2915733   3.679 0.000300 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 15.14 on 204 degrees of freedom
Multiple R-squared:  0.4945, Adjusted R-squared:  0.4697
F-statistic: 19.95 on 10 and 204 DF,  p-value: < 2.2e-16
```

Summary of the Model 2

```
Call:
lm(formula = y_transformed ~ I(Zn^(0.5)) + log(K) + log(Mn) +
    log(P) + log(Na) + log(Mg) + I(B^0.5) + log(Ca) + I(Fe^0.5) +
    I(Cu^(0.5)), data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-282060  -61261   -2766   48437  242658

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -414767.0   251251.3  -1.651  0.10032
I(Zn^(0.5))   23595.5    30550.0   0.772  0.44080
log(K)        38296.1    28901.6   1.325  0.18664
log(Mn)        6091.6    27646.9   0.220  0.82583
log(P)        37922.7    22761.5   1.666  0.09723 .
log(Na)      -12311.8    55747.1  -0.221  0.82543
log(Mg)       23083.8    32336.1   0.714  0.47612
I(B^0.5)       -419.9    22037.1  -0.019  0.98482
log(Ca)       14256.2    27998.4   0.509  0.61118
I(Fe^0.5)      3911.0     4036.9   0.969  0.33379
I(Cu^(0.5))  117007.8    43201.3   2.708  0.00733 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 85610 on 204 degrees of freedom
Multiple R-squared:  0.5265, Adjusted R-squared:  0.5033
F-statistic: 22.69 on 10 and 204 DF,  p-value: < 2.2e-16
```

Summary of the Model 2A

```
Call:
lm(formula = y_transformed ~ I(Zn^(0.5)) + log(K) + log(P) +
    I(Cu^(0.5)), data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-299911  -57316    -963   45234  239991

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -190920      79519  -2.401   0.0172 *
I(Zn^(0.5))    39622      25412   1.559   0.1205
log(K)         47372      25238   1.877   0.0619 .
log(P)         30105      17527   1.718   0.0371 *
I(Cu^(0.5))   142108      23735   5.987 9.13e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 84910 on 210 degrees of freedom
Multiple R-squared:  0.5205, Adjusted R-squared:  0.5114
F-statistic:    57 on 4 and 210 DF,  p-value: < 2.2e-16
```

Summary of the Model 2i

```
Call:
lm(formula = y_transformed ~ (I(Zn^(0.5)) + log(K) + log(Mn) +
    log(P) + log(Na) + log(Mg) + I(B^0.5) + log(Ca) + I(Fe^0.5) +
    I(Cu^(0.5)))^2, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-192732  -46385   -3404   41936  233067

Coefficients:
                        Estimate Std. Error t value Pr(>|t|)
(Intercept)              6640670    8985542   0.739   0.4610
I(Zn^(0.5))              2270775    1239963   1.831   0.0689 .
log(K)                   2401624    1464480   1.640   0.1030
log(Mn)                  1704624    1231671   1.384   0.1683
log(P)                  -1866165    1112824  -1.677   0.0955 .
log(Na)                   448593    3347110   0.134   0.8936
log(Mg)                 -2488421    1564790  -1.590   0.1138
I(B^0.5)                 868263    1247073   0.696   0.4873
log(Ca)                 -2136731    1134389  -1.884   0.0614 .
I(Fe^0.5)                -173130     204803  -0.845   0.3992
I(Cu^(0.5))              1778575    1675866   1.061   0.2902
I(Zn^(0.5)):log(K)         -5143     170679  -0.030   0.9760
I(Zn^(0.5)):log(Mn)       -50441     177763  -0.284   0.7770
I(Zn^(0.5)):log(P)       -110773     114774  -0.965   0.3359
I(Zn^(0.5)):log(Na)        12478     456545   0.027   0.9782
I(Zn^(0.5)):log(Mg)      -198839     220766  -0.901   0.3691
I(Zn^(0.5)):I(B^0.5)       84702     146985   0.576   0.5653
I(Zn^(0.5)):log(Ca)      -165647     149252  -1.110   0.2687
I(Zn^(0.5)):I(Fe^0.5)     -15731      27470  -0.573   0.5677
I(Zn^(0.5)):I(Cu^(0.5))   596334     261718   2.279   0.0240 *
log(K):log(Mn)           -245916     166796  -1.474   0.1424
log(K):log(P)              67449      50270   1.342   0.1816
log(K):log(Na)            348645     342242   1.019   0.3099
log(K):log(Mg)            147258     219112   0.672   0.5025
log(K):I(B^0.5)          -166223     127567  -1.303   0.1944
log(K):log(Ca)           -389104     181679  -2.142   0.0337 *
log(K):I(Fe^0.5)          -56707      23180  -2.446   0.0155 *
log(K):I(Cu^(0.5))        640110     262473   2.439   0.0158 *
log(Mn):log(P)             87230     128320   0.680   0.4976
log(Mn):log(Na)            36322     318962   0.114   0.9095
log(Mn):log(Mg)          -213252     182671  -1.167   0.2448
```

```
log(Mn):I(B^0.5)           71534     129735   0.551   0.5821
log(Mn):log(Ca)            52464     153277   0.342   0.7326
log(Mn):I(Fe^0.5)         -14553      20743  -0.702   0.4840
log(Mn):I(Cu^(0.5))        -8276     206426  -0.040   0.9681
log(P):log(Na)           -654461     274163  -2.387   0.0182 *
log(P):log(Mg)             68998     173142   0.399   0.6908
log(P):I(B^0.5)           166554      94671   1.759   0.0805 .
log(P):log(Ca)            317713     130149   2.441   0.0157 *
log(P):I(Fe^0.5)           34531      17930   1.926   0.0559 .
log(P):I(Cu^(0.5))       -448012     177173  -2.529   0.0124 *
log(Na):log(Mg)          -422622     393793  -1.073   0.2848
log(Na):I(B^0.5)           69132      88686   0.780   0.4368
log(Na):log(Ca)           116283     342645   0.339   0.7348
log(Na):I(Fe^0.5)          49136      43045   1.142   0.2554
log(Na):I(Cu^(0.5))       426078     557803   0.764   0.4461
log(Mg):I(B^0.5)          -95453     152874  -0.624   0.5333
log(Mg):log(Ca)           508291     197067   2.579   0.0108 *
log(Mg):I(Fe^0.5)          45476      23663   1.922   0.0564 .
log(Mg):I(Cu^(0.5))      -334409     258789  -1.292   0.1982
I(B^0.5):log(Ca)          -89162     135282  -0.659   0.5108
I(B^0.5):I(Fe^0.5)         -3323      17646  -0.188   0.8509
I(B^0.5):I(Cu^(0.5))       60115     210671   0.285   0.7757
log(Ca):I(Fe^0.5)          12084      13155   0.919   0.3597
log(Ca):I(Cu^(0.5))      -298476     194360  -1.536   0.1266
I(Fe^0.5):I(Cu^(0.5))     -62809      35238  -1.782   0.0766 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 82760 on 159 degrees of freedom
Multiple R-squared:  0.6551, Adjusted R-squared:  0.5358
F-statistic: 5.492 on 55 and 159 DF,  p-value: < 2.2e-16
```

Summary of the Model 2iB

```
Call:
lm(formula = y_transformed ~ log(P) + log(Na) + I(Cu^(0.5)) +
    log(P):log(Na), data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-282281  -51880   -5805   48000  219834

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)    -1636453     489387  -3.344 0.000978 ***
log(P)           531878     144942   3.670 0.000308 ***
log(Na)          569808     178514   3.192 0.001630 **
I(Cu^(0.5))      167617      17061   9.825  < 2e-16 ***
log(P):log(Na)  -169860      52574  -3.231 0.001433 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 84370 on 210 degrees of freedom
Multiple R-squared:  0.5266, Adjusted R-squared:  0.5176
F-statistic: 58.41 on 4 and 210 DF,  p-value: < 2.2e-16
```