UNIVERSITY OF NAIROBI
A world class university committed to Scholarly Excellence

FACULTY OF SCIENCE AND TECHNOLOGY
DEPARTMENT OF MATHEMATICS

MAIZE YIELD PREDICTION USING SIMPLE LINEAR REGRESSION MODEL

BY:

SHAMIM RASHID	I63/4255/2019
VICTORINE IMBUHILA	I63/4262/2019
OLWENY LYNN NICOLE ADHIAMBO	I63/137207/2019
JUSTUS ORIYA	I63/136922/2019
GABRIEL HAMGERA OBUWA	I63/4234/2019

SUPERVISOR: Dr. Musiga

June 9, 2023

Declaration

We declare that this research project is our original work and has not been present in any institution for any award or conferment of any degree.

NAME	REG NO.	SIGNATURE
SHAMIM RASHID	I63/4255/2019	....
VICTORINE IMBUHILA	I63/4262/2019	....
OLWENY LYNN NICOLE ADHIAMBO	I63/137207/2019	....
JUSTUS ORIYA	I63/136922/2019	....
GABRIEL HAMGERA OBUWA	I63/4234/201 9	....

This project has been submitted for examination with my approval as a University Supervisor.

DR. MUSIGA
Signed: .... Date: ....
LECTURER
SCHOOL OF MATHEMATICS, UNIVERSITY OF NAIROBI.

Dedication

We dedicate this project to our parents, lecturers, relatives, and friends for their firm support during our entire stay at the university.

Acknowledgment

We cannot express enough gratitude to God for seeing us through our four years on campus and most especially during our project. Secondly, we would like to appreciate our lecturers at the School of Mathematics, University of Nairobi for having taught us different Statistics skills which we have implemented in our project and our supervisor Dr. Musiga for guiding us entirely.

We would also like to thank our parents, guardians, and siblings for their able guidance and support in different aspects of our lives and through their support we were able to complete this project.

Finally, we would like to appreciate ourselves for putting in the time and effort to complete this project. For this we say, thank you!

Contents

Chapter 1
INTRODUCTION

1.1 Background of the Study

Drought and other severe climatic conditions immensely contribute to acute food insecurity in Africa, specifically in the Horn of Africa. Over the years, an increase in the population has resulted in the destruction of the environment, especially with the clearance of forests to accommodate the growing population. Farmers have realized reduced productivity both for themselves and for the market. Unsustainable use of the available land in arid and semi-arid areas and encroachment of settlement in arable land has exacerbated the issue. The employment of technology in analyzing land use and how best to maximize productivity is key to the realization of agricultural produce throughout the region.

1.2 Statement of the Problem

The world over, governments aim to attain food security. Homegrown agricultural produce assure both subsistence and affordable food prices. Thus, the problem statement is the determination of the relationship between Normalized Difference Vegetation Index(NDVI) and maize in Kenya. NDVI is used to determine the density of vegetation health. The higher the NDVI value, the healthier the vegetation and vice versa.

1.3 Objectives of the Study

1.3.1 General Objective

- The general objective is the prediction of maize yield using simple linear regression model.

1.3.2 Specific Objective

The specific objectives are:

- To perform exploratory data analysis to identify patterns, trends, and relationships between maize yield and NDVI.
- To fit and evaluate the fit of the model
- To use the model for prediction of future maize yields

1.4 Methodology

- The theory of Simple Linear Regression Model was applied in this study to analyze the relationship between NDVI and maize yield.
- R software was used to fit the models and also in the prediction of maize yield for the years 2023-2027
- The data for the study was collected from the Ministry of Agriculture, Kilimo House, consisting of maize yield records, and from the Kenya Space Agency, consisting of NDVI measurements.

1.5 Assumptions

The following assumptions are made in this work:

(i) The error term ε has zero mean.

(ii) The error term ε has constant variance σ^2.

(iii) The errors are uncorrelated.

(iv) The errors are normally distributed.

1.6 Justification of the Study

The study will enable the government and non-governmental organizations(NGOs) to predict maize yield for food security.

Chapter 2
LITERATURE REVIEW

In recent years, the use of satellite imagery has become a promising tool in addressing food insecurity in developing countries. By providing detailed and accurate information on land use, crop yields, and other indicators of food security, satellite imagery can help bridge the gap in food insecurity data and inform policy decisions. The following are the previous research that have addressed how to further improve food security by improving methods of yield estimation.

Ines et al. (2013) [3] developed a data assimilation-crop modeling framework that improved the prediction of crop yields at an aggregate scale. The authors modified the Decision Support System for Agro-technology Transfer – Cropping System Model (DSSAT-CSM)-Maize model. The core of the framework, Ensemble Kalman Filter (EnKF), was used to control crop model runs, assimilate remote sensing (RS) data, and update model state variables. Data was collected over a period of six years from 2003-2009 in Iowa, USA. The authors assimilated AMSR-E soil moisture and MODIS-LAI data independently and simultaneously. Assimilating the leaf area index, LAI, independently slightly improved the correlation of observed and simulated yields, but more significant improvements were seen when both were assimilated simultaneously. Results suggested that assimilation of LAI independently was preferable when conditions were extremely wet while assimilation of soil moisture and LAI would be more suitable under normal conditions.

Kasampalis et al. (2018) [4] reviewed the contribution of remote sensing data to crop growth models. Authors noted that crop models were classified into empirical models and dynamic models also known as process-based models. The empirical models were statistical models that used regression with one or more parameters. The regression models were often used to estimate crop yield in a particular environment so as to provide useful insights to policymakers about production options. Dynamic models simulated crop progression using differential equations through time. However, the process based models required significantly large inputs of data which in hindsight limited their usefulness for research. Remote sensing provided timely, non-destructive and accurate estimation of land resources.

Sayago and Bocco (2018) [7] conducted a study on the use of satellite images for crop yield estimation. They compared the performance of linear and non-linear models, specifically Multiple Linear Regression (MLR) and Neural Networks (NN), for estimating crop yields using satellite data. The authors used Landsat 8 and SPOT 5 images, which were preprocessed using the dark object subtraction method and the ENVI 4.6.1 software. The data consisted of 48,783 randomly selected surface reflectance values, with half of the data used for model training and the other half for model validation. The study found that the nonlinear models, specifically NN, outperformed MLR in estimating crop yields using satellite data. They concluded that the use of NN models provided accurate and reliable estimates of crop yield using satellite imagery, which has important implications for crop monitoring and management.

Torre et al. (2021) [8] reviewed the use of various models to estimate yield of rice. The primary inputs for rice- health estimation were the Leaf Area Index(LAI), radar backs-scatter, biomass, Normalized Difference Vegetation Index (NDVI), Canopy cover(CC), Fraction of Absorbed Photosynthetic Radiation(FAPR). The characteristics of rice at different stages of growth were detected using optical, radar sensors and satellite Imagery. Spectral responses from plant pigmentation such as chlorophyll strongly reflected near-infrared radiation (NIR) and green light while it strongly absorbed blue and red lights(R). The NDVI measured plant greenness with the value ranged from -1.0 to 1.0, the higher the NDVI value the healthier the plant. The NDVI time series imagery used Sentinel data from February 2019 to November 2019. The back-scatter coefficient which were sensitive to vegetation canopy structure, underlying ground surface structure and growth stages arose from the interaction of microwave signals with the target surface. The authors noted that empirical models that were employed in the past included machine learning, neural networks and random forests, multiple-linear regression and logistic regression. Crop models utilized both optical data and vegetation indices included ORYZA, WOFOST, and SIMRIW. The semi-empirical models involved the application both statistical methods and crop models. This had been made possible by the data-rich resources availed from ground and satellite observation. Assimilation of data aimed to combine the Canopy Cover (CC) and Leaf Area Index (LAI) with NDVI in order to optimize model parameters.

Ali et al. (2022) [1] reviewed crop yield prediction using Multi Sensors Remote sensing. Through the help of National Authority for Remote Sensing and Space Science(NARSS) and support from the Strategic academic leadership program (RUDN University) the authors identified the different types of sensors that are used to estimate crop yield production such as satellite imagery, aerial photography, RADAR and LIDAR and Field sensors. In conclusion it was found that remote sensing could be employed in all aspects of agricultural process from land preparation to harvesting. Moreover, Satellite data is still the most efficient remote sensing technique for monitoring national and regional changes. The use of satellite imagery and hyper-spectral data were found to retrieve crop biophysical and biochemical parameters.

A more recent study conducted by Roznik et al. (2022) [6] investigated how crop yield estimation could be improved by applying higher resolution satellite imagery and high resolution cropland masks. The data used included Normalized Difference Vegetation Indices (NDVI) which were derived using satellite images at different resolutions. The crop data layer was obtained from the United States Department of Agriculture (USDA). NDVI and crop yield data was collected over a period of 11 years for four crops in 48 US states from the year 2008 to 2018. Cubic spline regression method was used . The results obtained indicated that the accuracy of crop yield estimation models could be moderately improved using higher satellite resolution NDVI.

In conclusion, previous works that have been done used a crop simulation model along with remotely sensed data to improve the accuracy of crop yield prediction, a comparison of linear and non-linear models for crop yield estimation using satellite images and higher resolution satellite imagery and cropland masks to enhance crop yield estimation. Some of the researchers reviewed the integration of remote sensing data into crop models to enhance the prediction of crop yields, remote sensing-based estimation of rice yields and multi-sensor remote sensing data to predict crop yield. In this work, the Simple Linear Regression model is used to analyze vegetation health (NDVI) and agricultural productivity.

Chapter 3
SIMPLE LINEAR REGRESSION MODEL

3.1 Introduction

In this chapter, the theory of linear regression is presented, specifically, the Simple Linear Regression Model. The theory is presented in the following subsections.

3.2 Simple Linear Regression Model

Simple linear regression is a method for modelling the relationship between two variables, where one variable (the independent variable), is used to predict the value of the other variable (the dependent variable).It is a model with a single predictor variable x that has a relationship with a response variable y that is a straight line. We denote it using the equation

$$y = \beta_0 + \beta_1 x + \varepsilon \tag{3.1}$$

where, the intercept β_0 and the slope β_1 are unknown constants and ε is a random error component. The errors are assumed to have mean zero and unknown variance σ^2. Additionally, we usually assume that the errors are uncorrelated. This means that the value of one error does not depend on the value of any other error. It is convenient to view the regressor x as controlled by the data analyst and measured with negligible error, while the response y is a random variable. That is, there is a probability distribution for y at each possible value for x. The mean of this distribution is

$$E(y|x) = \beta_0 + \beta_1 x \tag{3.2}$$

and the variance is

$$Var(y|x) = Var(\beta_0 + \beta_1 x + \varepsilon) = \sigma^2 \tag{3.3}$$

Thus, the mean of y is a linear function of x although the variance of y does not depend on the value of x. Furthermore, because the errors are uncorrelated, the responses are also uncorrelated. The parameters β_0 and β_1 are usually called regression coefficients. These coefficients have a simple and often useful interpretation. The slope β_1 is the change in the mean of the distribution of y produced by a unit change in x. If the range of data on x includes $x = 0$, then the intercept β_0 is the mean of the distribution of the response y when $x = 0$. If the range of x does not include zero, then β_0 has no practical interpretation.

3.3 Estimation of Parameters

Suppose that we have n pairs of data, say $(y_1, x_1), (y_2, x_2), ..., (y_n, x_n)$ and that the parameters β_0 and β_1 are unknown and must be estimated using sample data.

3.3.1 Estimation of β_0 and β_1

The method of least squares is used to estimate β_0 and β_1. We write

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, i = 1, 2, ..., n \tag{3.4}$$

where this is a sample regression model, written in terms of the n pairs of data (y_i, x_i) (i = 1, 2, ..., n). Thus, the least-squares criterion is

$$s(\beta_0, \beta_1) = \sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2 \tag{3.5}$$

The least-squares estimators of β_0 and β_1 are obtained by minimizing the total squared prediction errors, thus

$$\frac{\partial S}{\partial \beta_0} = -2\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0 \tag{3.6}$$

and

$$\frac{\partial S}{\partial \beta_1} = -2\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)x_i = 0 \tag{3.7}$$

Simplifying these two equations yields,

$$n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i \tag{3.8}$$

$$\hat{\beta}_0 \sum_{i=1}^{n} x_i + \hat{\beta}_1 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} y_i x_i \tag{3.9}$$

the least-squared normal equations. The solutions to the normal equations are

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \tag{3.10}$$

and

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} y_i x_i - \frac{[\sum_{i=1}^{n} y_i][\sum_{i=1}^{n} x_i]}{n}}{\sum_{i=1}^{n} x_i^2 - \frac{[\sum_{i=1}^{n} x_i]^2}{n}} \tag{3.11}$$

where $\bar{x}$ and $\bar{y}$ are the averages of y_i and x_i, respectively. The fitted simple linear regression model gives a point estimate of the mean of y for a particular x.

3.3.2 Estimation of σ^2

The fitted simple linear regression model is

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x \tag{3.12}$$

which gives a point estimate of the mean of y for a particular x.
The difference between the observed value y_i and the corresponding fitted value $\hat{y}_i$ is a residual. Mathematically the i^{th} residual is

$$e_i = y_i - \hat{y}_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i), i = 1, 2, ..., n \tag{3.13}$$

Residuals play an important role in investigating model adequacy and in detecting departures from the underlying assumptions.

In addition to estimating β_0 and β_1, an estimate of σ^2 is required to test hypotheses and construct interval estimates pertinent to the regression model. Ideally we would like this estimate not to depend on the adequacy of the fitted model. This is only possible when there are several observations on y for at least one value of x or when prior information concerning σ^2 is available. When this approach cannot be used, the estimate of σ^2 is obtained from the residual or error sum of squares,

$$SS_{Res} = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \tag{3.14}$$

A convenient computing formula for SS_{Res} may be found by substituting $\hat{y}_i$=$\hat{\beta}_0$+$\hat{\beta}_1 x_i$ into the above equation and simplifying, yielding

$$SS_{Res} = \sum_{i=1}^{n} y_i^2 - n\bar{y}^2 - \hat{\beta}_1 S_{xy} \tag{3.15}$$

But

$$\sum_{i=1}^{n} y_i^2 - n\bar{y}^2 = \sum_{i=1}^{n} (y_i - \bar{y})^2 = SS_T \tag{3.16}$$

is just the corrected sum of squares of the response observations, so

$$SS_{Res} = SS_T - \hat{\beta}_1 S_{xy} \tag{3.17}$$

The residual sum of squares has n-2 degrees of freedom.

3.4 Hypothesis Testing

Hypothesis testing is a statistical tool used to make decisions about population parameters based on sample data. In this section, we discuss hypothesis testing in the context of a simple linear regression model. In a simple linear regression model, hypothesis testing is used to determine whether there is a significant linear relationship between the predictor and the response variables. We focus on testing the null hypothesis that the slope coefficient of the regression model is equal to zero, that is, there is no linear relationship between the predictor variable and the response variable.

3.4.1 The t-tests

To test whether there is a significant linear relationship between the predictor and the response variables, we need to test the hypothesis that the slope is different from zero, that is, not equal to zero. The hypotheses are:

$$H_0 : \beta_1 = 0 \text{ vs } H_1 : \beta_1 \neq 0$$

where, β_1 is the slope coefficient of the regression model and the hypotheses is a two sided alternative.

The null hypothesis for the test, $\beta_1 = 0$, means that there is no relationship between the predictor and response variables. The alternative hypothesis, $\beta_1 \neq 0$, indicates that there is a non-zero slope and therefore a significant linear relationship between the predictor and response variables.
The t-test statistic for the hypothesis test is:

$$t = \frac{\hat{\beta_1}}{s.e(\hat{\beta_1})} \tag{3.18}$$

where, $\hat{\beta_1}$ is the estimate of the slope obtained from fitting the regression model and $s.e(\hat{\beta_1})$ is the estimated standard error or the standard error of $\hat{\beta_1}$. The t-test statistic follows the $t(n-2)$ distribution.The null hypothesis,H_0, is rejected if $|t| > t_{\alpha/2,n-2}$. Rejecting the null hypothesis implies that, there is a statistically significant linear relationship between the response and predictor variables. This therefore implies that the predictor variable is of value in explaining the variability in the response.

Alternatively, we could use the p-value approach for the decision making. We use the $t(n-2)$ distribution to obtain the p-value for the test. If the p-value is less than the chosen level of significance, α, reject the null hypothesis and conclude that there is evidence of a significant linear relationship between the predictor and response variables. Otherwise, we fail to reject the null hypothesis and conclude that there is not enough evidence to indicate a significant linear relationship.

It should be noted that the t-test assumes that the errors in the regression model are normally distributed and have constant variance. If these assumptions are not met, the results of the test may not be valid and alternative tests, such as non-parametric tests, may be needed. Additionally, there may be other factors besides the predictor variable that influence the response variable, and failing to account for these factors may also bias the results of the analysis.

3.4.2 Analysis of Variance (ANOVA)

Analysis of variance approach is another way to test the significance of regression. This method is based on partitioning the total variability in the response variable. The variability in the response can be partitioned into two components:

(i) Variability accounted for by the regression model referred to as the explained variation.

(ii) Variability due to the error which is referred to as unexplained variation.

In other words, the total variation is expressed as the sum of the explained variation and the unexplained variation.

$$\Sigma_{i=1}^{n}(y_i - \bar{y})^2 = \Sigma_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + \Sigma_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{3.19}$$

The left-hand side of the equation is the corrected sum of squares of the observations, SST, which measures the total variability in the observations. The two components of SST measure, the amount of variability in the observations accounted for by the regression line and the residual variation left unexplained by the regression line, respectively. Symbolically we write:

$$\text{SST} = \text{SSR} + \text{SSE} \tag{3.20}$$

The degrees of freedom breakdown is as follows: SST has n-1 degrees of freedom, SSR has 1 degree of freedom and SSE has n-2 degrees of freedom.
In this case, we use the F-test to test our hypotheses:

$$H_0 : \beta_1 = 0 \text{ vs } H_1 : \beta_1 \neq 0$$

The ratio of the explained to unexplained variation is used in determining if the fitted model is statistically significant.
The F-test statistic is:

$$F_0 = \frac{\text{SSR}/1}{\text{SSE}/n-2} = \frac{\text{MSR}}{\text{MSE}} \tag{3.21}$$

The F-test statistic follows the $F(1, n-2)$ distribution. To test the hypotheses, we compute the test statistic F_0 and reject H_0 if: $F_0 > F_{\alpha,1,n-2}$ The results obtained can be presented in tabular form:

Table 1: Analysis of Variance table.

Source of variation	Degrees of freedom	Sum of squares	Mean sum of squares	F-ratio,F_0
Regression	1	SSR	MSR=SSR/1	MSR/MSE
Residuals (Error)	n-2	SSE	MSE=SSE/n-2	
Total	n-1	SST		

3.5 Prediction of New Observations

An important application of the regression model is prediction of new observations y corresponding to a specified level of the regressor variable x. If x_0 is the value of the regressor variable of interest, then

$$\hat{y_0} = \hat{\beta}_0 + \hat{\beta}_1 x_0 \tag{3.22}$$

is the point estimate of the new value of response y_0. We now develop a prediction interval for the future observation y_0. The random variable

$$\psi = y_0 - \hat{y_0} \tag{3.23}$$

is normally distributed with mean zero and variance

$$Var(\psi) = Var(y_0 - \hat{y_0}) = \sigma[1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}] \tag{3.24}$$

Since y_0 is independent of $\hat{y_0}$, then, the standard error of $\psi = y_0 - \hat{y_0}$ is the appropriate statistic on which to base a prediction interval. Thus, the $100(1-\alpha)$ percent prediction interval on a future observation at x_0 is

$$\begin{aligned} \hat{y_0} - t_{\alpha/2,n-2}\sqrt{MS_{Res}(1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}})} \leq y_0 \\ \leq \hat{y_0} + t_{\alpha/2,n-2}\sqrt{MS_{Res}(1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}})} \end{aligned} \tag{3.25}$$

3.6 Coefficient of Determination

The coefficient of determination, R^2 is a summary of the strength of the relationship between the x_i and the y_i in the data. It is calculated by:

$$R^2 = \frac{SS_R}{SS_T} = \frac{SS_{Res}}{SS_T} \tag{3.26}$$

The total sum of squares, SS_T, is a measure of the variability in y without considering the effect of the regressor variable x and SS_{Res} is a measure of the variability in y remaining after x has been considered, R^2 is the proportion of variation explained by the regressor x.

Given $0 \leq SS_{Res} \leq SS_T$, it follows that $0 \leq R^2 \leq 1$. The values that are close to 1 imply that most of the variability in y is explained by the regression model.

The coefficient of determination only measures the goodness-of-fit of the linear regression model, and it doesn't provide information on whether the model meets the assumptions of a linear model. Additionally, a large value for the coefficient of determination may not necessarily mean that the regression model has a good predictive power, as it may suffer from over fitting or lack generalizability to new data points.

Therefore, when interpreting a linear regression model, it is important to consider other factors besides R^2, such as residual analysis, model assumptions, and the meaningfulness of the independent variable(s) in explaining the dependent variable.

3.7 The Residuals

The most common plot used to assess the assumptions of a linear regression model is the plot of residuals versus the fitted values. A null plot with no pattern indicates that the model meets the assumptions of a linear model. If there is a curvature in the plot, it may suggest that the linear model is inappropriate and may require a more flexible model such as a quadratic or cubic polynomial regression model. If there is an increase or decrease in the average magnitude of residuals with the fitted values, it may indicate non-constant residual variance, which can be addressed by using weighted least squares or robust regression models.

Chapter 4
APPLICATION OF THE MODEL AND DISCUSSIONS

4.1 Introduction

In this chapter, the Simple Linear Regression model is applied to examine and predict the relationship between the predictor(NDVI) and response(yield) variables. It helps in the identification of the presence of a linear relationship between the predictor and response variables, allowing us to estimate how changes in the predictor variable impact the response variable. Vegetation health is the overall well-being of plants in terms of how well the plants are growing and thriving in a particular area. It takes into account factors like their overall condition, growth rate and ability to resist diseases or environmental stresses, indicating the overall health and vitality of plant ecosystems. Agricultural productivity is the amount of crops or agricultural products that are produced from a given area of land or resources. It shows the efficiency and effectiveness of agricultural practices in generating sufficient yields to meet demands and sustain food production.
The next section presents the data for the study.

4.2 Data for the Study

The data for the study, maize yield, was obtained from the Ministry of Agriculture, Kilimo House. The data obtained was for a period of five years from 2018-2022. To study maize yield in the country, five counties were selected because they are considered the bread basket of the country due to their high agricultural output, maize in specific. The selection was guided by the historical maize yield data. The selection was guided by the historical maize yield data. The counties selected for the study were Bungoma, Nakuru, Nandi, Trans-Nzoia and Uasin Gishu. The table below shows the maize yield per hectare for the five counties between 2018-2022.

Table 2: Table showing maize yield for the five counties from 2018-2022

County	Yield/ha(2018)	Yield/ha(2019)	Yield/ha(2020)	Yield/ha(2021)	Yield/ha(2022)
Bungoma	3.16	3	3.66	2.89	2.98
Nakuru	1.14	3	3.19	2.77	3.22
Nandi	1.01	3	2.80	2.39	1.96
Trans-Nzoia	5.09	4	4.60	3.90	3.57
Uasin Gishu	4.26	3	3.07	3.68	3.50

The Normalized Difference Vegetation Index(NDVI) is a commonly used vegetation index in remote sensing and is calculated from satellite or aerial imagery. NDVI provides information about the health and density of vegetation cover in a given area.
Satellite images were obtained from Google Earth Engine, as directed by the Kenya Space Agency. Mean NDVI values were then obtained for each year(2018-2022) in every county.NDVI

is an indicator of vegetation health and density, allowing us to assess the overall vegetation conditions in these counties. The legend explains variation of NDVI values in the specified counties. The images below show NDVI maps for the 5 counties.

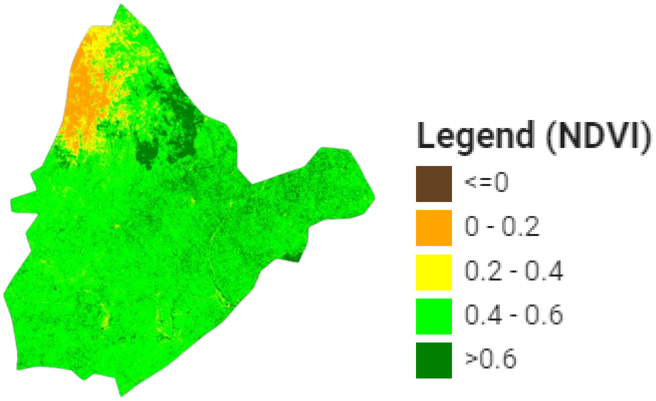

Figure 1: NDVI Image for Bungoma County

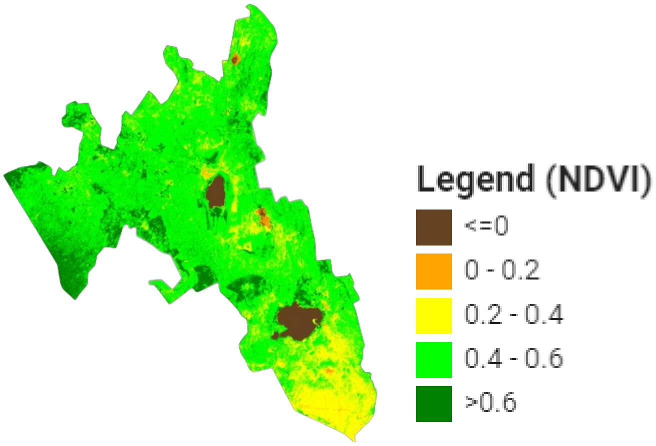

Figure 2: NDVI Image for Nakuru County

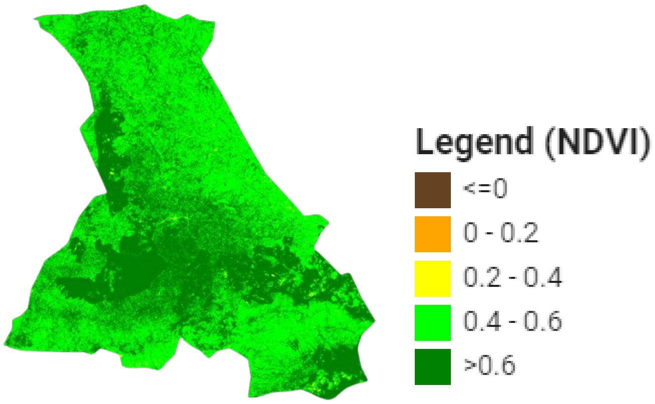

Figure 3: NDVI Image for Nandi County

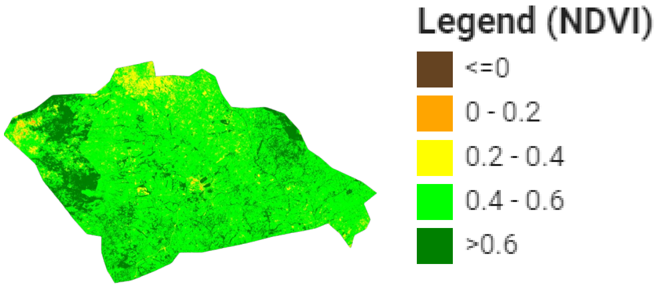

Figure 4: NDVI Image for Trans Nzoia County

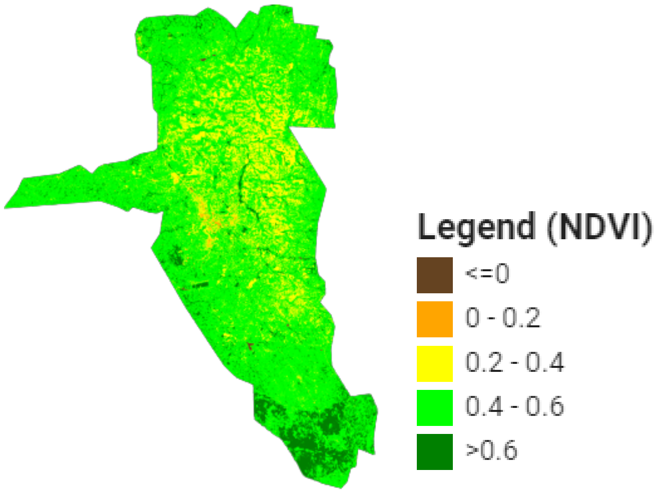

Figure 5: NDVI Image for Uasin Gishu County

The table below shows mean NDVI values for the five counties between 2018-2022

Table 3: Table showing NDVI for the five counties from 2018-2022

County	NDVI 2018	NDVI 2019	NDVI 2020	NDVI 2021	NDVI 2022
Bungoma	0.56842	0.356925	0.480692	0.353525	0.297571
Nakuru	0.3033913	0.34985	0.3917917	0.39055	0.357579
Nandi	0.2955	0.331462	0.285267	0.206318	0.2223
Trans Nzoia	0.71	0.633	0.67	0.6136	0.4815
Uasin Gishu	0.657882353	0.4365	0.452947	0.384	0.364

In the following section, the regression and correlation plots are analyzed to assess the relationships between variables.

4.3 Exploratory Data Analysis

Simple Linear Regression Models quantify relationships between response and predictor variables using a straight line. Scatter plots are used to determine whether the relationship is linear or not. The following plots were obtained using the average NDVI and Yield/Ha per year.

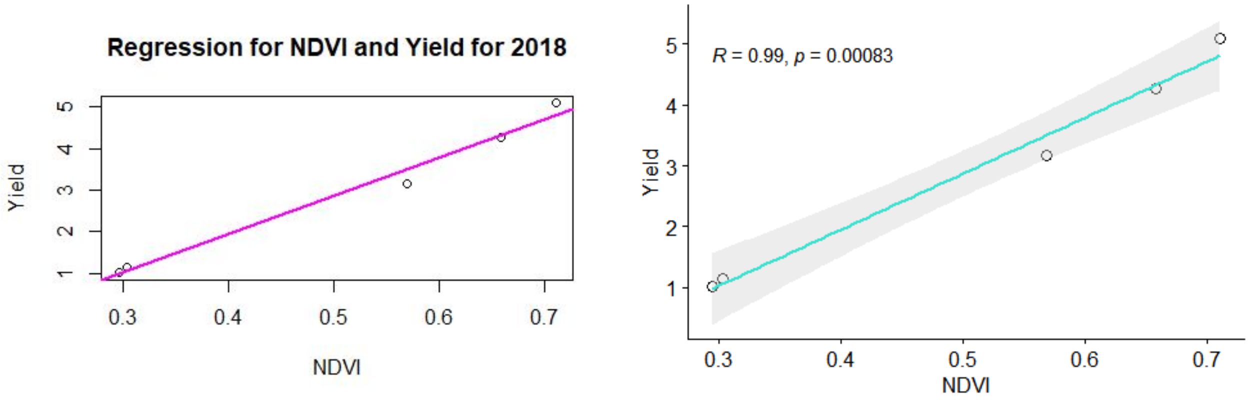

Figure 6: Regression and Correlation plots for NDVI and Yield/Ha for 2018

In the regression plot, there is an upward trend in the plot, which indicates that both the NDVI and yield are increasing. This indicates that there is a positive linear relationship between the NDVI and yield. The correlation plot gives the pearson correlation coefficient as 0.99. Therefore there is a correlation between Yield/Ha and NDVI for the year 2018.

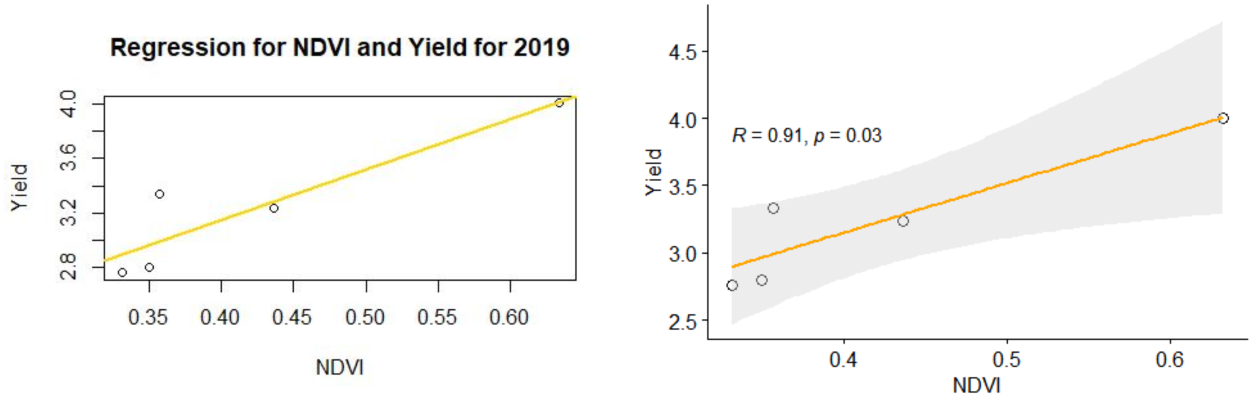

Figure 7: Regression and Correlation plots for NDVI and Yield/Ha for 2019

From the above plot, there is a rise in the regression line, which indicates that both the NDVI and yield are increasing. This indicates that there exists a positive correlation between the NDVI and yield for the year 2019. The correlation plot gives the pearson correlation coefficient as 0.91.

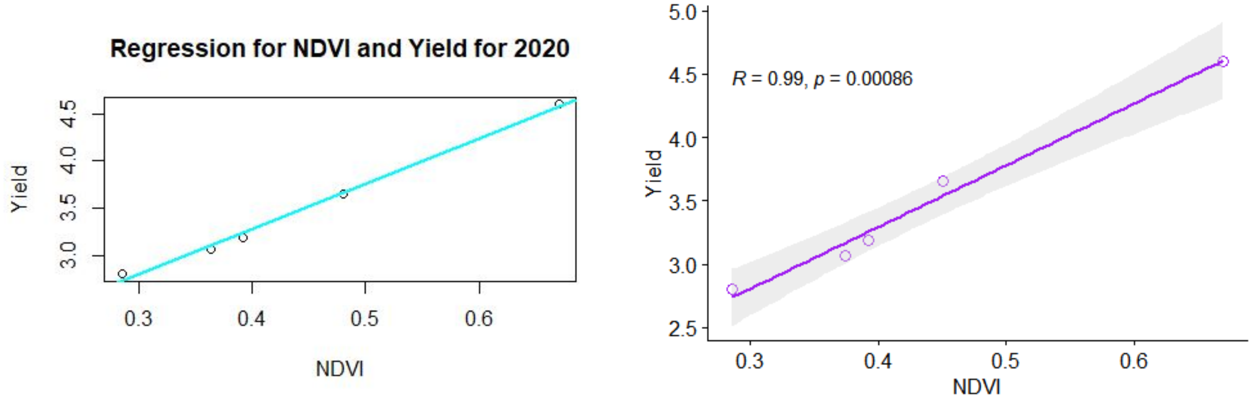

Figure 8: Regression and Correlation plots for NDVI and Yield/Ha for 2020

The regression plot shows an upward trend for both the NDVI and yield. This means that there is a positive linear relationship between NDVI and yield for the year 2020. The correlation plot gives the pearson correlation coefficient as 0.99 which indicates that there is a correlation between Yield/Ha and NDVI for the year 2020.

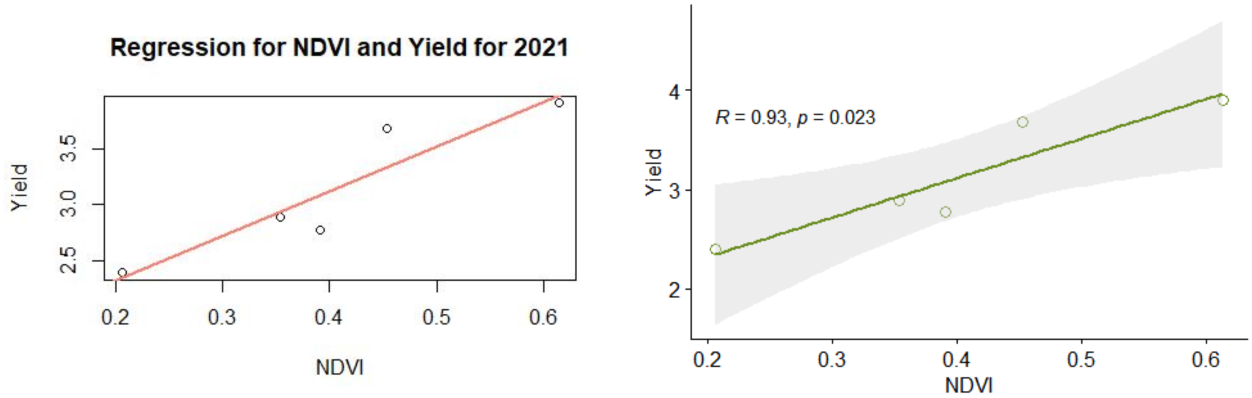

Figure 9: Regression and Correlation plots for NDVI and Yield/Ha for 2021

The plot indicates that there is a correlation between Yield/Ha and NDVI for the year 2021. The correlation plot gives the pearson correlation as 0.93. Hence a positive linear relationship between Yield/Ha and NDVI for the year 2021.

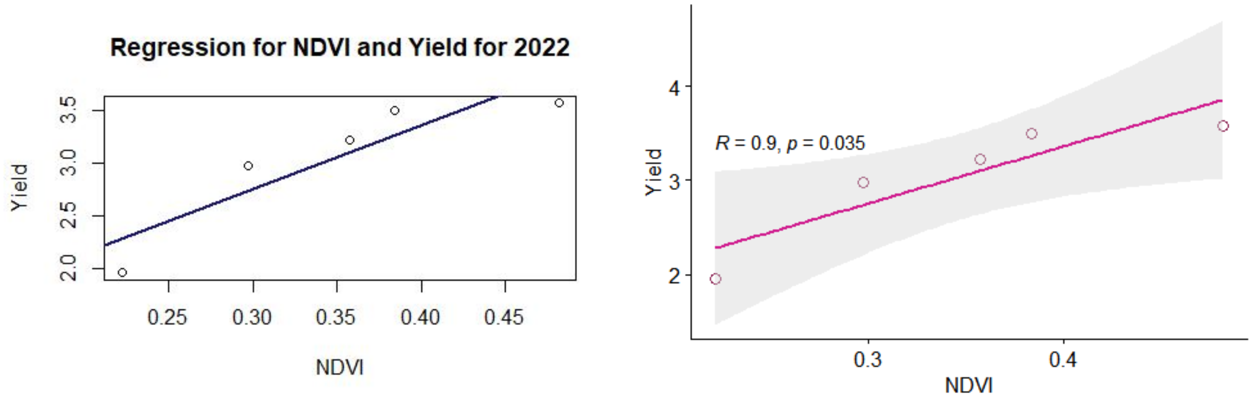

Figure 10: Regression and Correlation plots for NDVI and Yield/Ha for 022

There is a positive linear relationship between yield and NDVI for the year 2022. The correlation plot gives the pearson correlation coefficient as 0.90.

In the next section, the model parameters of a simple linear regression model are analyzed to assess the strength and significance of the relationship between NDVI and yield.

4.4 Estimation of Model Parameters and Test of Fit

A simple linear regression model is fit when there is a positive or negative linear relationship between the response and predictor variables. In this case, we have a positive linear relationship between NDVI and yield for the years 2018-2022. When the variables for each year are fitted into the linear regression model, at 0.05 level of significance, we find:

2018 Model:

```
                           Residuals:
           Min        Q1        Median        Q3        Max
        -0.33931  0.08978   0.03646    0.28173   -0.06866

                        Coefficients:
                   Estimate  Std.Error t value Pr(>|t|)
        (Intercept)  -1.7558  0.3606    -4.869   0.016551 *
        2018NDVI     9.2465  0.6721    13.758   0.000831 ***
                                ---
  Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

      Residual standard error: 0.2637 on 3 degrees of freedom
      Multiple R-squared:  0.9844,Adjusted R-squared:  0.9792
       F-statistic: 189.3 on 1 and 3 DF,  p-value: 0.0008311
```

Interpreting the Results
The y intercept is $\beta_0 = -1.7558$ and the NDVI coefficient is $\beta_1 = 9.2465$, therefore the model for the year 2018 is:

$$\text{Yield} = -1.7558 + 9.2465\text{NDVI}$$

This means that a 0.1 increase in NDVI results in a 0.92465 units increase in the yield. The **standard error** represents the average separation between the observed values and the regression line. The variation in the intercept can be up to 0.3606 and the variation in the 2018 NDVI is 0.6721. The **t-test statistic** is 13.758 and the critical value is 3.182. Therefore, we reject H_0 and conclude that there is a significant linear relationship between the yield and NDVI for the year 2018. **Residual standard error** is 0.2637 with 3 degrees of freedom. The **Multiple R-squared** and **Adjusted R-squared** are 0.9844 and 0.9792 respectively. The **F-statistic** is 189.3 with 1 and 3 degrees of freedom while the critical value is 10.128. Therefore, we reject H_0 and conclude that the fitted model is a significant fit. The **p-value** is 0.0008311 which is less than 0.05. This implies that the fitted model is statistically significant.

2019 Model:

```
                          Residuals:
          Min       1Q       Median     3Q         Max
      0.345235  -0.164705 -0.133575 -0.004225 -0.042729

                        Coefficients:
                 Estimate   Std. Error t value  Pr(>|t|)
     (Intercept) 1.6770     0.4108     4.083    0.0265 *
      2019NDVI   3.6809     0.9419     3.908    0.0298 *
                                ---
 Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

    Residual standard error: 0.2352 on 3 degrees of freedom
    Multiple R-squared:  0.8358,Adjusted R-squared:  0.7811
      F-statistic: 15.27 on 1 and 3 DF,  p-value: 0.02976
```

Interpreting the Results
The y intercept is $\beta_0 = 1.6770$ and the NDVI coefficient is $\beta_1 = 3.6809$, therefore the model for the year 2019 is:

$$\text{Yield} = 1.6770 + 3.6809\text{NDVI}$$

This means that a 0.1 increase in NDVI results in a 0.36809 units increase in the yield. The **standard error** represents the average separation between the observed values and the regression line. The variation in the intercept can be up to 0.4108 and variation in the 2019 NDVI is 0.9419. The **t-test statistic** is 3.908 and the critical value is 3.182. Therefore, we reject H_0 and conclude that there is a significant linear relationship between the yield and NDVI for the year 2019. **Residual standard error** is 0.2352 with 3 degrees of freedom. The **Multiple R-squared** and **Adjusted R-squared** are 0.8358 and 0.7811 respectively. The **F-statistic** is 15.27 with 1 and 3 degrees of freedom while the critical value is 10.128 therefore, we reject H_0 and conclude the fitted model is a significant fit. The **p-value** is 0.02976 which is less than 0.05. This implies that the fitted model is statistically significant.

2020 Model:

```
                                Residuals:
           Min       1Q       Median     3Q        Max
       -0.008572 -0.049288  0.071017  0.025969 -0.039127

                             Coefficients:
                        Estimate   Std.Error t value Pr(>|t|)
        (Intercept)    1.36007    0.08864   15.34 0.000601 ***
        2020NDVI       4.79668    0.19369   24.77 0.000144 ***
                                    ---
    Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

         Residual standard error: 0.05701 on 3 degrees of freedom
         Multiple R-squared:  0.9951,Adjusted R-squared:  0.9935
          F-statistic: 613.3 on 1 and 3 DF,  p-value: 0.0001444
```

Interpreting the Results
The y intercept is $\beta_0 = 1.36007$ and the NDVI coefficient is $\beta_1 = 4.79668$, therefore the model for the year 2020 is:

$$\text{Yield} = 1.36007 + 4.79668\text{NDVI}$$

This means that a 0.1 increase in NDVI results in a 0.479668 units increase in the yield. The **standard error** represents the average separation between the observed values and the regression line. The variation in the intercept can be up to 0.08864 and variation in the 2020 NDVI is 0.19369. The **t-test statistic** is 24.77 and the critical value is 3.182. Therefore, we reject H_0 and conclude that there is a significant linear relationship between the yield and NDVI for the year 2020. **Residual standard error** is 0.05701 with 3 degrees of freedom. The **Multiple R-squared** and **Adjusted R-squared** are 0.9951 and 0.9935 respectively. The **F-statistics** is 613.3 with 1 and 3 degrees of freedom while the critical value is 10.128.Therefore, we reject H_0 and conclude that the fitted model is a significant fit. The **p-value** is 0.0001444 which is less than 0.05. This implies that the fitted model is statistically significant.

2021 Model:

```
                          Residuals:
             Min      1Q      Median     3Q      Max
        -0.03809 -0.30599  0.05080 -0.06443  0.35771

                       Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
     (Intercept)     1.5187     0.3959   3.836   0.0312 *
     2021NDVI        3.9884     0.9322   4.278   0.0235 *
                                ---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

    Residual standard error: 0.2767 on 3 degrees of freedom
    Multiple R-squared:  0.8592,Adjusted R-squared:  0.8122
      F-statistic:  18.3 on 1 and 3 DF,  p-value: 0.02345
```

Interpreting the Results
The y intercept is $\beta_0 = 1.5187$ and the NDVI coefficient is $\beta_1 = 3.9884$, therefore the model for the year 2021 is:

$$\text{Yield} = 1.5187 + 3.9884\text{NDVI}$$

This means that a 0.1 increase in NDVI results in a 0.39884 units increase in the yield. The **standard error** represents the average separation between the observed values and the regression line. The variation in the intercept can be up to 0.3959 and the variation in the 2021 NDVI is 0.9322. The **t-test statistic** is 4.278 and the critical value is 3.182. Therefore, we reject H_0 and conclude that there is a significant linear relationship between the yield and NDVI for the year 2021. **Residual standard error** is 0.2767 with 3 degrees of freedom. The **Multiple R-squared** and **Adjusted R-squared** are 0.8592 and 0.8122 respectively. The **F-statistic** is 18.3 with 1 and 3 degrees of freedom while the critical value is 10.128 therefore, we reject H_0 and conclude that the fitted model is a significant fit. The **p-value** is 0.02345 which is less than 0.05. This implies that the fitted model is statistically significant.

2022 Model:

```
                          Residuals:
            Min      1Q    Median   3Q       Max
          0.2424  0.1222 -0.3198 -0.2824  0.2376

                       Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
      (Intercept)     0.9250     0.5978   1.547   0.2196
      2022NDVI        6.0872     1.6644   3.657   0.0353 *
                                 ---
  Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

      Residual standard error: 0.3225 on 3 degrees of freedom
      Multiple R-squared:  0.8168,Adjusted R-squared:  0.7557
        F-statistic: 13.38 on 1 and 3 DF,  p-value: 0.03531
```

Interpreting the Results

The y intercept is $\beta_0 = -0.9250$ and the NDVI coefficient is $\beta_1 = 6.0872$, therefore the model for the year 2022 is:

$$\text{Yield} = 0.9250 + 6.0872\text{NDVI}$$

This means that a 0.1 increase in NDVI results in a 0.60872 units increase in the yield. The **standard error** represents the average separation between the observed values and the regression line. The variation in the intercept can be up to 0.5978 and the variation in the 2022 NDVI is 1.6644. The **t-test statistic** is 3.657 and the critical value is 3.182. Therefore, we reject H_0 and conclude that there is a significant linear relationship between the yield and NDVI for the year 2022. **Residual standard error** is 0.3225 with 3 degrees of freedom. The **Multiple R-squared** and **Adjusted R-squared** are 0.8168 and 0.7557 respectively. The **F-statistic** is 13.38 with 1 and 3 degrees of freedom while the critical value is 10.128 therefore, we reject H_0 and conclude that the fitted model is a significant fit. The **p-value** is 0.03531 which is less than 0.05. This implies that the fitted model is statistically significant.

Prediction of New Observations

The table below are maize yield values predicted for the years 2023-2027:

```
    County      2023     2024     2025     2026     2027
   Bungoma 3.5000757 2.990854 3.665791 2.928679 2.736324
    Nakuru 1.0494881 2.964811 3.239365 3.076349 3.101603
     Nandi 0.9765212 2.897126 2.728401 2.341557 2.278136
TransNzoia 4.8091950 4.007068 4.573840 3.965963 3.855933
UasinGishu 4.3272893 3.283764 3.104230 3.325215 3.262433
```

In order to check the accuracy of the prediction of yield, the R^2 value is calculated:

```
[1] 0.7102979
```

The above output indicates that observed and predicted values are highly correlated.

4.5 Discussions

The exploratory data analysis revealed a positive linear relationship between the Normalized Difference Vegetation Index (NDVI) and Yield/Ha for the years 2018-2022. This relationship was observed through regression plots, which showed an upward trend in the data points for both NDVI and yield, indicating an increase in both variables.

Additionally, the correlation plots indicated a strong correlation between NDVI and yield for each year, with Pearson correlation coefficients ranging from 0.90 to 0.99. These coefficients suggest a high degree of correlation between NDVI and yield, further supporting the presence of a positive linear relationship.

Based on the simple linear regression models fitted for each year, we obtained the equations for predicting yield based on NDVI. These equations showed that a 0.1 increase in NDVI resulted in a corresponding increase in yield, with the magnitude of the increase varying for each year.

The statistical analysis provided further evidence for the significance of the linear relationship. The t-test statistics and associated p-values indicated that the NDVI coefficients were statistically significant for all years.

The multiple R-squared values, which represent the proportion of variance in yield explained by NDVI, were high for all years, ranging from 0.7811 to 0.9935. These values indicate a strong relationship between NDVI and yield and suggest that NDVI can explain a significant portion of the variability in yield.

Overall, the analysis consistently demonstrated a positive linear relationship between NDVI and yield for the years 2018-2022. The findings suggest that as NDVI increases, the yield per hectare tends to increase as well. These results can be valuable for understanding the relationship between vegetation health (NDVI) and agricultural productivity (yield) and can potentially aid in predicting and optimizing crop yields in the future, providing valuable insights for agricultural practices and decision-making.

Chapter 5
CONCLUSIONS

5.1 Introduction

In this chapter, we make conclusions on our findings of our study and then give necessary recommendations.

5.2 Summary

In this study the theory of Simple Linear Regression Model was used in the analysis of the relationship between NDVI and maize yield. Statistical software R was used in this study. The main objective was to demonstrate the potential of utilizing satellite imagery, specifically NDVI, to estimate maize yield. The study incorporated NDVI data from five counties and maize yield data from the same five counties, in the years 2018-2022. A significant positive linear relationship between NDVI and maize yield was observed, indicating a significant correlation between NDVI and maize yield.

5.3 Conclusions

From the results of the model, we can say that increase in NDVI causes an increase in maize yield. The results obtained from the analysis indicate that there is a direct relationship between NDVI and crop yield.

Specifically, it was observed that higher NDVI values corresponded to higher crop yields, while lower NDVI values were associated with lower crop yields. This finding suggests that an increase in NDVI, as measured by satellite imagery, can be indicative of improved vegetation health and, consequently, enhanced agricultural productivity.

Conversely, a decrease in NDVI may signify poorer vegetation conditions and lower crop yields. These results highlight the potential of utilizing NDVI data from satellite imagery as a tool to predict and monitor crop yield, offering valuable insights for agricultural management and interventions.

5.4 Recommendations

Based on the results of this study, it is recommended that use of machine learning algorithms such as Random Forest and artificial neural networks can be explored as an alternative to traditional statistical models for analyzing NDVI data and predicting crop yield. It is also recommended that the use of satellite imagery data can help us analyze how land topology affects vegetation health and growth in Kenya.

References

[1] Ali A.M., Abouelghar M., Belal A.A., Saleh N., Yones M., Selim A.I, Amin M.E.S., Elwesemy A., Kucher D.E., Maginan S. and Savin I., Crop Yield Prediction Using Multi Sensors Remote Sensing, *The Egyptian Journal of Remote Sensing and Space Sciences*, 25, 711-716, 2022.

[2] Gebremariam G. G., Tesfamariam E. H. and Gebrehiwot K., Predicting the Adoption of Drought-tolerant Maize Varieties in Kenya Using Machine Learning,*Agricultural Systems*, 171, 19-27,2019.

[3] Ines A.V.M., Das N.N., Hansen J.W. and Njoku E.G., Assimilation of Remotely Sensed Soil Moisture and Vegetation with a Crop Simulation Model for Maize Yield Prediction, *Remote Sensing of Environment*, 138, 149-164, 2013.

[4] Kasampalis D.A., Alexandridid T.K., Deva C., Challinor A., Moshou D. and Zalidis G., Contribution of Remote Sensing on Crop Models, *Journal of Imaging*, 52, 1-19, 2018.

[5] Montgomery D.C., Peck E.A. and Vining G.G., Introduction to Linear Regression Analysis, John Wiley & Sons, 2012.

[6] Roznik M., Boyd M. and Porth L., Improving Crop Yield Estimation by Applying Higher Resolution Satellite NDVI Imagery and High-resolution Cropland Masks, *Remote Sensing Applications: Society and Environment*,25, 100-693, 2022.

[7] Sayago S. and Bocco M., Crop Yield Estimation Using Satellite Images: Comparison of Linear and Non-Linear Models, *Agriscientia*, 35, 1-9, 2018.

[8] Torre D.M.G.D., Gao J. and Macinnis-Ng C., Remote Sensing-based Estimation of Rice Yields using Various Models, *Geo-Spatial Information Science*, 24:4, 580-603, 2021.

[9] Weisberg S., Applied linear regression, John Wiley & Sons, 2014.