Is the threat against the Tree of life a threat to the wallet?

- A study investigating the coconut lethal yellowing disease’s effect on the farmers’ income

Authors: Hanna Hammarbacken & Max Segerlund
Research Paper, 15c
Fall 2015
Supervisor: Inger Persson
Abstract
Coconuts are one of the most economically important plants in Mozambique, where millions of people depend on income from coconuts. The coconut lethal yellowing disease (CLYD) is a highly destructive disease that ever since the early 90’s causes coconut palms in Mozambique to stop producing fruit and leave the coconut farmers with only empty stems. This thesis examines the disease's effect on the farmers’ income, both from coconuts and other complementary sources, since the vendible harvest should decrease with the incidence of the disease. The method used is multivariate linear regression, where several income variables are used as dependent variables. Two models are created, one only interpreted for the sample of 488 observations and one aiming at generalizing the results. By this study, it cannot be confirmed that the incidence of CLYD has a significant effect on coconut farmers’ income. The results from the sample analysis do however show that the income is affected by the degree of the disease, which is an incentive for continued research in the field.

Keywords: Coconut lethal yellowing disease, CLYD, mutlivariate linear regression, income, Mozambique, coconut agriculture.
# Table of contents

1. **INTRODUCTION** ................................................................................................................................. 4

2. **DESCRIPTION OF VARIABLES** .......................................................................................................... 6

3. **METHODOLOGY** ................................................................................................................................. 8
   3.1 **MULTIVARIATE LINEAR REGRESSION ANALYSIS** ................................................................. 8
   3.2 **ASSUMPTIONS AND CONSIDERATIONS** ................................................................................. 10
   3.3 **THE MODELS** ....................................................................................................................... 13
   3.4 **ALTERNATIVE METHODS** .................................................................................................... 14
   3.5 **DATA** ........................................................................................................................................ 15

4. **RESULTS** ............................................................................................................................................. 17
   4.1 **DESCRIPTIVE STATISTICS OF THE DATA** ............................................................................. 17
   4.2 **ASSUMPTIONS FOR MODEL 1** ............................................................................................... 21
   4.3 **ASSUMPTIONS FOR MODEL 2** ............................................................................................... 24
   4.4 **RESULTS FOR MODEL 1** ......................................................................................................... 24
   4.5 **RESULTS FOR MODEL 2** ......................................................................................................... 29
   4.6 **DISCUSSION OF THE RESULTS FROM THE TWO MODELS** ....................................... 33

5. **CONCLUSION** ....................................................................................................................................... 35

6. **REFERENCES** ....................................................................................................................................... 37

APPENDIX .................................................................................................................................................. 38
1. Introduction

Coconuts and the coconut palms (*Cocos nucifera*) are often seen as a symbol for the exotic and tropic life, but the plant is also one of the most economically important plants in many low-latitude developing countries (Schnell & Priyadarshan, 2012). Around the world approximately 10 million families depend on coconuts as their main source of food and income (Eziashi & Omamor, 2010). Mozambique is one of the major producers of coconuts in both Africa and the entire world, and it’s believed that between 14 and 30 percent of the Mozambican population’s survival is either directly or indirectly dependent on the coconuts, both as an income-generating product but also as a food supplier. Especially for the small scale farmers, the great importance of high yield cannot be enough emphasised. (Vaz et al., 2013)

A threat to the coconut palms’ well-being and existence is the fast spreading and highly destructive coconut lethal yellowing disease (CLYD) (Bila et al., 2015; Vaz et al., 2013), a disease caused by the cell wall-less bacterium of phytoplasma (Harrisson et al., 2014). It is known to be transmitted by insect vectors, especially plant hoppers, and shows a number of symptomatic steps that lead up to the death of the palm tree. The first sign of an infected palm is partial or total fruit drop and blackening of new inflorescence followed by discolouration of the leaves, mostly yellowing. A third symptom is that the stem apical tissues rot and finally, within 3-6 months after the first signs of the disease, the palm crown dies, leaving the palm with only its naked stem and the farmers without any coconuts to harvest. (Bila et al., 2015)

The current outbreak in Mozambique was first detected in the country's northeast area in 1992 (Bila et al., 2015). Due to the importance of the coconut, the outbreak cause loss in production of coconuts and its derivatives, which in turn could cause serious economic damages for the small farmers, both on the households’ economies and the country’s coconut industry as a whole. To cope with the risks associated with income depending solely on coconut production, farmers often use the strategy of having multiple sources of income besides their coconut plantations, such as producing other fruits or crops, fishing, carpentry, commerce, metalworking and livestock. (Vaz et al., 2013)

There is reason to believe that some kind of relationship exists between the spread and severity of CLYD and the farmers’ income situation. This relationship is what this thesis will
address. The main aim is to investigate to what extent the palm disease CLYD has affected the coconut farmers’ household income. It is also reasonable to believe that the severity of CLYD will affect the farmers’ search for other income generating activities, why we also will evaluate how the disease has affected the farmers’ income from other activities besides coconut agriculture. These interests result in two main research questions:

- *To what extent have the farmers’ income from coconut production been affected by the coconut lethal yellowing disease?*
- *How has the extent of the coconut lethal yellowing disease affected the farmers’ incomes originating from other activities?*

These issues are researched by using data from a study conducted in Mozambique, the year of 2013. The original sample consists of 499 randomly selected households, with data from both direct observations at the plantations and from survey answers (Vaz et al., 2013). The analysis is performed with multivariate linear regression, where several income variables are used as dependent variables.

The remainder of this thesis is organized as follows. The next section contains a description of the variables included in the analysis, followed by a third section presenting our chosen method to perform the analysis, including a presentation of the assumptions and some alternative statistical methods. Section 3 also contains a more detailed presentation of the data used for this thesis. In section 4, the results and the actual analysis are presented, followed by the final and 5:th concluding section for this thesis.
2. Description of variables

In order to investigate the relationship between the households’ income and the severity of CLYD, we use several variables. Since there is reason to believe that a household’s income is depending on several sources of income, three dependent variables are used.

- **Income from Coconuts** (incomecoc), a numerical variable measuring the income from coconuts.
- **Income from other farming activities** (incomeotherfarming), a numerical variable measuring the income from other farming activities.
- **Income from other activities and pension** (incomeother), a numerical variable measuring the income from other activities than coconuts and farming i.e. other employment, pensions and offers.

To explain the variation in the dependent variables, a number of explanatory variables are used. Due to the main objective of this thesis’, to investigate how CLYD affects the households’ income, we use a variable measuring the severity of CLYD. Also, a number of complementary explanatory variables are used as controlling variables.

- **The CLYD disease incidence level** (CLYD) is a numerical variable measuring the percentage of infected plants (including dead palms) on each farm, ranging from 0 to 100.
- **Male** (Male) is a dummy variable set to 1 if the head of household is a male and set to 0 if the head of household is a female.
- **Education** (Educ) is a dummy variable denoting if the head of household is educated or not. If the head of household has any kind of education, including preschool, the variable is assigned the value 1, and 0 if the head of household isn’t educated.
- **Number of producing coconut trees** (ProdTree) is a numerical variable measuring the number of coconut trees that has been producing for the last 12 months.
- **Sold crops** (SoldCrop) is a dummy variable set to 1 if the household has sold other crops besides coconuts during the last 12 months, and set to 0 if not.
- **Medium and large animals** (ML_Animals) is a dummy variable set to 1 if the household has owned medium or large livestock animals during the last 12 months, and set to 0 if not.
• *Sold animals* (SoldAnimals) is a dummy variable set to 1 if the household has sold any animals during the last 12 months, and set to 0 if not.

• *Outside jobs* (OutsideJob) is a dummy variable set to 1 if anyone in the household has had a paid employment during the last 12 months, and set to 0 if not.

The relationship between the dependent and explanatory variables presented above are analysed using the statistical technique of multivariate linear regression. The methodology along with data and model description are discussed in the next section.
3. Methodology

The main objective in this thesis is to investigate how the coconut lethal yellowing disease (CLYD) affects the coconut farmers’ income. The second issue of interest is to investigate how CLYD affects the farmers’ income originating from complementary sources to coconuts. In order to investigate these issues, three income variables are used as dependent variables, as described in section 2. Due to the correlation between the income variables (see table A1 in appendix), a multivariate technique is preferred to use instead of several univariate models (Stevens, 2009). In contrast to the dependent variables that are measured on a numerical scale, the explanatory variables are both categorical and numerical. Due to the multiple number of dependent numerical variables and the explanatory variables being of numerical and categorical nature, multivariate linear regression analysis is an appropriate method to use, and will be addressed to perform the analysis for this thesis.

3.1 Multivariate linear regression analysis

The statistical technique of multivariate linear regression is an approach suitable for handling several dependent and explanatory variables, where each explanatory variable is assigned a weight to be able to estimate an explanatory model for the dependent variables. The weights in the model are estimated using the mathematical procedure of least squares. For the simple linear regression case, the least squares procedure finds the line that best fits the scores for each subject on the dependent and explanatory variables so that the sum of squared estimated errors of prediction is minimized. If the error term \( \hat{e}_i = y_i - \hat{y}_i \), where \( y_i \) is the actual score on the dependent variable for the \( i \)th individual and \( \hat{y}_i \) is the estimated score for the same individual, the least squares procedure chooses the estimates that makes the sum of all squared error terms \( \hat{e}_i^2 \) as small as possible (see equation 1). (Stevens, 2009)

\[
\hat{e}_1^2 + \hat{e}_2^2 + \cdots + \hat{e}_n^2 = \sum_{i=1}^{n} \hat{e}_i^2 = \text{min}
\]

Essentially, when minimizing the sum of squared prediction errors, it is equivalent to maximizing the correlation between the observed and predicted scores for \( y \) (Stevens, 2009). The higher the correlation coefficient is, the stronger the relationship and the greater the predictive accuracy between the explanatory and response variables (Hair et al, 2014).
\[ Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \varepsilon \]  

An example of a univariate regression model is presented in equation 2; \( Y \) is the dependent variable, \( \beta_0 \) the intercept, \( X_i \) the explanatory variables, \( \beta_i \) the regression coefficients and \( \varepsilon \) the error term. The model is interpreted as follows. The estimated intercept in the model (\( \beta_0 \)) is the average value of the dependent variable (\( Y \)) when all the explanatory variables (\( X_i \)) are zero, and its sign is based on the characteristics of the explanatory variables. The regression coefficients, the \( \beta_i \)'s, are separately interpreted as the estimated marginal change in the dependent variable (\( Y \)) for a unit change in the explanatory variable (\( X_i \)), holding everything else constant. If the explanatory variable (\( X_i \)) is a dummy variable, the regression coefficient (\( \beta_i \)) is interpreted in comparison with the dummy’s reference group. For instance, if we consider the variable \textit{Male} (presented in section 2), the regression coefficient (\( \beta_{\text{Male}} \)) is interpreted as the estimated change in the dependent variable (\( Y \)) if the head of the household is male in comparison to if the head of household is female, holding everything else constant. (Hair et al., 2014) The regression coefficients are thus valuable when determining a variable’s effect on the dependent variable.

When moving from the univariate regression situation to multivariate regression, the difference is that the multivariate model simultaneously includes several dependent variables in the specification. The explanatory variables are identical for both the univariate and multivariate models; only the number of response variables and the number of residuals will differ (Haase, 2011). Also, there are some statistical reasons why it is preferable to use a multivariate approach rather than several univariate models. First, using fragmented univariate tests increases the probability of at least one false rejection, i.e. it generates a greatly inflated overall type I error rate. By performing a multivariate test, there is an overall test for multiple responses generated, and the issue of multiplicity is avoided. Secondly, when using univariate tests the information from the correlation between the response variables is ignored, which can be considered to be a loss of information. When performing a multivariate test, the correlations are incorporated through the covariance matrix. Third, a multivariate test can be more powerful in the sense that small differences on several variables might become a reliable overall difference when combining them together, but not when evaluated separately. When using a univariate test, this information will not be seized. (Stevens, 2009)
When going from several univariate models to a multivariate linear regression analysis, the step is not far. The actual procedure of the multivariate linear regression analysis is simply a hypothesis testing technique stating the following hypotheses (equation 3) for each explanatory variable:

\[ H_0: \beta_{11} = \beta_{12} = \cdots = \beta_{1m} = 0 \]

\[ H_a: \text{At least one of the } \beta_{1i} \text{ is different from zero} \quad (3) \]

In words, this means that for each explanatory variable included in the model there is a hypothesis test for whether at least one of the estimated coefficients is significantly different from zero. By this, an overall effect of each variable is tested simultaneously on the number of dependent variables. (Stevens, 2009) Essentially, multivariate linear regression calculates both the univariate and multivariate solutions.

When applying multivariate linear regression to actual data, it involves two types of research problems: predictive and explanatory. For this thesis, we focus on the explanatory part, where we examine the regression coefficients (such as their magnitude, sign and statistical significance) for each explanatory variable and attempt to develop a substantive reason for the effects of the explanatory variables. However, to some extent the two research problems are intertwined and one cannot really be disjoint from the other. (Hair et al., 2014)

To assess the prediction accuracy, how good the predictions are, of each of the univariate regression models produced with multivariate linear regression, the coefficient of determination, \( R^2 \), is used. It is calculated as the squared correlation between the actual and predicted values of the dependent variable and represents the effects of the explanatory variables combined in predicting the dependent variable. Also, since the measure is constructed by the squared correlation of actual and predicted values, it is also a measure of how much of the variance in the dependent variable that is explained by the explanatory variables in the model. (Hair et al, 2014)

### 3.2 Assumptions and considerations

In order to be able to generalize the results of the analysis, there are several assumptions that need to be fulfilled (Hair et al., 2014). This section will discuss the assumptions for
multivariate linear regression and how to detect any violations of the assumptions. The following assumptions are needed to justify the results (Haase, 2011):

- The true relationship is linear; the dependent variables are linear functions of the explanatory variables and the error terms.
- $\varepsilon \sim i.i.d. \text{multivariate } N(0, \sigma)$
  - The error terms are random variables with mean zero and constant variance.
  - The error terms are independent.
  - The error terms are multivariate normally distributed.

The first assumption is that the true relationship is linear, which means that the dependent variable ($Y_i$) should be a linear function of the explanatory variables ($X_i$) and the error term ($\varepsilon_i$) (Haase, 2011). This assumption needs to be fulfilled for all dependent variables and therefore for all univariate models. For this thesis, it means that all three income variables should be a linear function of the explanatory variables that are used (see section 2). This assumption is investigated by studying partial regression plots between each dependent variable and each explanatory variable (Hair et al., 2014). The plots are used to evaluate whether there are any non-linear patterns that suggest a violation of the assumption.

The second assumption is that the error terms are random variables with mean zero and constant variance, meaning that the error terms are homoscedastic (Haase, 2011). We investigate this by examining standardized residual plots (Stevens, 2009). The values of the standardized residuals relate to the t-distribution, which simplifies the assessment of the residuals importance. We plot the standardized residuals against the predicted values which is a useful graph to determine if the assumption is fulfilled or not (Stevens, 2009; Hair et al., 2014). The residual plot is compared to the null-plot (a plot showing how it would look like if all the assumptions are met), see figure A1 in appendix. If the standardized residuals are randomly spread around zero and resemble the null-plot there are no indications of violations of the assumption. The third assumption, that the error terms are independent, is also examined by studying the standardized residual plot. The aim is to see a complete random pattern showing that there is no dependency between the residuals. (Stevens, 2009)

The fourth and last assumption is that the error terms are multivariate normally distributed. The normality assumption is needed to justify the test statistics in order to be able to
generalize the results. The multivariate normality assumption is examined graphically with a multivariate normality plot, where the plot should resemble a straight line (see figure A2 in appendix). (Stevens, 2009; Hair et al., 2014)

Besides the four assumptions to justify and validate the results of a multivariate linear regression analysis, there are some other considerations to remember. One of these concerns multicollinearity between the explanatory variables. Multicollinearity is the situation when there exists moderate to high correlation between the explanatory variables. If the correlation between the explanatory variables is high, the variables essentially explain much of the same variance in the dependent variables. This limits the degree of determination ($R^2$), since some of the variance already is captured by one of the variables. Multicollinearity is also problematic because it makes it hard to determine each explanatory variable's importance for the model. Furthermore, multicollinearity also increases the variance of the regression coefficients. (Stevens, 2002) There are essentially two ways to determine whether multicollinearity is a problem or not. The first one is to examine the pairwise correlations between the explanatory variables, where a correlation of 0.7 or higher is an indication that multicollinearity might be a problem. The second way to determine the degree of multicollinearity is by calculating the variance inflation factor (VIF). This measure is the inverse of the tolerance value (see equation 4), where tolerance is the amount of variability captured by one specific explanatory variable that is not explained by the other explanatory variables. A rule of thumb is that a VIF value below 10 is accepted, which corresponds to a tolerance value above 0.10. (Hair et al., 2014)

$$VIF = \frac{1}{Tolerance} \quad (4)$$

The sample size is the last consideration to take into account. There should be at minimum 50 but preferably at least 100 observations, and a minimum of 5 observations per explanatory variable but preferably at least 15 or 20 observations per explanatory variable. (Hair et al., 2014)

The evaluations of the assumptions for the two models used in this thesis are presented in section 4.2 and 4.3, respectively. In short, the reason for using two models is due to violations of the assumptions in the first model. By that, we are prevented from generalizing the results
to the population. Therefore, a second model is created with the aim to generalize to the population. The two models are described in the following section.

3.3 The models

In this thesis, two models are used, where the first model contains variables in their original form. The second model is a modified model where the dependent variables have been transformed. The relationships investigated by the two models are presented in equations 5 and 6 respectively, and the univariate regression models for each relationship are presented in sections 4.4 and 4.5. As mentioned in the previous section, the reason for using two models is due to violations of the assumptions for the first model, which prevents us from generalizing the results. Hence, the results for model 1 will only be interpreted for the sample.

Relationship for model 1:

\[
\text{Income}_{\text{coc}} + \text{income}_{\text{others farming}} + \text{income}_{\text{others}} = CLYD + Male + Educ + ProdTree \\
+ ML\_Animal + SoldAnimals + SoldCrop + OutsideJob
\]  

(5)

Relationship for model 2:

\[
\ln(\text{Income}_{\text{coc}}) + \ln(\text{income}_{\text{others farming}}) = CLYD + Male + Educ + ProdTree \\
+ ML\_Animal + SoldAnimals + SoldCrop
\]  

(6)

In order to be able to say something about the population, we chose to create a second model where we transform the model by using natural logarithms of the dependent variables. However, the data contains a lot of observations that have zero income for one or more dependent variables. It would be preferable to keep these observations by transforming the zeros into ones, but this still violates the assumptions. This forces us to exclude these observations from the analysis, in order to transform the variables using logs and thereby fulfilling the assumptions. We also choose to exclude the dependent variable measuring income from other activities than farming for this model. Including this variable would lead to the violation of the sample size consideration, since the variable has so many households having zero income from other activities. Ultimately, this also leads to an exclusion of the explanatory variable \textit{OutsideJob}, since this variable mainly is relevant when explaining the deleted income variable. By transforming the dependent variables and excluding the
observations with zero income, we manage to create a model that fulfils the assumptions (see section 4.3).

When using the two models, the original sample size of 499 households was reduced to 359 households for model 1, due to missing values and exclusion of some outliers (see section 3.5 for discussion). For model 2, the sample size was reduced to 184 observations due to the same reasons as for model 1, with the addition that all households with zero income from coconuts and other farming activities were excluded. To be clear, this means that results from the second model only apply to the population of farmers involved in, and with incomes from, both coconut farming and complementary farming activities. Both models are also restricted by some income limitations due to the exclusion of a few outliers, as to be discussed further in section 3.5.

3.4 Alternative methods

There are some alternative methods that might serve the purpose of this thesis instead of using the chosen technique of multivariate linear regression. One of these is Seemingly Unrelated Regression (SUR), a similar technique to multivariate linear regression, but with the difference that it allows the researcher to use different models for each dependent variable (Timm, 2002). The advantage with this approach is that overfitting the model is avoided by only choosing the relevant explanatory variables for each dependent variable (Ibid). On the other hand does this require that the researcher have a stronger understanding and theoretical reasoning for how the model should look like.

Another technique that could have been appropriate to use is Analysis of Variance, ANOVA, where it is possible to compare the means of the income from coconuts before and after the incidence of CLYD. However, this would limit us to only investigate the effects on income from coconuts and not consider income from other activities, because the dataset only contains information about income before and after CLYD for the coconut income variable. Another disadvantage with ANOVA is that the assumptions for the technique do not hold for this dataset. This makes the entire analysis a waste, due to the main purpose of a hypothesis test, which is to generalize the results (Hair et al., 2014).
Canonical Correlations is a third statistical technique that could have been used. It allows investigation of the relationship between two sets of variates. However, the method does not distinguish between the variables effects in the variate to the same degree as for multivariate linear regression, where both the multivariate solution and the univariate solutions are estimated. (Hair et al., 2014) The multivariate linear regression approach makes it easier to interpret the estimates and how each explanatory variable affects the dependent variables simultaneously and individually.

Due to the reasons stated above, multivariate linear regression seems to be the best statistical technique to perform the analysis for this thesis.

Before presenting the results from the multivariate linear regression, a review of the data used for the analysis is presented. The data section describes the target population, how the data is conducted and a motivation for excluding a few outliers.

3.5 Data

The dataset used for this thesis comes from a study conducted in Mozambique 2013 - Impact of coconut lethal yellowing disease and the beetle (oryctes) on farming systems and household income in the coastal Provinces of Zambeze and Nampula. The data from the study was collected by fieldwork in 2012 and contains 499 randomly selected coconut farmers in Mozambique. The study is divided into two parts where one part consists of direct observations of households’ plantations with the main purpose of estimating the degree and severity of CLYD and the destructive beetle Oryctes. The second part of the data consists of survey results, where the households have answered questions about farming systems, income generating activities, coconut production and details about the household members. (Vaz et al., 2013)

The target population in the study was households currently involved in coconut production in the provinces Zambeze and Nampula. Enumeration areas were used as the sample unit to divide the two provinces in homogeneous areas that corresponds to the proportion of households in the different province districts. The sampling population was generated by data from the National Institution of Statistics in Mozambique (INE) from 2007, where the data provided information about the number of districts and the number of households within each
area. The data from INE was combined with the information about the areas coconut plantations to create the sample population. The final sample population consisted of 235 enumeration areas with 26 554 households. A random sampling technique was used, where 50 enumeration areas were selected with 10 households from each enumeration area. The data selection therefore consisted of 500 households. However, one of the selected households had to be excluded due to bad weather circumstances at the time for the information to be registered. The final dataset therefore consists of 499 randomly selected households and 534 farms, since some of the households own more than one farm. (Vaz et al., 2013)

When investigating the data and creating the models, some outliers are excluded. Some of them due to concerns about the values being miss-specified, while others are deleted to get a better model fit. Households with income from coconuts above 40 000 metical/year, or income from other farming activities above 60 000 metical/year, or income from other activities above 100 000 metical/year are excluded. All observations deleted are approximately at least 10 000 metical above the observation with the highest income kept for each income variable, some of the extreme values are more than 3 times the highest income. The implication of these restrictions is that the models cannot be used to explain household earnings above these levels. In total, 11 observations have been deleted leaving the final sample used in this thesis to consist of 488 households.

The data just described are evaluated in the following section with some descriptive statistics and analysed using the approach of multivariate linear regression.
4. Results

4.1 Descriptive statistics of the data

We start the analysis by looking at some descriptive statistics of some of the variables for the sample of 488 observations. They are used to get a greater understanding for how each variable is distributed between the households in the sample. We will not present descriptive statistics for all variables included in the analysis, but only for those variables where such figures can serve an explanatory purpose.

In order to describe the income distribution of the households in a contributing way, we chose to divide the income variables in five categories. Above, in figure 4.1.1, the distribution for income from coconuts is presented. As the bars show, the majority of the sample is within the lower income brackets, including households stating that they don’t have any income from coconuts. The majority of the households are thus low-income earners regarding income from coconuts, which will be considered when analysing the results.
In figure 4.1.2, the distribution of income from other farming activities is presented. Once again, five categories representing different income ranges are used to divide the population into groups for clarifying purposes. As for the variable measuring income from coconuts, the majority of the observations are within the lower income levels and it is quite rare to earn more than 5000 metical/year from other farming activities.

Figure 4.1.2: Distribution of income from other farming activities, measured in metical/year

Figure 4.1.3: Distribution of income from other activities, measured in metical/year
The last income variable, measuring income from other activities, is presented in figure 4.1.3. A big majority of the households don’t have any income from other activities, shown by the high bar representing the zero-income earners. For those who state they do have income generated by other activities, it is quite evenly distributed for each of the brackets.

<table>
<thead>
<tr>
<th>Variable</th>
<th>N</th>
<th>Mean</th>
<th>Std dev</th>
<th>Minimum</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td>IncomeCoc</td>
<td>461</td>
<td>2203.52</td>
<td>5333.67</td>
<td>0</td>
<td>40000.00</td>
</tr>
<tr>
<td>IncomeOtherFarming</td>
<td>481</td>
<td>2230.21</td>
<td>6391.05</td>
<td>0</td>
<td>55625.00</td>
</tr>
<tr>
<td>IncomeOther</td>
<td>439</td>
<td>4248.00</td>
<td>12750.72</td>
<td>0</td>
<td>96000.00</td>
</tr>
</tbody>
</table>

Table 4.1.1: Means, standard deviations and variable range for incomes from coconuts, other farming and other activities

In table 4.1.1, are the mean values for the three income variables presented. Income from other activities has the highest mean value of 4248 metical/year, while the mean value for income from coconuts and from other farming activities is 2204 and 2230 metical per year, respectively. It seems that it is most profitable to be engaged in other activities rather than farming of any kind, and that there for this sample is a large difference in income for the three different occupational categories.

The explanatory variable of greatest interest in this thesis is CLYD. In figure 4.1.4, the distribution of the households’ incidence level for the disease is presented. 41 percent of the
observations do not have any palms infected by CLYD, and 34 percent of the households are observed to have an incidence level of CLYD above 0 and up to 25 percent. Thus, 75 % of the households in this sample have either been spared or have a moderate level of CLYD, and only 25 percent of the observations have been observed to have an incidence level of the disease above 25 percent. However, the figure also shows that 59 % of the households in the sample actually are infected by the disease, which further strengthens our incentive to perform this study.

![Diagram 4.1.5: Distribution of number of producing trees](image)

A main explanatory variable to determine the income from coconuts is the number of producing trees. The distribution of the variable is presented in figure 4.1.5, where the number of trees has been grouped into five categories. As one can assume from previous figure 4.1.1, presenting the distribution for income from coconuts, the majority of the farmers are within the first two categories, indicating that the sample mainly consists of smaller scale farmers with a smaller number of producing trees.
Lastly, we present the education level of the head of household in figure 4.1.6. The majority of the head of households, 54 %, have at least gone to preschool, while 40 % of the sample does not have any kind of education. 6 % of the sample has not stated whether they have got any level of education. Considering this, it is quite a large part of the households involved in coconut farming that have an uneducated head of household, which could be an interesting feature to address later in the thesis when interpreting the results.

The diagrams above present a broader picture of the distribution of the most essential variables included in the models. The two following sections concern the assumptions for our chosen method of multivariate linear regression, for the two models.

4.2 Assumptions for model 1
This section will evaluate the first model with reference to the assumptions and look for violations of them. The first assumption for multivariate linear regression concerns, as presented in section 3.2, linearity of the relationship between the variables. This is evaluated by investigating partial regression plots on the dependent variables for each explanatory variable. The plots for model 1 are presented in figures A3-A5 in appendix. There are no clear deviations from linearity or any non-linear patterns observable in the plots. Thus, the first assumption can be considered to be fulfilled.
To check the second and third assumptions of multivariate linear regression, we investigate standardized residual plots to see if there are any violations concerning the error terms, plots are presented in figure 4.2.1. They clearly indicate a violation regarding the error terms’ constant variance. The variance increases with the predicted income for all the income variables, which is a quite commonly observed pattern for income variables. The implication of this violation is that the lower values of income will be quite accurately predicted, whilst the higher income levels will be associated with more contingency. The standard errors will also be inflated. (Stevens, 2009) However, the third assumption about independence of the error terms seems to be fulfilled, as we don’t see any signs of dependency in the figure.
The last assumption concerns the multivariate normal distribution of the error terms, which is diagnosed with a multivariate normality plot. For the assumption to be fulfilled, the plot should resemble a straight line, an example shown in figure A2 in appendix. The line visible in figure 4.2.2 does clearly not resemble a straight line, and the residuals cannot be stated to be multivariate normally distributed. By this, the assumption is clearly violated in this case.

As mentioned in section 3.2, the sample size and multicollinearity between the explanatory variables are two important considerations. The recommendation for sample size is to preferably be at least 100 observations in total and 20 observations per explanatory variable. In the first estimated model the sample size consists of 359 observations, which means that we roughly have 45 observations per explanatory variable. By this, the sample size is well over the preferable limit. Concerning multicollinearity, table A2 in appendix shows that all of the tolerance values for the variables are well above the recommended cut off limit of 0.10, and multicollinearity is therefore not considered to be a problem.
4.3 Assumptions for model 2

When continuing to the second model, where the two income variables measuring coconut production and other farming have been logged, all of the assumptions for multivariate linear regression are fulfilled; there are no obvious signs of deviations from linearity looking at the partial plots (see figure A6 and A7 in appendix), the error terms show no sign of being correlated to each other and there is no indication of heteroscedasticity (see figure A8 in appendix). Also, the error terms seem to follow a multivariate normal distribution (see figure A9 in appendix). Lastly, the tolerance values are all above the recommended cut off value of 0.10, as can be seen in table A3 in appendix, which indicates that there are no complications regarding multicollinearity.

For this model, 184 households are used, which is a sample size within the preferable range of number of observations included. By this final consideration, the second model is clear for statistical inference, and the results can be interpreted for the entire population, in contrast to model 1 that only is interpretable for the sample.

4.4 Results for model 1

When performing the analysis for the first model, we need to bear in mind that some of the assumptions for multivariate linear regression are violated. The implication of the violations is that we cannot interpret any of the results for the different hypothesis tests, not for the multivariate nor the univariate results, and that it will predict lower income levels better than higher levels, as discussed in section 3.2. Also, this causes the standard errors to be inflated. However, they will still be interpreted, but with caution. This somewhat limits the advantages of using a multivariate linear regression and we can only interpret the univariate results for each income variable, due to that the multivariate solution is a hypothesis test. The violations of the assumptions also prohibit us from generalizing the results and the analysis for model 1 are therefore only interpreted and valid for this sample of 359 observations.

The first univariate model is estimated for income from coconuts, see equation 7.

\[
\text{Income}_{cocos} = \beta_0 + \beta_1 CLYD + \beta_2 Male + \beta_3 Educ \\
+ \beta_4 ProdTree + \beta_5 ML_{Animal} \\
+ \beta_6 SoldAnimals + \beta_7 SoldCrop \\
+ \beta_8 OutsideJob + \epsilon
\]
The results of the regression are presented in table 4.4.1. Most of the signs of the estimates are as expected, based on previous knowledge and the thesis objective. The sign of education is however unexpected, but logical when giving it further consideration. With a negative estimate of -1095, it means that if the head of household has any type of education, a household on average earns 1095 metical/year less from coconuts than a household with a head of household without education, ceteris paribus. It is well known that education in general increases the level of income, but when considering the fact that many of the coconut farmers in Mozambique are uneducated, as seen previously in figure 4.1.7, and that a head of household with some kind of education might have other sources of income (besides coconut production) from an employment or activity that have higher educational requirements as primary source of income, the negative sign of the estimate feels reasonable. Nevertheless, the standard deviation for the variable is quite high, 542 metical, indicating that the prediction accuracy is quite low and that we should be careful when interpreting the results.

In this thesis we are particularly interested in the variable CLYD. The estimate for the variable shows a negative value of -7.44, with a standard error of 12.41. This means that a household’s income decreases with on average 7.44 metical per year for every percentage point’s increase in CLYD incidence, ceteris paribus. In relation to the mean income from coconut production (2204 metical/year, see table 4.1.4), the effect from CLYD is perceptual. Furthermore, if we consider that the data is heteroscedastic with increasing variance for larger
values of income from coconuts, the model is better at predicting income of lower values. For the small scale coconut farmers with low levels of income (see income distribution in figure 4.1.1), an income decrease of a 7.44 metical for every percentage increase in CLYD have a more severe affect on their total income from coconuts compared to larger scale farmers, and CLYD can thus be said to have a quite noticeable impact on the coconut farmers’ income.

Another interesting parameter estimate is for the variable Male; if the head of household is male, the household on average increases its income from coconuts with 334 metical/year compared to if the head of household is female, ceteris paribus. The number of producing trees is also a variable worth some attention, which is interpreted as for every unit’s increase in the number of producing trees, the income from coconuts increases with on average 68.24 metical/year, ceteris paribus. In comparison with the other estimates, the standard deviation for number of producing trees is relatively low, which indicates that the variable is quite accurate at predicting the income from coconuts. The remaining variables included in the model (ML_animals, SoldAnimals, SoldCrop and OutsideJob) all show positive estimates, indicating that income from coconuts will increase if the household has any of these features. They also have rather substantial standard deviations, which tells us that the estimates are associated with low prediction accuracy.

The goodness of fit for the model is assessed by the value of the coefficient of determination (R$^2$). The predicted model for income from coconuts shows an R$^2$ of 0.1497 in table 4.4.1. The predicted model therefore explains approximately 15 percent of the variation in income from coconuts in the sample. It is far from a perfect explanation degree, but it still explains a substantial proportion of the variation in the dependent variable.

Moving on to the second univariate model, which estimates the income from other farming activities, see equation 8.

\[
\text{Incomeotherfarming} = \beta_0 + \beta_1 \text{CLYD} + \beta_2 \text{Male} + \beta_3 \text{Educ} + \beta_4 \text{ProdTree} + \beta_5 \text{ML_Animal} + \beta_6 \text{SoldAnimals} + \beta_7 \text{SoldCrop} + \beta_8 \text{OutsideJob} + \epsilon
\]
The results from the regression are presented in table 4.4.2. The signs are a bit different compared to the first univariate model; some are as theoretically expected and some are not. If the gender of head of household is male, it has a negative effect on the household income in comparison to if a female were to be head of household, ceteris paribus. However, the standard error for the estimate (812 metical/year) is very high in comparison to the estimated value (-151 metical/year), indicating that the variable’s prediction accuracy is quite low. The variable measuring education has a slightly positive effect, unlike the first univariate model for income from coconuts. Even so, the standard error is very large in comparison to the estimated value, and the prediction accuracy is not eligible.

The sign for the CLYD parameter is again negative; the estimated value is -6.45 with a standard error of 12.74. This means that for every percentage point’s increase in CLYD the households’ income from other farming activities decrease on average by 6.45 metical per year, ceteris paribus. CLYD can thus be said to have a negative effect on income from other farming activities, in contrast to our initial beliefs, even though there isn’t a direct relationship between the disease and other farming activities.

*SoldAnimals* is the variable that has the strongest effect on income from other farming activities, households that have sold livestock are on average earning 2916 metical/year more than households that haven’t sold livestock, holding everything else constant. If the household

<table>
<thead>
<tr>
<th>Variable</th>
<th>Estimate</th>
<th>Standard error</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>713.11</td>
<td>868.17</td>
</tr>
<tr>
<td>CLYD</td>
<td>-6.45</td>
<td>12.74</td>
</tr>
<tr>
<td>Male</td>
<td>-151.26</td>
<td>811.60</td>
</tr>
<tr>
<td>Educ</td>
<td>18.51</td>
<td>555.95</td>
</tr>
<tr>
<td>ProdTree</td>
<td>-6.20</td>
<td>9.58</td>
</tr>
<tr>
<td>ML_Animal</td>
<td>1061.35</td>
<td>759.35</td>
</tr>
<tr>
<td>SoldAnimals</td>
<td>2916.03</td>
<td>631.30</td>
</tr>
<tr>
<td>SoldCrop</td>
<td>1706.43</td>
<td>535.12</td>
</tr>
<tr>
<td>OutsideJob</td>
<td>-626.65</td>
<td>719.97</td>
</tr>
</tbody>
</table>

Table 4.4.2: Parameter estimates and standard errors for univariate model for income from other farming activities
has owned medium and/or large livestock animals or has sold crops also affects the income to a large extent, an expected average increase of 1061 metical/year and 1706 metical/year, respectively. Something that has a negative effect on the income from other farming activities is an increasing number of producing coconut trees. For a one unit (tree) increase of producing trees, the income from other farming activities decreases with on average 6.20 metical/year, ceteris paribus. This is a quite intuitive result, since that the larger a coconut plantation is, the more time consuming it is, and less time is left to dealing with other farming activities. Also, the larger a coconut plantation is, the less farming area will be left for other farming activities. Lastly, if any household member has a job outside the farm, the income from other farming activities decreases with on average 627 metical/year, compared to if no one from the household has had an outside job or other activity, ceteris paribus. Since other activities than farming generate a higher income on average, as table 4.1.1 show, this result is quite reasonable. Also, the model shows an $R^2$ of 0.09, see table 4.4.2, meaning that it explains approximately 10 percent of the variation in income from other farming activities.

The third univariate model estimated is for income from other sources, more specifically income from employment, pensions and offers. The predicted model is shown in equation 9.

\[
\text{Income}_{\text{other}} = \beta_0 + \beta_1 \text{CLYD} + \beta_2 \text{Male} + \beta_3 \text{Educ} \\
+ \beta_4 \text{ProdTree} + \beta_5 \text{ML\_Animal} \\
+ \beta_6 \text{SoldAnimals} + \beta_7 \text{SoldCrop} \\
+ \beta_8 \text{OutsideJob} + \varepsilon
\] (9)

<table>
<thead>
<tr>
<th>Variable</th>
<th>Estimate</th>
<th>Standard error</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>-2179.84</td>
<td>1808.21</td>
</tr>
<tr>
<td>CLYD</td>
<td>41.42</td>
<td>26.53</td>
</tr>
<tr>
<td>Male</td>
<td>1111.43</td>
<td>1690.39</td>
</tr>
<tr>
<td>Educ</td>
<td>1325.09</td>
<td>1157.93</td>
</tr>
<tr>
<td>ProdTree</td>
<td>2.93</td>
<td>19.95</td>
</tr>
<tr>
<td>ML_Animal</td>
<td>-599.05</td>
<td>1581.55</td>
</tr>
<tr>
<td>SoldAnimals</td>
<td>1244.44</td>
<td>1314.85</td>
</tr>
<tr>
<td>SoldCrop</td>
<td>699.06</td>
<td>1114.54</td>
</tr>
<tr>
<td>OutsideJob</td>
<td>21432.16</td>
<td>1499.54</td>
</tr>
</tbody>
</table>

$R^2=0.3824$

Table 4.4.3: Parameter estimates and standard errors for univariate model for income from other activities
The estimates from the regression are to be seen in table 4.4.3. As we assumed previously, education in this model have the opposite sign in comparison to the first univariate model: a positive estimate of 1325. This means that households with an educated head of household on average earn 1325 metical more per year from other activities than households with an uneducated head of household, ceteris paribus. The standard error for the estimate is quite large (1158 metical/year), but it is still reasonable to state that education has a positive effect on income from activities besides farming activities.

CLYD has a positive estimate of 41.42 with a standard error of 26.53, indicating that farmers that are affected by CLYD on average increase their income from other activities with 41.42 metical/year for every percentage point’s increase in CLYD, ceteris paribus. This could indicate that farmers affected by CLYD increase their income from employment (or pensions/offers) in order to cope with decreasing income from coconuts, reasoning in accordance with our initial beliefs.

OutsideJob is undoubtedly the variable with the greatest impact on income from other activities. The estimated value is 21432 with a relatively moderate standard error of 1500. This means that if any member in the household has a job outside the farming activities, the household’s income on average increases with 21432 metical/year in comparison to the situation where none of the household members have a job besides the farm, ceteris paribus. The remaining parameters have positive estimates, except for the variable ML_animal, which has a negative estimate. This could be an indication of that other activities, such as employment and pensions, mostly are a complementary source to other farming activities, since the estimates for the variables that mainly relate to the other income categories are almost entirely positive. They all also have large standard deviations, indicating that the prediction accuracy is not very good and there is great uncertainty in the estimates.

The model for income from other activities shows the highest degree of determination of the three univariate models. With an $R^2$ of 0.38 it means that the explanatory power of this model is approximately 38 percent, which is a quite good result.

The results for model 1 can be summarized to be both surprising and as expected. CLYD shows expected effects on both income from coconuts and income from other activities, but
has an opposite sign compared to what was expected regarding income from other farming activities. The following section continue with presenting and interpreting the results for the second model.

4.5 Results for model 2

Besides the variable measuring whether any of the households’ members has had an employment outside the farming activities, the same explanatory variables as in model 1 are included in the second model. Considering the dependent side of the model, the variable measuring income from other activities has been dropped (see section 3). In order to fulfil the assumptions of multivariate linear regression, it is necessary to exclude all the households with zero income from the two dependent variables and transform the variables using logarithms (see section 3 and 4.3). The implication of these changes is that the results from model 2 only are valid for households with incomes from both coconuts and other farming activities. The aim of the second model is to be able to generalize the results to the population, and we are therefore interested in analysing the significance of the variables, both the overall and individual significance. The standard significance level of 5 % is used, since there is nothing to suggest that the results would have dire consequences if a type 1 error were to be made.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Wilks’ Lambda, Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>CLYD</td>
<td>0.4328</td>
</tr>
<tr>
<td>Male</td>
<td>0.8974</td>
</tr>
<tr>
<td>Educ</td>
<td>0.7146</td>
</tr>
<tr>
<td>ProdTree</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>ML_Animal</td>
<td>0.0888</td>
</tr>
<tr>
<td>SoldAnimals</td>
<td>0.2365</td>
</tr>
<tr>
<td>SoldCrop</td>
<td>0.0119</td>
</tr>
</tbody>
</table>

Table 4.5.1: P-values for Wilks’ Lambda testing overall effect of variables

When looking at the overall effects of the explanatory variables, presented in table 4.5.1, only two of them show significant p-values at the 5 % level: SoldCrop with p-value 0.0119 and ProdTree with p-value <0.0001. Interpreting this, we can say that the coefficients for these two variables are significantly different from zero; a farmer selling other crops besides
coconuts, and the number of producing coconut trees have significant effects on income from coconuts and from other farming activities simultaneously. The variable of main interest for this thesis, CLYD, does not have a significant overall effect. With a p-value of 0.4685 it is not even close to being significant, and it cannot be stated that the parameter value for CLYD is significantly different from zero.

Moving on to the univariate results, the first estimated model concerns the logarithm of income from coconuts, presented in equation 10.

\[
\ln (Income_{coc}) = \beta_0 + \beta_1 CLYD + \beta_2 Male + \beta_3 Educ + \beta_4 ProdTree + \beta_5 ML_{Animal} + \beta_6 SoldAnimals + \beta_7 SoldCrop + \varepsilon
\]  

\[(10)\]

The results from the regression are presented in table 4.5.2. The univariate model for the variable measuring the log of coconut income, is highly significant with a p-value of 0.0001. Looking closer at the separate effects of the explanatory variables, only the variable measuring number of producing trees is significant at a 5 % level, leaving the rest of the variables to be insignificant. Essentially, this means that we only can say that the number of producing coconut palms significantly contributes to explaining a farmer's income originating from coconut production. This is a very reasonable result, since there is a direct linkage
between incomes from coconuts, i.e. how much money a farmer gets from selling the
products, and how many trees that give coconuts, i.e. how much products there is to sell.
Interpreting the parameter estimate for the variable, the expected relative change in coconut
income is 0.02 (2 %) when increasing the number of producing trees with one tree, ceteris
paribus. The standard error, 0.0036, is quite low, which indicates good prediction accuracy.

The variable of main interest, CLYD, shows a non-significant p-value of 0.5403 stating that
this sample cannot conclude the effect of the coconut lethal yellowing disease to be
significantly different from zero. This result is in contrast to the theoretical reasoning behind
this thesis, that the presence and severity of CLYD should have an effect on the income of
coconut farmers. The overall degree of determination for the model, the $R^2$, is 15.39 percent,
an acceptable level of explanatory power.

The second estimated univariate model is for the logarithm of income from other farming
activities, presented in equation 11.

$$\ln(\text{incomeotherfarming}) = \beta_0 + \beta_1 \text{CLYD} + \beta_2 \text{Male} + \beta_3 \text{Educ} + \beta_4 \text{ProdTree} + \beta_5 \text{ML\_Animal} + \beta_6 \text{SoldAnimals} + \beta_7 \text{SoldCrop} + \epsilon$$  (11)

<table>
<thead>
<tr>
<th>Variable</th>
<th>Estimate</th>
<th>Standard error</th>
<th>Pr &gt; t</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>5.9283</td>
<td>0.4829</td>
<td>&lt; 0.0001</td>
</tr>
<tr>
<td>CLYD</td>
<td>0.0068</td>
<td>0.0065</td>
<td>0.2986</td>
</tr>
<tr>
<td>Male</td>
<td>0.0547</td>
<td>0.4176</td>
<td>0.8959</td>
</tr>
<tr>
<td>Educ</td>
<td>0.0126</td>
<td>0.2707</td>
<td>0.9630</td>
</tr>
<tr>
<td>ProdTree</td>
<td>-0.0019</td>
<td>0.0035</td>
<td>0.5795</td>
</tr>
<tr>
<td>ML_Animal</td>
<td>0.8053</td>
<td>0.3624</td>
<td>0.0276</td>
</tr>
<tr>
<td>SoldAnimals</td>
<td>0.4680</td>
<td>0.2842</td>
<td>0.1014</td>
</tr>
<tr>
<td>SoldCrop</td>
<td>0.7466</td>
<td>0.2843</td>
<td>0.0094</td>
</tr>
</tbody>
</table>

Table 4.5.3: Parameter estimates, standard errors and p-value for students t-test for logged income from other farming activities
Looking at the second univariate model for the logged variable measuring income from other farming activities, it shows a p-value of 0.0354 telling us that it is a significant model at the 5% level, see the results in table 4.5.3. The variable measuring whether a farmer has sold other crops besides coconuts is significant at the 5% level with a p-value of 0.0094. The parameter estimate is 0.7466; the expected relative change in income from other farming activities is calculated to be $1.1098 \left( e^{0.7466} - 1 \right)$, 111%, in comparison to if the household hasn’t sold crops, ceteris paribus (Halvorsen & Palmquist, 1980). This is from a theoretical point of view a reasonable result; to retrieve income from other farming activities, it is quite necessary to actually sell the harvest from the production. The only other variable that is significant at a 5% level is if the farmers have owned medium or large livestock animals. That variable shows a p-value of 0.0276, which also is reasonable from a theoretical point of view because medium and large livestock animals are related to other farming activities. The estimated value is 0.8053, and the expected relative change in income from other farming activities is calculated to be $1.2374 \left( e^{0.8053} - 1 \right)$, 124%, in comparison to if the household has not owned medium or large livestock animals, ceteris paribus (Ibid). The overall fit for the model, $R^2$, is 0.0809.

The results for model 2 shows the intuitive result that the number of producing trees and whether a household sell their crops from other farming activities or not are overall significant variables when estimating a households income. The model does not succeed in determining that the effect from the incidence of CLYD is different from zero. Before concluding this thesis, a short discussion of the results from the two models is presented in the following section.

### 4.6 Discussion of the results from the two models

Summarizing the results of the analysis, they show both surprising and expected elements. Both models confirm already theoretically generic relationships: that the number of producing trees has an impact on how much income is generated by coconut production, and that a household that has sold crops in the last 12 months besides coconuts have a higher income from other types of farming compared to households that haven’t sold other crops. The unexpected elements concern the main variable for this thesis, CLYD. We are not able to confirm the hypothesis that the incidence of coconut lethal yellowing disease does have a significant effect on any type of income concerning coconut farmers with complementary
income from other farming activities. This could be due to the fact that the model only includes households that actually have some kind of income from any type of farming activity, including both coconut farming and other kinds of farming. Since all households with zero income have been excluded, the households that possibly have lost all their income from coconuts due to the disease are not included. Important information might be lost with these observations when trying to state a relationship between the farmers’ income and CLYD. The first model, focusing on the sample only and including the households with zero income, did show differences in income depending on different CLYD incidence levels, which could be an incentive to perform future research with new data.

The fact that there are large differences in the dataset concerning income levels for all three sources, could also be a reason for the non-significant effect of CLYD. The model might be inefficient when taking into account the size of the farm, and by that the size of the income, in relation to the severity of CLYD. Preforming a similar study focusing on more homogeneous groups might be an interesting angle to apply in the future. Another reason to ponder is the fact that the dataset used for this study might have omitted variables important when determining a household’s income, which could cloud the effect of the disease. Also, to get a wider understanding and better grounds for the economic situation of the coconut farmers, it could be rendering to also look at the costs of the households. For instance, the level of CLYD might increase the costs associated with coconut farming, and by that have a more severe impact on the households’ economic situation. To collect more detailed data describing the situation both before and after the disease would be another interesting aspect, to once again get a more accurate description of the impact of the disease. This would allow the researcher to perform and interpret a MANOVA to compare the mean income before and after CLYD, if the assumptions are non-violated for the method.

Lastly, the CLYD incidence levels for a majority of the farms included in this sample (75 %) have an incidence level of 25 % or less. This could mean that, depending on how the farmers decide to approach the disease, the spread and extent of CLYD could become a much more severe problem in the future, which could cause larger damages on the farmers’ incomes. Together with the elucidated scenarios above, there are clear motifs for new studies and future inquires.
5. Conclusion

The coconut lethal yellowing disease (CLYD) is a destructive and infectious disease that, ever since the early 90’s, haunts the Mozambican farmers involved in coconut production. The disease causes the palms to drop all their fruit and leaves, and leaves the stem bleak. If the disease infects a plantation, there is a risk for the farmer to suffer a loss in income from selling coconuts and its associated derivatives. The main objective with this thesis have been to investigate how this disease affects the farmers’ income from coconut production, and also further examine the effect of the disease regarding the farmers’ incomes from other activities, which could be affected due to the farmers need to cope with reduced incomes from coconuts.

We have chosen to use the statistical technique of multivariate linear regression to perform the analysis of interest.

This study is conducted using two models, one model regarding only the specific sample for this study, due to that some of the assumptions for the chosen technique were violated, and one model aiming at generalizing the results to the entire population of coconut farmers in Zambeze and Nampula in Mozambique, with some income restrictions as discussed in section 3.5. The first model shows that the increasing degree of CLYD has a negative effect on income from coconuts on average, holding everything else constant. If the head of household is male, the income from coconut production increases in comparison to if the head of household is female, holding everything else constant. The variable measuring CLYD also has a negative impact on income from other farming activities, and indicate increasing levels of income from other activities with increasing levels of CLYD. These results are however not generalizable and thus only valid for interpretation within this specific sample, due to violations of some of the assumptions.

The second model has the log of income from coconuts and the log of income from other farming activities as dependent variables, and the results are only valid for households with incomes from both coconuts and other farming activities within the income range discussed in section 3.5. Two of the explanatory variables have an overall significant effect on the logged income variables at a 5 % level; the number of producing trees and if the household has sold crops or not. Looking at the univariate cases, the number of producing trees is significant for the logged income from coconuts-variable and it has a positive effect on income. For the univariate model for the logged income from other farming, the dummy variable denoting
whether the household has sold crops or not is significantly different from zero, together with the variable indicating if the household has owned medium or large animals for the last 12 months or not. Both variables have a positive effect on income from other farming activities. The variable measuring the incidence level of CLYD is not even close to being significant, not on the overall level nor for the univariate cases. By that, it cannot be stated that the incidence level of CLYD has a significant effect on a farmer’s income from coconuts or from other farming activities, and our beliefs when initiating this thesis, that the incidence level of CLYD would have an impact on income from coconuts and from other farming activities, cannot be confirmed. However, the results from the sample analysis (model 1) show that CLYD has a negative impact on income from coconuts. The variable’s positive estimate for income from other activities could also be an indication that CLYD has affected the farmers’ overall income situation, forcing them to supplement their incomes with other activities due to the decreasing levels of income from coconut production.

Finally, the aim of this thesis was to investigate how CLYD has affected the farmers’ income from coconuts and other income activities. The results show some indications of CLYD having an effect on the income from coconuts in a negative manner, and income from other activities in a positive way. However, the estimates for CLYD in the model with non-violated assumptions are non-significant, prohibiting us from generalizing the results to the whole population and significantly establishing something about this relationship. However, due to the results from model 1, we still believe that further research is necessary to determine if and, if so, what kind of relationship the coconut lethal yellowing disease has with coconut farmers’ income.
6. References


Vaz, K. et al. (2013), Impact of coconut lethal yellowing disease and the beetle (oryctes) on farming systems and household income in the coastal Provinces of Zambeze and Nampula, Maputo: Verde Azul Consult Ld
Appendix

<table>
<thead>
<tr>
<th>Correlations</th>
<th>IncomeCoc</th>
<th>IncomeOtherFarming</th>
<th>IncomeOther</th>
</tr>
</thead>
<tbody>
<tr>
<td>IncomeCoc</td>
<td>1</td>
<td>0.0231</td>
<td>0.0239</td>
</tr>
<tr>
<td>IncomeOtherFarming</td>
<td>0.0231</td>
<td>1</td>
<td>0.0097</td>
</tr>
<tr>
<td>IncomeOther</td>
<td>0.0239</td>
<td>0.0097</td>
<td>1</td>
</tr>
</tbody>
</table>

*Table A1: Correlations between income variables*

*Figure A1: The null plot of residuals*

*Figure A2: Example of a multivariate normality plot of residuals*
Figure A3: Partial Regression Plots for incomecoc for model 1

Figure A4: Partial Regression Plots for incomeotherfarming for model 1

Figure A5: Partial Regression Plots for incomeother for model 1
Table A2: Tolerance values for model 1

<table>
<thead>
<tr>
<th>Variable</th>
<th>Tolerance</th>
</tr>
</thead>
<tbody>
<tr>
<td>CLYD</td>
<td>0.9583</td>
</tr>
<tr>
<td>Male</td>
<td>0.9175</td>
</tr>
<tr>
<td>Educ</td>
<td>0.9379</td>
</tr>
<tr>
<td>ProdTree</td>
<td>0.9669</td>
</tr>
<tr>
<td>ML_Animal</td>
<td>0.9383</td>
</tr>
<tr>
<td>SoldAnimals</td>
<td>0.9659</td>
</tr>
<tr>
<td>SoldCrop</td>
<td>0.9984</td>
</tr>
<tr>
<td>OutsideJob</td>
<td>0.9872</td>
</tr>
</tbody>
</table>

Figure A6: Partial regression plots for model 2

Figure A7: Partial regression plots for model 2
Figure A8: Standardized residual plots for model 2

Figure A9: Multivariate normality plot for model 2
<table>
<thead>
<tr>
<th>Variable</th>
<th>Tolerance</th>
</tr>
</thead>
<tbody>
<tr>
<td>CLYD</td>
<td>0.9129</td>
</tr>
<tr>
<td>Male</td>
<td>0.9231</td>
</tr>
<tr>
<td>Educ</td>
<td>0.9579</td>
</tr>
<tr>
<td>ProdTree</td>
<td>0.9116</td>
</tr>
<tr>
<td>ML_Animal</td>
<td>0.9209</td>
</tr>
<tr>
<td>SoldAnimals</td>
<td>0.9061</td>
</tr>
<tr>
<td>SoldCrop</td>
<td>0.9240</td>
</tr>
</tbody>
</table>

*Table A3: Tolerance Values for model 2*