2015

Linear Regression Analysis on Net Income of an Agrochemical Company in Thailand

Supichaya Sunthornjittanon
Portland State University

10.15760/honors.137

Linear Regression Analysis on Net Income of an Agrochemical Company in Thailand.

by

Supichaya Sunthornjittanon

An undergraduate honors thesis submitted in partial fulfillment of the requirements for the degree of

Bachelor of Science

in

University Honors

and

Mathematics and Statistics

Thesis Adviser

Jong Sung Kim

Portland State University

2015

Abstract:

The purpose of this research is to analyze the ABC Company's data and verify whether the regression analysis methods and models would work effectively in the ABC Company based in Bangkok, Thailand. After the data are collected, models are created to examine the contribution of each of the company's financial factors to the net income of the company. The final model is selected using Stepwise Regression Methods. A linear regression line and equation for the model are generated to help observe and predict future trends. The model also shows which variables play the most important roles in the company's net income.

Table of Contents

Research Question 4
Introduction 4
Methodology 13
    Model Selection 14
Findings 18
Analysis 22
Discussion 23
Suggestions for Further Study 24
Appendix 26
Bibliography 28

Research Question:

How does the linear regression analysis depict the company's net income? What factors in the company play significant roles in its net income?

Introduction:

The purpose of this research is to analyze the ABC Company's data of the past nine years and verify whether the regression analysis methods and models studied and developed in the U.S. would work effectively in ABC Company based in Bangkok, Thailand.

There are two main elements in this research that one should be familiar with. These elements include some background information about the company and the statistically tools used in the analysis.

Agriculture has always played a major role in the economy of both Thailand and all of Southeast Asia. Its exports are very successful internationally. As a result, many businesses related to agriculture are booming there. Due to the fact that agricultural production is so large there, innovations like chemical fertilizers, pesticides, and herbicides have been produced to protect the plants as they grow and increase yields (Narakon, 2014). ABC Company is one of Thailand's biggest producers, importers, and distributors of pesticides and fertilizers. The company's products can be categorized into five main sections, which are fungicide, pesticide, herbicide, fertilizer, and others. The company's customers include Heinz Thailand for their tomato farm, Lay's for their potato farm, as well as other dealers in Thailand (Sun, 2014).

In order to understand this research, one should be familiar with the materials used in the analysis. The following statistical tools and techniques are used in the analysis.

Regression analysis:

Regression analysis is a technique used in statistics for investigating and modeling the relationship between variables (Douglas Montgomery, Peck, & Vinning, 2012)**.**

Simple linear regression:

Simple linear regression is a model with a single regressor x that has a relationship with a response y that is a straight line. This simple linear regression model can be expressed as

$$y = \beta_0 + \beta_1 x + \varepsilon$$

where the intercept β_0 and the slope β_1 are unknown constants and ε is a random error component .

Multiple linear regression:

If there is more than one regressor, it is called **multiple linear regression.** In general, the response variable y may be related to k regressors, x_1, x_2,…,x_k, so that

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + \varepsilon$$

(Douglas Montgomery, Peck, & Vinning, 2012).

Least Squares Estimation:

The method of least squares is used to estimate β_0, β_1,… β_k. That is, we estimate β_0 and β_1 so that the sum of the squares of the differences between the

observations y_i and the straight line is a minimum (Douglas Montgomery, Peck, & Vinning, 2012).

R-squared:

R-squared is a measure in statistics of how close the data are to the fitted regression line. It is also known as the coefficient of determination, or the coefficient of multiple determinations for multiple regression. It is the percentage of the response variable variation that is explained by a linear model.

$$\mathrm{R-squared} = \frac{\text{Explained variation}}{\text{Total variation}}$$

R-squared is always between 0 and 100%. 0% means the model explains none of the variability of the response data around its mean. 100% indicates that the model explains all the variability of the response data around its mean. Generally, the higher the R-squared, the better the model fits the data (Frost, 2013).

Analysis of variance (**ANOVA**):

Analysis of variance (**ANOVA**) is a collection of statistical models used in order to analyze the differences between group means and their associated procedures. In the ANOVA setting, the observed variance in a particular variable is partitioned into components attributable to different sources of variation. The following equation is the Fundamental Analysis-of-Variance Identity for a regression model.

$$\sum_{i=1}^{n}(y_i - \overline{y})^2 = \sum_{i=1}^{n}(\widehat{y_i} - \overline{y})^2 - \sum_{i=1}^{n}(y_i - \widehat{y_i})^2$$

$$SS_T = SS_R + SS_{Res}$$

(Douglas Montgomery, Peck, & Vinning, 2012).

Statistical Hypothesis:

Statistical hypothesis are statements about relationships. The statistical hypothesis testing is the use of statistics to determine the probability that a given hypothesis is true (Iyanaga& Kawada, 1980). The null hypothesis is denoted by H_0. The alternative hypothesis is the negation of the null hypothesis, denoted by H_1 or H_a (Wyllys, 2003).

Testing Significance of Regression:

H_0: $\beta_1 = \beta_2 = \ldots = \beta_k = 0$

H_1 : at least one $\beta_i \neq 0$

The hypotheses are related to the significance of regression. Failing to reject H_0 implies that there is no linear relationship between x and y. On the other hand, if H_0 is rejected, it implies that at least one β_i show a significant relationship to y (Douglas Montgomery, Peck, & Vinning, 2012).

F-test:

An **F-test** is a statistical test in which the test statistic is based on the F-distribution under the null hypothesis. It is most often used when comparing statistical models that have been fitted to a data set, in order to identify the model that best fits the population from which the data were sampled. In this research, the F-test is used to test the significance of the model.

The test statistics F_0 can be computed by $\frac{MS_R}{MS_{RES}}$ follows the $F_{k,n-k-1}$ distribution. Reject H_0 , if $F_0 > F_{k,n-k-1}$. The test statistic F_0 can usually be obtained from the ANOVA table (Experiment Design and Analysis Reference, n.d.).

Test on Individual Regression Coefficients (*t* Test):

The **t-test** is used to check the significance of individual regression coefficients in the multiple linear regression model. Adding a significant variable to a regression model makes the model more effective, while adding an unimportant variable may make the model worse. The hypothesis statements to test the significance of a particular regression coefficient, β_j , are:

$$H_0: \beta_j = 0$$

$$H_1 : \beta_j \neq 0$$

The test statistic for this test has the t-distribution:

$$T_0 = \frac{\hat{\beta}_j}{se(\hat{\beta}_j)}$$

where the standard error, $se(\hat{\beta}_j)$, is obtained. One would fail to reject the null hypothesis if the test statistic lies in the acceptance region:

$$-t_{\alpha/2,\, n-2} < T_0 < t_{\alpha/2,\, n-2}$$

This test measures the contribution of a variable while some other variables are included in the model (*Experiment Design and Analysis Reference*, n.d.).

P-value:

P-value or calculated probability is the estimated probability of rejecting the null hypothesis (H_0) of a study question when that hypothesis is true.

VIF:

VIF (the variance inflation factor) for each term in the model measures the combined effect of the dependences among the regressors on the variance of the term. Practical experience indicates that if any of the VIFs exceeds 5 or 10, it is an indication that the associated regression coefficients are poorly estimated because of multicollinearity (Douglas Montgomery, Peck, & Vinning, 2012).

Model Selection:

Stepwise Regression Methods is an approach to selecting a subset of effects for a regression model by either adding or deleting regressors one at a time. These methods can be divided into three categories:

1.) **Forward Selection**: This procedure begins with the assumption that there is no regressor in the model other than the intercept. The goal is to find an optimal subset by inserting regressors into the model one at a time.

2.) **Backward Elimination**: This procedure is the opposite approach from the forward selection. First, we begin with a full model with K candidate regressors. Then, the partial F statistic (or a t statistic) is computed for each regressor. If the regressor with the smallest partial F or t value is less than the preselected F value, that regressor is removed from the model. Fit model with K-1 predictors and the procedure is repeated.

3.) **Stepwise Regression**: It is a method that allows moves in either direction, dropping or adding variables at the various steps. It combines both forward selection and backward elimination. We perform two steps in forward selection and a backward step. Then, perform another forward step and another backward step. We continue until no action can be taken in either direction (Douglas Montgomery, Peck, & Vinning, 2012).

Residuals:

The difference between the observed value of the dependent variable (y) and the predicted value ($\hat{y}$) is called the **residual** (e). Each data point has one residual.

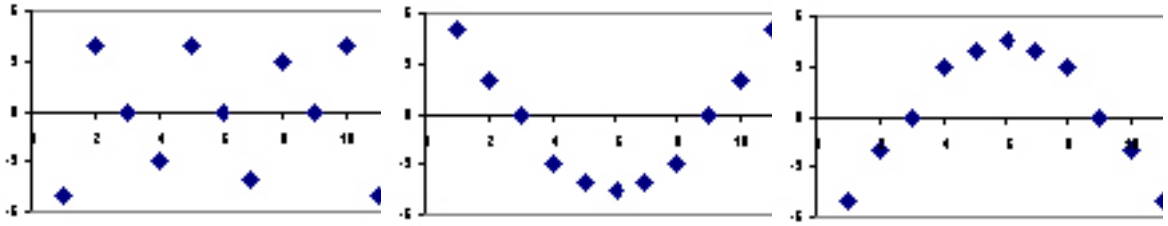
$$\text{Residual} = \text{Observed value} - \text{Predicted value}$$

$$e = y - \hat{y}$$

Both the sum and the mean of the residuals are equal to zero. That is, $\Sigma e = 0$ and $e = 0$ (Douglas Montgomery, Peck, & Vinning, 2012).

Residual Diagnostics:

A **residual plot** is a graph that shows the residuals on the vertical axis and the independent variable on the horizontal axis. If the points in a residual plot are randomly dispersed around the horizontal axis, a linear regression model is appropriate for the data; otherwise, a non-linear model is more appropriate. The residual plots below show three typical patterns.



Random pattern **Non-random: U-shaped** **Non-random: Inverted U**

Figure 1 Residual Plot (Berman, n.d.)

The first plot shows a random pattern (or no pattern), indicating a good fit for a linear model. The other plot patterns are non-random (U-shaped and inverted U) (Berman, n.d.).

Checking normality:

Histogram:

The **Histogram** of the Residual can be used to check whether the variance is normally distributed. A symmetric histogram as shown in the figure 2 below, which is evenly distributed around zero, indicates that the normality assumption is likely to be true ("Graphic Residual Analysis," n.d.). The typical "bell-curve" is the ideal indication as to normality. When this cannot be obtained, a symmetrical histogram is sufficient.

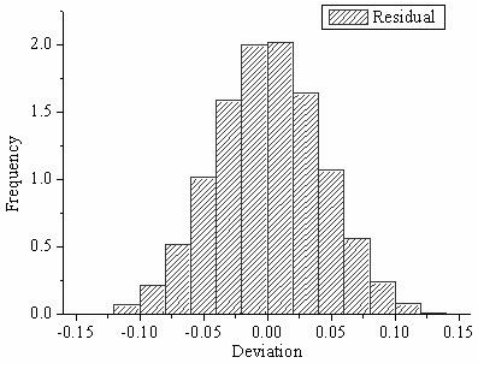


Figure 2 A symmetric histogram ("Graphic Residual Analysis," n.d.)

Normality Plot:

The normality plots below show some of the ways that residuals can deviate from normality.

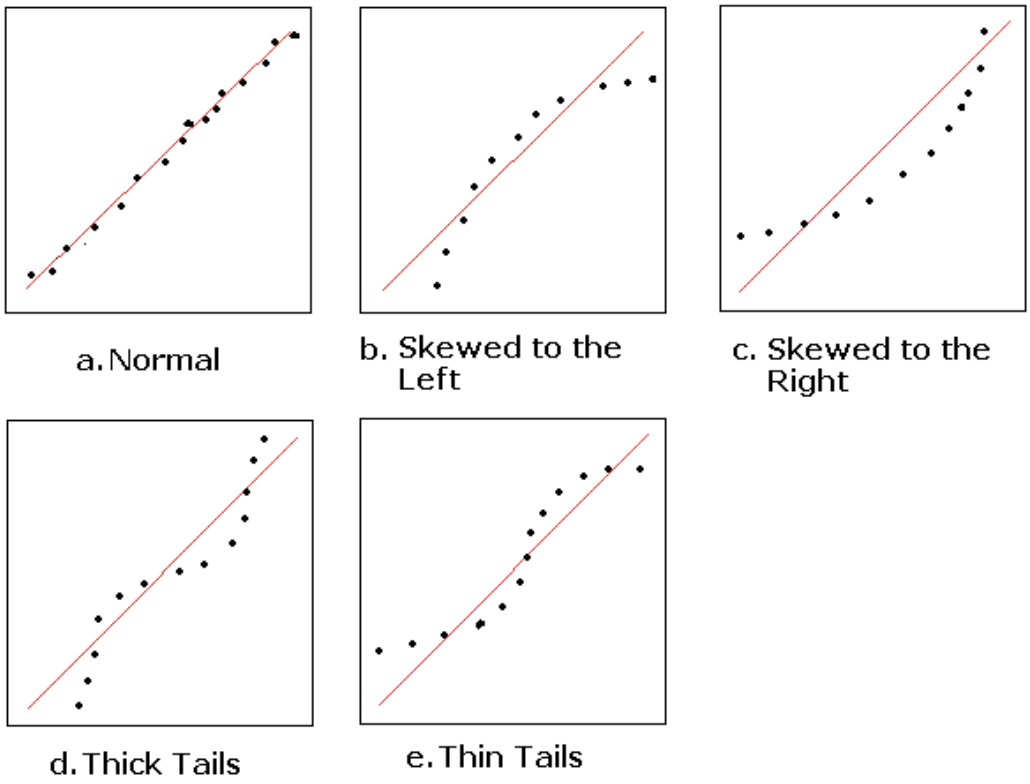


Figure 3 Normality Plot ("Residual Analysis," n.d.)

Minitab:

Minitab is a program used in this research. Minitab is a general-purpose statistical software package designed for easy interactive use. Minitab is not only well suited for instructional applications, but is also powerful enough to be used as a primary tool for analyzing research data ("What is Minitab, and where can I access it?," 2014).

Methodology:

The methods and models are created based on the company's available data and regression analysis methods, using the company's sales numbers and profits for the past nine years (2005-2013). All data for this analysis were supplied directly by the accounting department of ABC Company in Thailand for educational purposes with no funding received.

The data provided includes nine factors of the company, which are Net Income, Net Sales, Financial Cost, Cost of Goods Sold, Income from Fungicide (FUNG), Income from Pesticide (IN), Income from Herbicide (HER), Income from Fertilizer (FER), and Income from all other products manufactured and distributed by the company (Other). Within each category of product produced by the company (fungicide, herbicide, etc.), data are collected for each individual product within that particular product line.

When each multiple regression model is created, R-squared, P-value, VIF, and residuals are analyzed. When a simple regression model is created, R-squared value, P-value, and residuals are observed.

Stepwise Regression Methods are used in the model selection process, which is shown in the following model selection section.

Model Selection:

1.) Forward Selection Method: (From Minitab)

Regression Analysis: Net income versus Sales-net, Finance Cost, Cost of goods, FUNG, IN, ...

Forward Selection of Terms

α to enter = 0.05

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	1	3.07036E+12	3.07036E+12	87.70	0.000
FUNG	1	3.07036E+12	3.07036E+12	87.70	0.000
Error	7	2.45081E+11	35011506795		
Total	8	3.31544E+12			

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
187114	92.61%	91.55%	86.73%

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	1831745	220169	8.32	0.000	
FUNG	1.354	0.145	9.36	0.000	1.00

Regression Equation

Net income = 1831745 + 1.354 FUNG

Figure 4 Forward selection results from Minitab for the ABC Company data.

2.) Backward Elimination:

The net income is perfectly explained by all of the variables combined, so the standard error is zero. The test of statistics is undefined when regressing Net Income with the eight main factors on linear regression (as shown in appendix A). In order to further use Backward Elimination in pinpointing a factor's contribution to the net income of a company, every variable, but one (K-1) are selected to generate the initial model for this method.

Regression Analysis: Net income versus Finance Cost, Cost of good, FUNG, IN, HER, FER, Other

Backward Elimination of Terms

α to remove = 0.05

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	1	3.07036E+12	3.07036E+12	87.70	0.000
FUNG	1	3.07036E+12	3.07036E+12	87.70	0.000
Error	7	2.45081E+11	35011506795		
Total	8	3.31544E+12			

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
187114	92.61%	91.55%	86.73%

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	1831745	220169	8.32	0.000	
FUNG	1.354	0.145	9.36	0.000	1.00

Regression Equation

Net income = 1831745 + 1.354 FUNG

Figure 5 Backward selection results from Minitab for the ABC Company data

3.) Stepwise Selection: (From Minitab)

Regression Analysis: Net income versus Sales-net, Finance Cost, Cost of goods, FUNG, IN, ...

Stepwise Selection of Terms

α to enter = 0.05, α to remove = 0.05

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	1	3.07036E+12	3.07036E+12	87.70	0.000
FUNG	1	3.07036E+12	3.07036E+12	87.70	0.000
Error	7	2.45081E+11	35011506795		
Total	8	3.31544E+12			

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
187114	92.61%	91.55%	86.73%

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	1831745	220169	8.32	0.000	
FUNG	1.354	0.145	9.36	0.000	1.00

Regression Equation

Net income = 1831745 + 1.354 FUNG

Figure 6 Stepwise selection results from Minitab for the ABC Company data.

Findings:

Simple Regression: Net income versus Fung

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	1	3.07036E+12	3.07036E+12	87.70	<0.0001
FUNG	1	3.07036E+12	3.07036E+12	87.70	<0.0001
Error	7	2.45081E+11	3.50115E+10		
Total	8	3.31544E+12			

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
187113.620	92.61%	91.55%	86.73%

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	1831745	220169	8.32	<0.0001	
FUNG	1.3543	0.1446	9.36	<0.0001	1

Regression Equation

Net income = 1831745 + 1.3543 FUNG

Figure 7 Final Minitab Regression Analysis

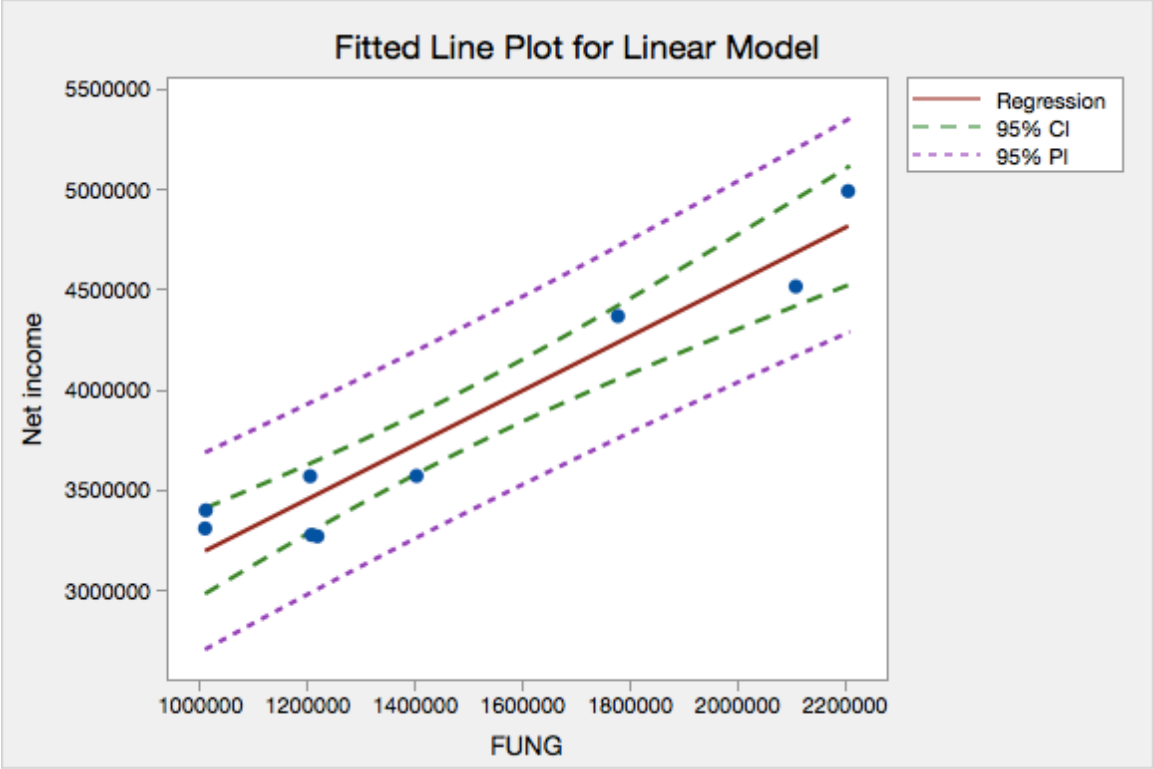


Figure 8 Plot of Fitted Line for ABC Company's findings.

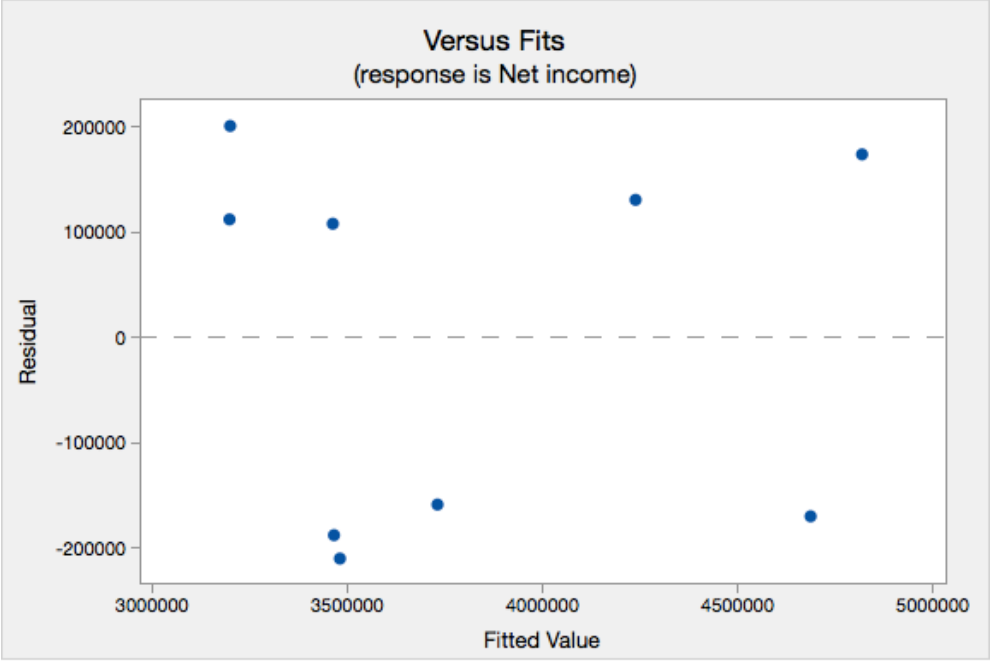


Figure 9 Plot of Residuals versus Fitted Values for the ABC Company data.

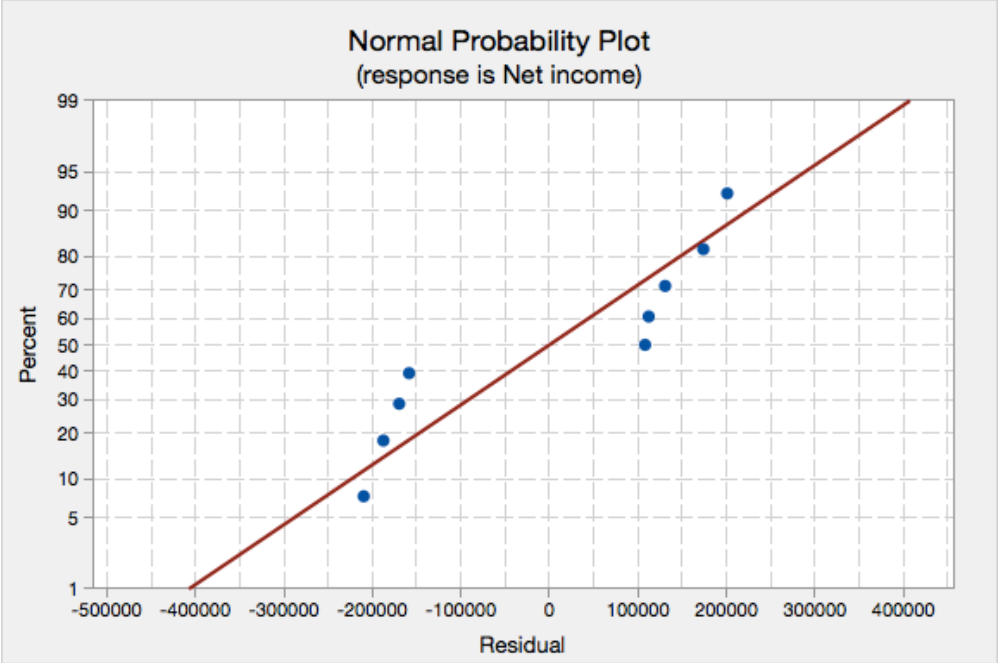


Figure 10 Normality Probability Plot of the ABC Company data.

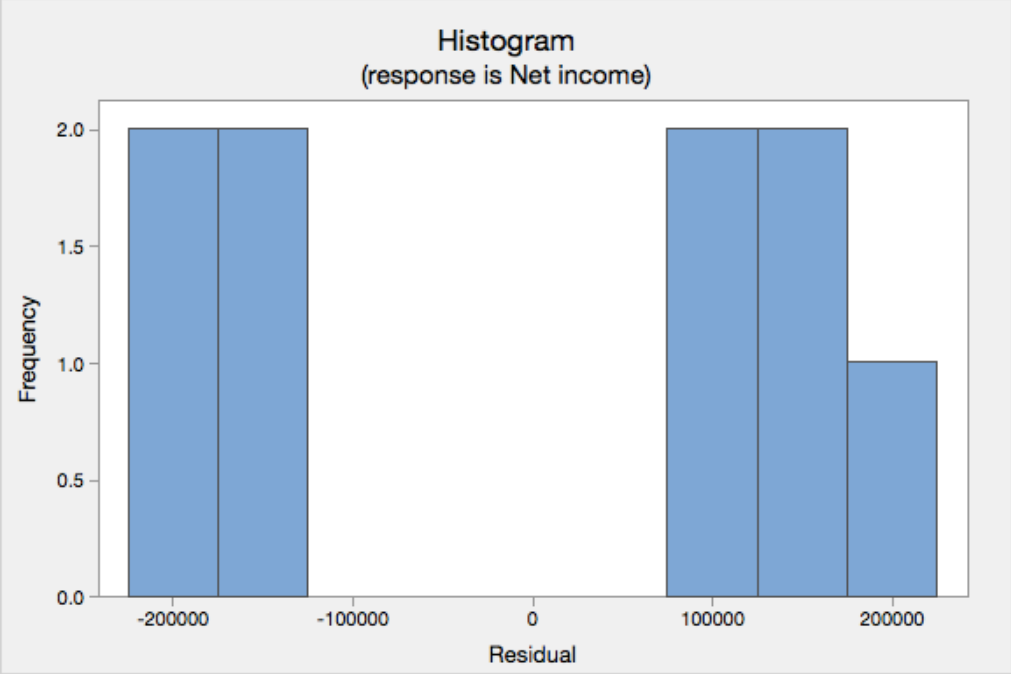
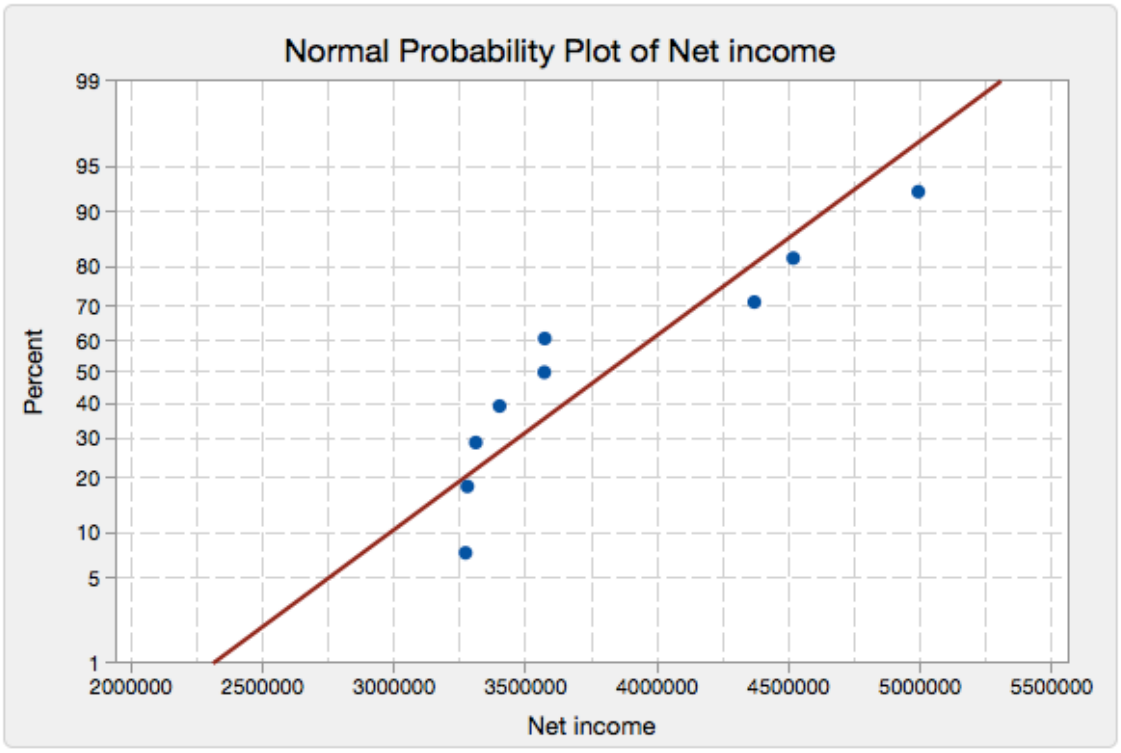


Figure 11 Histogram of the ABC Company data.

Normality Test:



Descriptive Statistics

N	Mean	StDev
9	3809073	643762

Anderson-Darling Test

Null hypothesis	H_0: Data follow a normal distribution
Alternative hypothesis	H_1: Data do not follow a normal distribution

AD-Value	P-Value
0.78	0.0264

Figure 12 Normality Test (Anderson-Darling Test)

Analysis:

The hypothesis in this case is

H_0 **:** $\beta_1 = 0$ (The relationship between Net Income and Income from Fungicide is not significant).

H_1 : $\beta_1 \neq 0$ (The relationship between Net Income and Income from Fungicide is significant).

According to the findings of the final model (Net Income versus Income from Fungicide), it can be seen that the p-value from the Analysis of Variance table (figure 7) is less than 0.0001, which is very low. The null hypothesis can be rejected, which means there is significant relationship between Net Income and Income from Fungicide at any significant level. From the Model Summary table, R-squared value is 92.61%, which indicates that 92.61% of the variation of Net Income is explained by the model.

Moreover, the residual plot in figure 9 shows no pattern, which implies that this model shows independence of the residuals. In examining figure 10, the normality plot could appear to indicate deviation from normality. However, it must be considered that the x-axis depicts data on a much smaller scale than the actual data and the number of the data points examined total only nine. Considered these factors, the normality assumption seems to be satisfied. Similarly, the histogram (as shown in figure 11) appears to be symmetric, which seems to depict normality.

Normality test, Anderson-Darling, is also conducted. However, normality test is actually rejected because there are not many data points, which create a

gap in the net income. Usually, to test normality, the sample size should be large. Even though according to normality test, the normality assumption is not satisfied, the model satisfies the goal of finding the pattern and trend of the company's net income and identifying which factors play statistically significant roles in the net income.

Therefore the equation that can best depict the company's net income is

Net income = 1831745 + 1.3543 FUNG

Discussion:

After model selection method is processed, the consensus has shown that only the Income from Fungicide plays a statistically significant role in the net income of the company.

After the significant category is found deeper analysis is conducted of the fungi category. Each individual fungicide product line is then regressed on the net income. After running these additional models, it was discovered that the Net Income from Fungicide remained the most significant model. Time is also taken into account to see if it plays some role in the net income, but after the analysis, it was found that time is not significant in this case.

From the model and research, it can be concluded that Income from Fungicide plays the most significant role in the company's net income. It can be used as a tool to estimate the company's future net income. The company should pay close attention to this product line. Based on the study findings, some suggestions for the company, which may prove beneficial in increasing future net income may include lowering the cost of production of fungicide, manufacturing

increased volume of fungicide, increasing the advertising budget for this product line, and increasing negotiations surrounding the product.

However, it is important to note that even though fungicide product might be significant, some practical factors may make it challenging to manage in reality. Some of these issues include limited supplies of the product, a limitation in company's production capacity, a limitation in company's budget, and other challenges that may require more investment than would prove worthwhile.

Some factors are uncontrollable and may have caused unexpected an result in the net income. Some of these factors that may include, but are not limited to, natural disasters in the past year (flooding, earthquakes, etc.), Thailand's potentially volatile economy, unstable political systems, fluctuations in exchange rates, changing government policies, etc.

Suggestions for Further Study:

The ABC Company has been in existence for more than 15 years. The company has faced many ups and downs and changes throughout those years. All appropriate data should be taken into account and examined when the model is created.

Since there is limitation in accessing the company's data, only 9 years of data sets are used in this thesis. In order to create a better and more accurate model and result, more data points from the company should be examined and analyzed.

Since the normality test, Anderson-Darling, does not satisfy the assumption in this case, one could go further to conduct piecewise analysis.

There are clearly two separate net income groups on the scatter plot (as shown in figure 8) due to the small sample size.

Due to my personal interest, I hope to continue the study and analysis of data as related to ABC Company. I recommend to ABC Company to further the study of models relating to the net income of the company.

Appendix A

Multiple Regression: Net Income versus Net Sales, Financial cost, Cost of goods sold, Income from Fungicide, Income from Pesticide, Income from Herbicide, Income from Fertilizer, and Income from all other products.

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	8	3.31544E+12	414429530847	*	*
Sales-net	1	7.60666E+10	76066610168	*	*
Finance Cost	1	9.93295E+08	993294929	*	*
Cost of good sold	1	7.18686E+10	71868575714	*	*
FUNG	1	3.88639E+10	38863868441	*	*
IN	1	5.65584E+10	56558449084	*	*
HER	1	7.68714E+10	76871431031	*	*
FER	1	6.38368E+10	63836798067	*	*
Other	1	5.64754E+10	56475422438	*	*
Error	0	0	*		
Total	8	3.31544E+12			

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
*	100.00%	*	*

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	-25783069.1	*	*	*	
Sales-net	3.83698008	*	*	*	11819.9265
Finance Cost	2.99045252	*	*	*	41.4014
Cost of good sold	-3.98550912	*	*	*	10619.1025
FUNG	-2.84515356	*	*	*	348.6572
IN	-1.43293289	*	*	*	192.0251
HER	-10.2095816	*	*	*	1045.0083
FER	2.11818853	*	*	*	193.8126
Other	16.2027791	*	*	*	102.8735

Regression Equation

Net income = − 25783069.1 + 3.83698008 Sales-net + 2.99045252 Finance Cost − 3.98550912 Cost of good sold − 2.84515356 FUNG − 1.43293289 IN − 10.2095816 HER + 2.11818853 FER + 16.2027791 Other

Figure 13 Minitab result with all variables in the model

Note: Residual plots could not be displayed because MSE = 0 or the degrees of freedom for error = 0.

Bibliography

Berman, H. (n.d.). Residual Analysis in Regression. *Stat Trek*. Retrieved from http://stattrek.com/regression/residual-analysis.aspx

Data. (2014, September 25). ABC Company's Accounting Department.

Douglas Montgomery, Peck, E., & Vinning, G. (2012). *Introduction to Linear Regression Analysis* (5th ed.). Wiley.

Experiment Design and Analysis Reference. (n.d.). ReliaSoft. Retrieved from http://reliawiki.org/index.php/Experiment_Design_and_Analysis_Reference

Frost, J. (2013, May 30). Regression Analysis: How Do I Interpret R-squared and Assess the Goodness-of-Fit? Retrieved from http://blog.minitab.com/blog/adventures-in-statistics/regression-analysis-how-do-i-interpret-r-squared-and-assess-the-goodness-of-fit

Graphic Residual Analysis. (n.d.). *OriginLab Corporation*. Retrieved from http://www.originlab.com/doc/Origin-Help/Residual-Plot-Analysis

Iyanaga, S., & Kawada, Y. (1980). *Statistical Estimation and Statistical Hypothesis Testing.* (Vol. Appendix A, Table 23). Cambridge, MA: MIT Press.

Narakon, S. (2014, December 1). Accounting Department Manager at ABC Company.

Residual Analysis. (n.d.). *DePaul University*. Retrieved from http://facweb.cs.depaul.edu/sjost/csc423/documents/resid-anal.htm

Sun, S. (2014, October 12). Marketing Director at ABC Company.

What is Minitab, and where can I access it? (2014). *Indiana University*. Retrieved from https://kb.iu.edu/d/cagq

Wyllys, R. (2003). STATISTICAL HYPOTHESES. *The University of Texas At Austin School of Information*. Retrieved from https://www.ischool.utexas.edu/~wyllys/IRLISMaterials/stathyp.pdf