A WEIGHTED LEAST SQUARES ESTIMATION OF THE POLYNOMIAL REGRESSION MODEL ON PADDY PRODUCTION FOR THE MUDA AGRICULTURE AND DEVELOPMENT AUTHORITY (MADA) AREA

by

ROSLIZA BINTI MUSA

Dissertation submitted in partial fulfillment
of the requirements for degree
of Master of Science in Statistics

January 2016

ACKNOWLEDGEMENT

Alhamdulillah. First of all, I would like to express my gratitude to His Almighty for giving me the strength and courage to complete this dissertation this semester as part of the requirement for the degree of Master of Science in Statistics.

I would like to express my special appreciation and many thanks to my dissertation supervisor, Puan Zalila Binti Ali who gave me guidance, suggestions and encouragement throughout completing this dissertation.

Deepest thanks to the Muda Agricultural and Development Authorities (MADA), for providing the data required in this study and also to the School of Mathematical Sciences of Universiti Sains Malaysia for providing me guidance in writing the dissertation.

Also special thanks to my husband, kids, family and friends who provided endless support during the study period. Last but not least, I would like to thank everyone involved directly or indirectly in completing this dissertation.

TABLE OF CONTENT

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATION

ABBREVIATION	DESCRIPTION
ANOVA	Analysis of Variance
BLS	Barat Laut Selangor (Northwest Selangor)
DAS	Day After Seeding
DOA	Department of Agriculture
FAO	Food and Agriculture Organization
GDP	Gross Domestic Product
GNI	Gross National Income
IADA	Integrated Agriculture Development Area
K	Kalium
KADA	Kemubu Agriculture Development Authority
KETARA	North Terengganu Integrated Agriculture Development
MADA	Muda Agricultural and Development Authorities
MARDI	Malaysia Agricultural Research and Development Institute
MLE	Maximum Likelihood Estimation
MSE	Mean Square Error
MSR	Mean Square Regression
N	Nitrogen
NAP	National Agriculture Policy
NKEA	Agriculture National Key Economic Area
NPK	Nitrogen, Phosphorus, Kalium mixture
P	Phosphorus
PPS	Paddy Production Survey
R&D	Research and Development
SSE	Sum of Square Error
SSL	Self-Sufficiency Level
SSR	Sum of Square Regression
SST	Sum of Square Total
VIF	Variance Inflation Factor

SUATU ANGGARAN PEMBERAT BAGI SUATU MODEL REGRESSI POLINOMIAL TERHADAP PENGELUARAN PADI DI KAWASAN LEMBAGA KEMAJUAN DAN PERTANIAN MUDA (MADA)

ABSTRAK

Hubungan garis lengkung antara pembolehubah bersandar dan pembolehubah tak bersandar boleh diwakili oleh model regresi polinomial. Model ini digunakan untuk mengkaji hubungan antara pembolehubah bersandar dan pembolehubah peramal yang hadir dalam bentuk kuasa dua dan ke atas. Model regresi polinomial ialah satu model khas bagi model regresi linear berganda. Pembinaan model regresi polinomial mengambil ciri-ciri pembinaan model regresi linear berganda dalam aspek kaedah penganggaran parameter regresi, pentaabiran bagi model regresi berganda, pemilihan pembolehubah peramal dan model diagnostik. Anggaran pemberat digunakan sebagai alternatif untuk mengatasi masalah varian tak malar. Kajian ini menggunakan model regresi polinomial bersama anggaran pemberat untuk mengkaji pengeluaran padi di kawasan Lembaga Kemajuan dan Pertanian Muda (MADA) berdasarkan ciri-ciri penanaman dan persekitaran. Keputusan analisis menunjukkan pengeluaran padi dipengaruhi oleh varieti padi, keluasan tanaman, amaun baja sebatian, kitaran pembajaan baja sebatian selepas tanam, kitaran pembajaan baja NPK selepas tanam, purata suhu, purata hujan, serangan perosak, purata hujan peringkat kedua, serangan perosak peringkat kedua dan saling tindakan antara keluasan tanaman dengan amaun

baja sebatian, varieti padi dan kitaran pembajaan baja NPK selepas tanam, serangan perosak dengan kitaran pembajaan baja NPK selepas tanam.

ABSTRACT

The curvilinear relationship between a dependent variable and several independent variables can be represented by a polynomial regression model. This model is used to study the relationship between response variable and predictor variable which contain square and higher-order term. Polynomial regression model is a special case of multiple regression model. The building of polynomial regression model has the same characteristics as multiple linear regressions in term of parameter estimation, regression inference, variable selection and model diagnostic. Weighted least square estimation is used as a remedy for non-constant variance. This study used polynomial regression model with weighted least square estimation to investigate paddy production of different paddy lots based on environmental and cultivation characteristics in Muda Agriculture and Development Authority (MADA) area. The result shows the factors that affecting paddy production are paddy variety, acreage, amount of mixture fertilizer, mixture fertilizer application after seeding, NPK fertilizer application after seeding, average temperature, average rainfall, pest and disease, average rainfall squared, pest and disease squared also interaction between acreage with amount of mixture fertilizer, paddy variety with NPK fertilizer application after seeding, pest and disease with NPK fertilizer application after seeding.

CHAPTER 1

INTRODUCTION

1.1 Background of Regression Analysis

In many situations, the variation in the experimental measurements of a variable is caused to a great extent by other related variables whose magnitude change over the course of the experiment. As an example, the student obesity measured by races, gender, age, eating habit and medical condition. Hence, obesity is measured by body weight and denoted as dependent variable, whereas races, gender, age, eating habit and medical condition are denoted as independent or predictor variables. Regression analysis is a body of statistical methods dealing with the information on the relationship between a dependent variable and several independent variables.

The relationship between a dependent variable and one or more independent variables could be linear or non-linear and can be represented by a regression model which consists of several parameters (Sweet A. *et al.*, 2012). If the relationship between dependent variables and independent variables show a straight line, it is assumed to have a linear relationship and expressed mathematically by using a linear regression model. A non-linear relationship is curvilinear and is expressed mathematically by using a non-linear regression model such as a logarithmic model, a logistic model or polynomial model. Various of fields that used this type of analysis in their research such as biological (Wagner, 2013), agricultural (Karkacier *et al.*, 2006), economic (Chen,

2010) and engineering (Yitagesu *et al.*, 2008) fields. In economic studies, a linear regression becomes the predominant empirical tool to describe the relationship between dependent and independent variables such as to predict consumption spending, labor demand and labor supply. Exponential regression is one of the nonlinear regression models which is used by agricultural and biological researchers to describe the growth of the plant or bacteria (Miguez & Archontoulis, 2014).

The parameter of a regression model can be estimated by using either the least squares method or the maximum likelihood method (Marill, 2004). The least squares estimation method estimate the parameters by minimizing the sum of squares of the error terms in the model. The method of Maximum Likelihood Estimation (MLE) can be performed when the distribution of the error terms is known and belong to a certain parametric family of probability distribution. The maximum likelihood estimation uses the product of the probability densities as the measure of consistency of the parameter value with the sample data (Myung, 2003).

Regression model building involves two main processes which are selection of the independent variables and model diagnostics. The significance of a single predictor variable is determined by using a *t*-test for small sample data or a z-test for a large sample data. Variables selection process is a step to identify the important independent variables to be included in the regression model. The automatic selection procedures consist of the enter method, forward method, backward method and stepwise method. These methods develop a sequence of regression models, at each step either adding or deleting predictor variables. The criterion for adding or deleting predictor variables is based on the *F* statistics. The all-possible regression procedure considers all possible subsets of potential predictor variables and identifying a few good subsets according to

several criterions. The different criteria used for comparing the different subset are the coefficient of multiple determinations (R^2), mean residual sum of square (s^2) and Mallow C_p (Draper & Smith, 1998).

Model diagnostic is important to determine whether the model is appropriate and meets all the assumptions before performing any inference. If the assumptions are violated, it will have an effect on the validity of the prediction and lead to a faulty conclusion. The assumptions for regression analysis consist of linearity of data, normality and homogeneity of error term, multicollinearity among predictor variables, outliers and influential value. These assumptions can be checked by using graphical plot or statistical test. Scatter plot and residual plot can be used for checking the linearity between dependent variables and independent variables. For normality assumption, the diagnostic tool can be either graphical method or statistical test such as normality plot and Shapiro-Wilk test. The residual plot and statistical tests such as Levene test or the Bartlett test can be used to check the homogeneity of variance. Multicollinearity exists when the predictor variables are highly correlated and this problem will have an effect on the parameter estimated. Variance inflation factor (VIF) is an indicator of the existence of multicollinearity and measures how much the variance of an estimated regression coefficient is increased because of collinearity (Graham, 2003). In certain situations, the outlier values in the sample data can affect the estimated model and this problem is detected by leverage value and Cook's Distance to identify the possible influential outliers (Ghosh & Vogt, 2012). Cook's Distance test provides the information either the value should be retained or removed from the sample.

Inferences related to linear regression consist of the inference for interval estimation of the mean response and prediction interval of a new observation. Interval

estimation is used on the sample data to calculate an interval of possible value of an unknown population parameter while prediction interval is an estimate of an interval in which future observation will fall (Bingham & Fry, 2010).

1.2 Background of Case Study

This study applied the regression techniques on paddy production based on the environmental and cultivation characteristics in the Muda Agricultural and Development Authorities (MADA) of Kedah and Perlis area. An overview of paddy industry, paddy cultivation activities and factors that affect paddy production is presented in this section.

1.2.1 The Paddy industry

Rice is currently grown in over a hundred countries that produce more than 715 million tons of paddy rice annually (480 million tons of milled rice). Rice production are among the highest in Asian populations which China, India,Indonesia, Bangladesh, Vietnam, Myanmar, Thailand, the Philippines, Japan, Pakistan, Cambodia, the Republic of Korea, Nepal, and Sri Lanka and Asian countries account for 90% of the world's total rice production. Rice provides up to 50% of the dietary caloric supply for millions living in poverty in Asia and is, therefore, critical for food security (FOASTAT, 2014). Rice remains one of the most protected food commodities in world trade and being special importance nutrition for population in Asia and part of America Latin. It is also the primary source of income and employment for more than 200 million household across countries in developing world (FOA, 2013).

Two species of rice are important for human consumption: Oryza sativa and Oryza glaberima. Oryza sativa is more widely grown in more than 100 countries, including in Asia, North and South America, the European Union, the Middle East, and Africa and, is the staple food of an estimated 3.5 billion people worldwide. The long-grained indica variety is widely grown in tropical and sub-tropical Asia. Paddy rice is the end product of the harvesting and threshing of rice grains. The paddy rice is made up of an outer husk layer, germ and bran layers, and the endosperm. The weight of white rice only accounts for 72% of the total weight of paddy (Mitchell, 2009).

Agriculture sector in Malaysia is broad, encompassing industrial crops such as oil palm and rubber, and agro food industry (food that is produced by agriculture) such as paddy and livestock. Paddy is one of the important crops and its production is a strategic industry in Malaysia because it supplies the staple food of Malaysians. Due to its importance as a security crop, paddy industry has been given special attention and consideration by the government in order to sustain and increase the total of paddy production (Ramli *et al.*, 2012). Paddy was selected as one of the sub-sectors with the aim to strengthen the productivity of Malaysia's paddy farms in order to establish Malaysia's long term food security while increasing the income of paddy farmers (Economic Transformation Programme, 2010). The government intervention and attention to the farmers are in the form of financial allocation, fertilizer subsidy, technology application and other incentives. It was reported that the sub-sector under Agriculture National Key Economic Area (NKEA) accounted for 82% of agriculture's contribution to Malaysia Gross National Income (GNI) in 2009.

Global food crisis had caused a short supply in rice production and that situation was very critical when the price increased greatly and the exporters were reluctant to

sell their rice to other countries. Malaysian government transformed the paddy industry and its policy to ensure stability in national food security and reduce dependency on importation (Rabu & Mohd Shah 2013).

Table 1.1: Gross Domestic Product (GDP) by Sector, 2000-2010.

SECTORS	2000		2005		2010		Average yearly growth rate (%)		
	RM (mil)	%	RM (mil)	%	RM (mil)	%	2001 - 2005	2006 - 2010	2001 - 2010
AGRICULTURE	30,647	8.6	35,835	8.0	40,680	7.3	3.2	2.6	2.9
Industrial crop	18,759	5.3	22,031	4.9	21,822	3.9	3.3	-0.2	1.5
Agrofood industry	11,888	3.3	13,804	3.1	18,858	3.4	3.0	6.4	4.7
MINING	37,617	10.6	42,427	9.5	40,338	7.2	2.5	-1.0	0.7
MANUFACTURING	109,998	30.9	137,940	30.7	154,621	27.7	4.6	2.3	3.5
CONTRUCTION	13,971	3.9	14,685	3.3	18,220	3.3	1.0	4.4	2.7
SERVICES	175,649	49.3	230,043	51.2	320,559	57.4	5.5	6.9	6.2

Source: National Agro-food Policy 2011-2020

National Agro Food Policy 2011-2020 was formulated to replace The Third National Agriculture Policy (NAP 3) 1998-2010. The main focuses are to ensure sufficient food supplies and increase agriculture's contribution to Gross Domestic Product (GDP) (Dasar Agromakanan Negara 2011-2020, 2011). In year 2010, agro food industry contributed about RM 18.9 billion to GDP (Table 1.1) and 4.5% or RM 849 million (Table 1.2) was contributed by paddy production.

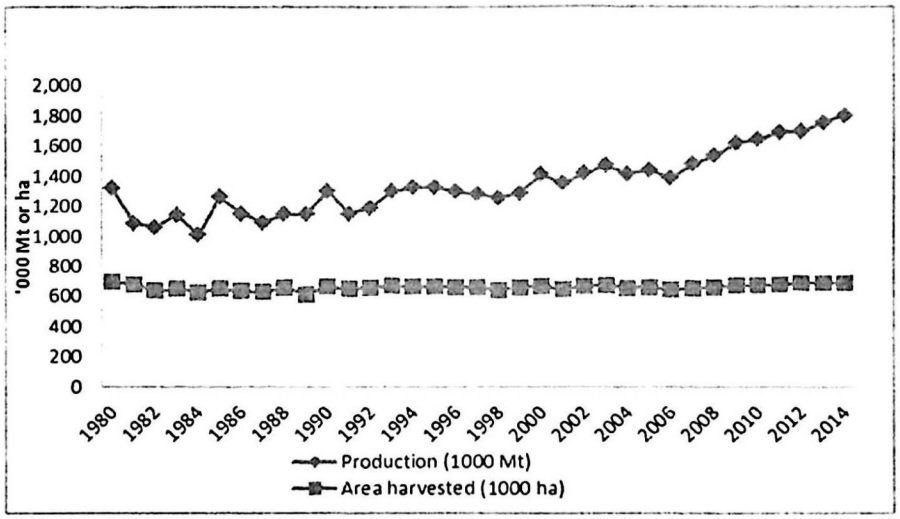
Being a staple food and crucial to the Malaysian diet, rice is very important for stability and population growth. Stability and security of rice supply in national level is measured through self-sufficiency level (SSL) (Arhad & Hameed, 2009). Rice production in Malaysia showed increasing trend since 1980 to 2014 but the amount is still not sufficient (Figure 1.1).

Table 1.2: Value Added Agro Food Industry And Agro-Based, 2000-2010

SECTORS	2000		2005		2010		Average yearly growth rate (%)		
	RM (mil)	%	RM (mil)	%	RM (mil)	%	2001 - 2005	2006 - 2010	2001 - 2010
AGROFOOD INDUSTRY	**11,888**	**100.0**	**13,805**	**100.0**	**18,858**	**100.0**	**3.0**	**6.4**	**4.7**
Plant	3,578	30.1	4,169	30.2	5,787	30.7	3.1	6.8	4.9
Paddy	738	6.2	797	5.8	849	4.5	1.5	1.3	1.4
Fruits	957	8.0	1,327	9.6	1,928	10.2	6.8	7.8	7.3
Vegetables	1,883	15.8	2,045	14.8	3,010	16.0	1.7	8.0	4.8
Fishing	**5,083**	**42.8**	**5,434**	**39.4**	**7,285**	**38.6**	**1.3**	**6.0**	**3.7**
Livestock	**2,382**	**20.0**	**3,261**	**23.6**	**4,730**	**25.1**	**6.5**	**7.7**	**7.1**
Others	**845**	**7.1**	**941**	**6.8**	**1,056**	**5.6**	**2.2**	**2.3**	**2.3**
AGRO-BASED INDUSTRY	**7,077**	**100.0**	**9,698**	**100.0**	**12,685**	**100.0**	**6.5**	**5.5**	**6.0**

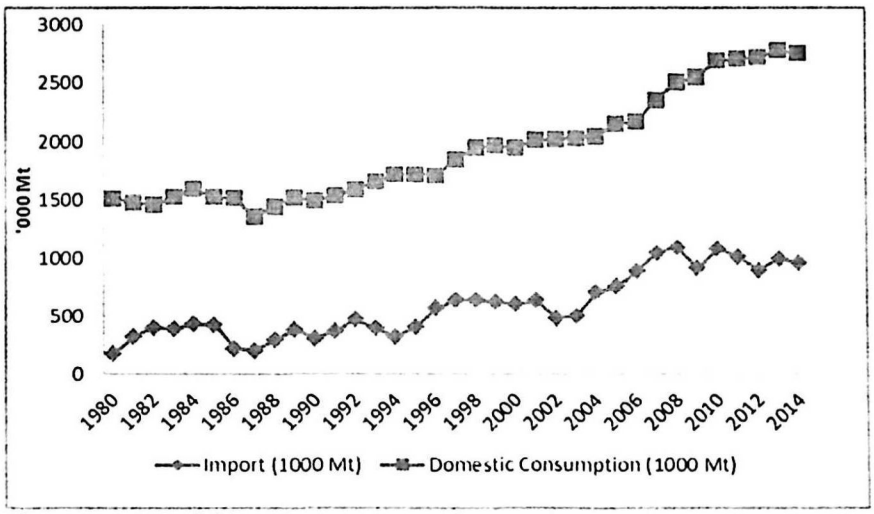
Source: National Agro-food Policy 2011-2020

Malaysia produced 1.678 million tons of rice in the year 2013 and the SSL for local rice in that year was 71% (Agrofood Statistics, 2014). Nonetheless, it was still not sufficient for the local market demand and the shortage was imported from the main producing countries such as Thailand and Vietnam. The imported rice is increase gradually due to the increasing domestic consumption (Figure 1.2) even though the per capita consumption dropped from 147.9 kg in 1960 to 91.7 kg in 2009 (Vengedasalam et al. 2011). The total imported rice for the year 2014 considerably increased from 596,200 tons in 2000 to more than 950,000 tons. This figure dramatically increase due to an increase in total consumption which resulted from an increase in population (Anon, 2014). Currently SSL for local rice is set to exceed 70% to ensure national food security and for the year 2020, the government targeted to achieve 100% SSL to fulfill the rapid growth of Malaysia population (Express, 2015).



Source: United State Department of Agriculture, PSD Online (2015)

Figure 1.1: Area Harvested and Rice Production, 1980 – 2014



Source: United State Department of Agriculture, PSD Online (2015)

Figure 1.2: Domestic Consumption and Import of Rice

1.2.2 Paddy Cultivation in Malaysia

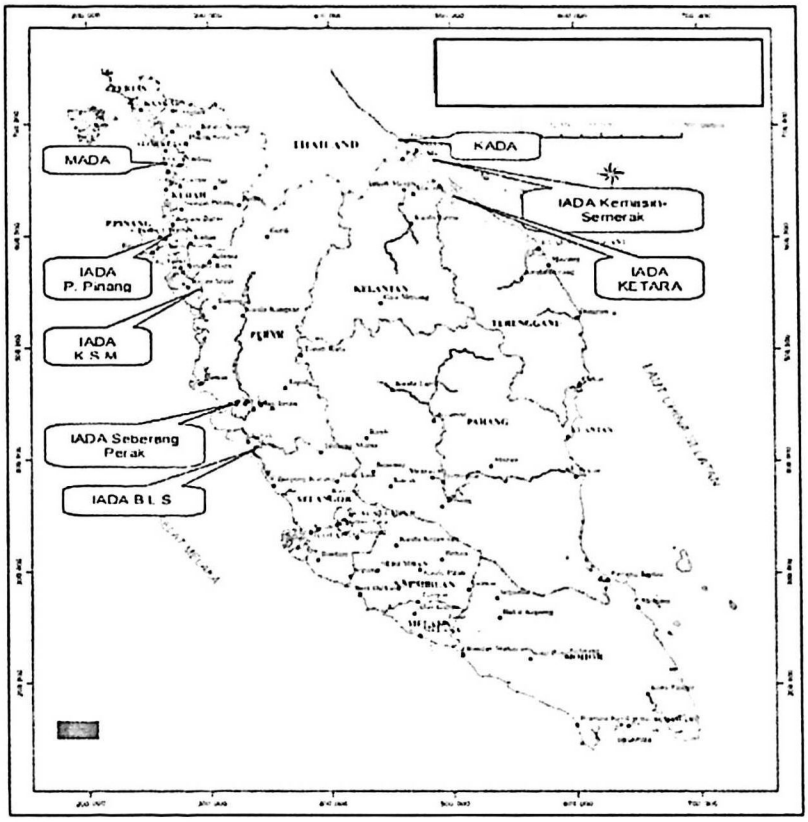
Paddy cultivation in Malaysia consists of two types of paddy which are wetland paddy (*Padi Sawah*) and dry land paddy (*Padi Huma*). 97.8% of rice production in Malaysia was contributed by wetland paddy (Paddy Statistics of Malaysia, 2014). Wetland paddy is the primary paddy type planted in Peninsular Malaysia and the main

contributor for national rice. Water and rainfall are the main requirements for this type of cultivation. Dry land paddy is mostly planted in Sabah and Sarawak and does not depend purely on rainfall. Total paddy production from dry land is low compared to wetland paddy because of limited input such as fertilizer and variety of seed.

Wetland paddy is mostly planted in granary and outside the granary area. Granary area refers to major irrigation schemes recognized by government in the National Agricultural Policy as the main paddy producing area while the areas outside granary are fully dependent on the water from rainfall and the nearest river. The government decided to develop eight main granaries for the purpose of ensuring adequate supplies of local rice that is readily available at a reasonable price. The eight granary areas (Figure 1.1) are located in Peninsular Malaysia which are Muda Agriculture and Development Authority (MADA), Kemubu Agriculture Development Authority (KADA), Integrated Agriculture Development Area (IADA) Kerian, Integrated Agriculture Development Area (IADA) BLS, Integrated Agriculture Development Area (IADA) Pulau Pinang, Integrated Agriculture Development Area (IADA) Seberang Perak, Integrated Agriculture Development Area (IADA) KETARA and Integrated Agriculture Development Area (IADA) Kemasin Semerak (Paddy Statistics of Malaysia, 2014).

Generally, wetland paddy is cultivated in two seasons which are the main season and off season. The main season is between August and February of the following year. This season is also known as wet season because of the northeast monsoon which brings heavy rain and paddy is grown without depending wholly on any irrigation system. Meanwhile, paddy planted in off season normally depends on an irrigation system

because the decrease in the quantity of rainfall due to the Southwest monsoon. Therefore this season has a drier weather and it happens between March and July.



Source: Soil Management Division, Department of Agriculture Peninsular Malaysia

Figure 1.3: Granary Areas in Peninsular Malaysia

Double cropping is mostly applicable for areas inside granary and mini granary since water requirement provided by the irrigation system is sufficient and adequate. Double cropping of paddy in MADA area has been effected since 1970 with the completion of the first stage Muda Irrigation which covered approximately 33,600 hectares for the off season crop (MADA, 2015). However, certain area outside the granary can only be cultivated once a year especially in areas which depend wholly on rainfall.

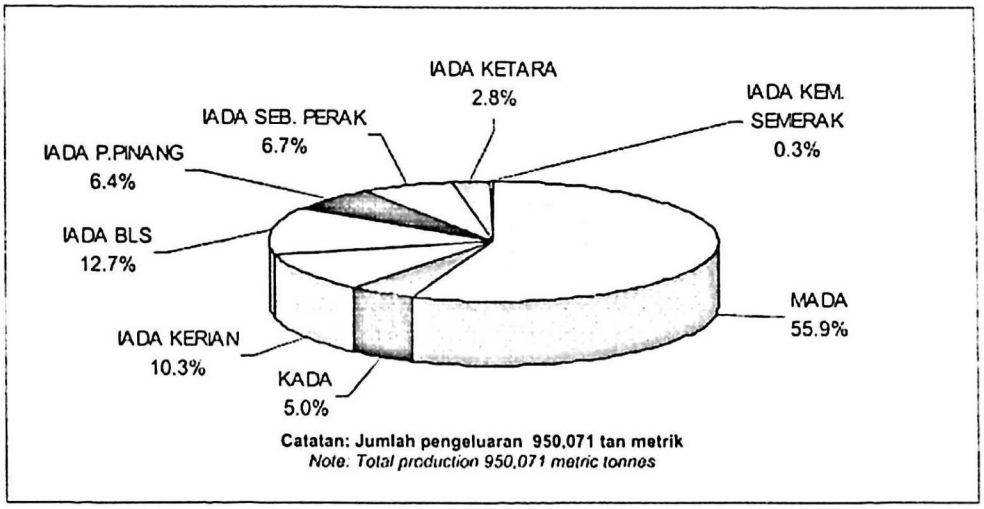
The most popular paddy varieties among farmer in granary and outside the granary area are MR 219 and MR 220. The varieties have the capability to produce high yield, easier to manage and resistant to pest and disease. The granary area contributed about 80.9 % or 0.950 million metric tons to the total paddy production in Peninsular Malaysia and the average production is 5,355 kilograms per hectare (Paddy Production Survey Report for off season 2013, 2014). Nonetheless, the average production is still far from potential yield which is 10 tons per hectare and there is still a wide productivity gap within and between scheme. This target is stated by the government in order to achieve beyond the current national SSL. IADA BLS had the highest average paddy production (Table 1.3) among other granary areas which was 6,362 kilograms per hectare, followed by IADA Pulau Pinang, 5,864 kilograms per hectare and MADA area which was 5,566 kilograms per hectare. Even though MADA area contributed about 55.9% to paddy production (Figure 1.4) and was the highest paddy production contributor among granary areas, the average paddy production was still low if compared to IADA BLS and IADA Pulau Pinang. This area has a potential to increase the paddy production because it has the largest planted area which is approximately 96,000 hectares.

Various actions and effort have been undertaken by the government and related agencies toward increasing paddy production by optimizing the agriculture land use. This is because agriculture land showed a decreasing trend compared to 30 years ago resulting from rapid growth in industrialization and urbanization. Strategies and projects such as breeding new cultivars, upgrading irrigation system, reviewing planting practice and pesticide cycle are undertaken to increase paddy production (Toriman *et al.*, 2013).

Table 1.3: Average Yield of Wetland Paddy by Granary Area, Peninsular Malaysia, Off Season 2013

Granary Area	Off Season 2013		Off Season 2012		Off Season 2011	
	Yield (Kg/Ha)	Area (Ha)	Yield (Kg/Ha)	Area (Ha)	Yield (Kg/Ha)	Area (Ha)
MADA	5,566	95,366	5,296	95,893	5,039	96,526
KADA	3,856	12,362	3,678	18,405	4,095	27,577
IADA KERIAN	4,655	20,972	4,517	26,594	4,116	26,594
IADA BLS	6,362	18,934	6,152	18,934	6,052	18,731
IADA P.PINANG	5,864	10,305	5,812	10,305	5,913	10,305
IADA SEBERANG PERAK	4,613	13,843	4,874	8,233	4,794	8,233
IADA KETARA	5,506	4,876	5,471	4,876	5,637	4,923
IADA KEM SEMERAK	3,180	748	3,912	1,155	2,883	1,870
Total	5,355	177,406	5,116	184,395	4,907	194,759

Source: Department of Agriculture Peninsular Malaysia



Source: Department of Agriculture Peninsular Malaysia

Figure 1.4: Production of Wetland Paddy by Granary Area, Peninsular Malaysia, Off Season 2013

Since 1970, Malaysia Agricultural Research and Development Institute (MARDI) has been directly involved in research development and technology improvement especially in crop agriculture including developing new technology for

agriculture and breeding activities for new paddy cultivar. Those agencies are responsible for producing high quality paddy variety for high yield. 41 paddy varieties (Table 1.4) launched and planted throughout the country. Each variety has its own advantage and tolerance to abiotic or biotic factor.

Table 1.4: Paddy variety planted in Malaysia

Number	Variety	Year	Number	Variety	Year
1	Malinja	1964	22	MR 103	1990
2	Mahsuri	1965	23	MR 106	1990
3	Ria	1966	24	P.HItam 9	1990
4	Bahagia	1968	25	MR 123	1991
5	Murni	1972	26	MR 127	1991
6	Masria	1972	27	MR 159	1995
7	Jaya	1973	28	MR 167	1995
8	SM 1	1974	29	MR 185	1997
9	SM 2	1974	30	MR 211	1999
10	PM 1	1974	31	MRQ 50	1999
11	Setanjung	1979	32	MR 219	2001
12	Sekencang	1979	33	MR 220	2003
13	Sekembang	1979	34	MRQ 74	2005
14	Kadaria	1981	35	MR 232	2006
15	P. Siding	1981	36	MR 220 CL-1	2010
16	Manik	1984	37	MR 220 CL-2	2010
17	Muda	1984	38	MR 253	2010
18	Seberang	1984	39	MR 263	2010
19	Makmur	1985	40	MRQ 76	2012
20	MR 84	1986	41	MR 269	2013
21	MR 81	1988			

Source: Malaysian Agricultural Research and Development Institute (MARDI)

1.2.3 Factors affecting paddy production

Paddy production depends on various factors such as environmental, socio-economic, technology and farm management. These factors affected the paddy

production in different ways and differ significantly among countries, regions and areas within the countries (Alam *et al.*, 2011).

Paddy is cultivated in a variety of water regimes and soil types, such as saline, alkaline, and acid–sulphur soils (IRRI, 2013). Technology improvement was resulted in an increase in agriculture production since three decade. About 84% of the rice-production growth has been attributed to modern farming technologies that have produced semi-dwarf, early-maturing rice varieties that can be planted up to three times per year and are responsive to nitrogen fertilizers (Muthayya *et al.*, 2014).

Paddy production remains highly dependent on the environmental factors such as abiotic and biotic factors (Alam *et al.*, 2011). The abiotic factors consist of uncontrollable nature such as rainfall, temperature, humidity and type of soil (Hui *et al.*, 2012). Biotic factors are also known as limiting factors and refer to biological organism within an ecosystem. This type of factor consists of fungus, bacteria, plant and animal. It might affect the crop either in a beneficial or harmful way (Wisz *et al.*, 2013).

Empirical study on the impact of abiotic factor toward food production has been done throughout the world on various types of crop based on the impact of climate and type of soil. The uncertainties of the climate will affect future crop production especially paddy which is a high water demand plant. The extreme weather such as drought is expected to influence its water use requirement. Though rainfall has a major influence on the yield of crops, yields are not always directly proportional to the amount of rainfall. The suitable amount of rainfall for paddy cultivation in Malaysia is between 200 and 300 mm. Less than the minimum requirement or excess above optimum will reduce the crop yields especially at the end of crop cycle (Alam *et al.*, 2011). In

irrigation area, low rainfall and humidity could be overcome by adequate water supply so they do not affect paddy production directly (Toriman *et al.*, 2013). By using linear mixed regression model, a study on rice production in Tanzania showed that the increased rainfall variability has a negative impact on yield (Rowhani *et al.*, 2011).

Temperature is essential to living organisms because they will be able to function properly and survive at the right temperature. Distribution of crop plant and vegetation is influenced by the temperature of the place. The study done in Philippines indicated that the night time temperature had significant impact on rice production due to yield decrease by 10 percent when the temperature increases 1°C. The relationship between yield attributes and climatic parameters were evaluated by using correlation and partial correlation analysis (Peng *et al.*, 2004). Time series regression model were used to examine the effect of climatic factors on paddy production in IADA North-West Selangor and found that 1% increase in temperature will lead to a 3.44% decrease in current paddy production and a 0.03 % decrease in the next season (Alam *et al.*, 2014).

The Ricardian production function was used to evaluated the climate impact on the farmer income and the result indicated that temperature, rainfall, farm size, educational knowledge, land area and value of labor input have impact on rice production per hectare. Minimal increase in temperature during the main season will increase net revenue by RM 4.78 per hectare while during off season, net revenue is decreased by RM 3.02 per hectare. Increasing rainfall level during off season also will increase net revenue by RM 1.32 and reduce revenue per hectare by RM 1.01 during the main season Masud *et al.*, (2012). The polynomial regression is utilized when involving climatic factors such as temperature and rainfall because these factors have a non-monotonic effect on production (Lobell *et al.*, 2007).

The types of soil determine the availability of nutrients required by the plant. Organic matter and mineral are two types of natural sources of nutrient. Nutrient exchanges between organic matter, water and soil are essential to soil fertility and need to be maintained for sustainable production purposes. Failing to do so will result in broken nutrient cycle and finally a decline in soil fertility (Bot & Benites, 2005). Paddy is suitable planted at the area with type of soil such as clay, silty clay, clay loam and sandy loam. The rate of nitrogen application depends on the level of soil fertility and the research found that higher yield could be achieved in relatively more fertile soil which has higher clay and soil nitrogen content Samy (1982) . Soil reaction is measured by pH (hydrogen ion concentration) of soil and it also affects the growth of plant. The best soil pH for paddy growth is between 5.5 and 6.5 and the acidity of soil can be treated by liming method (Azman *et al.*, 2014). Soil fertility also determines the performance of paddy variety. The result from correlation analysis showed that paddy variety MR 263 is suitably cultivated at low and high fertility soil because it produced high yield at both soil zone compared to paddy variety MR 253 (Othman *et al.*, 2012). Since the level of soil fertility will reflect on paddy yield, identification of the type of soil is very important in order to select the suitable paddy variety.

In order to ensure that paddy has sufficient nutrient during the growing season, sufficient fertilizer application at the right time is very important. The type of fertilizer and amount of application are very crucial at different stage of paddy growth (McKenzie,1998). Nitrogen (N) fertilizer uptake varies among paddy varieties and to achieve high paddy yield it would be best to determine the suitable nitrogen (N) dose for each variety (Chaturvedi, 2005). Paddy variety MR 269 and MR 284 responded to nitrogen in different ways. During off season, both varieties contribute higher yield with Nitrogen application of up to 200 kg/ha but it showed quadratic manner in the main

season due to a decrease in yield when the amount of fertilizer increased subsequently from 100 kg/ha to 200 kg/ha (Ahmad Arif et al., 2014). Excessive fertilizer application also contributes to paddy lodging which reduces paddy yield directly (Suhaimi *et al.*, 1994).

The main challenges for paddy cultivation are pest and disease attacked such as animal, bacteria, fungus and weed which resulted in low paddy yield. Based on the field experiment in China, a correlation analysis showed that sheath blight intensity was positively related to lodging while Nitrogen fertilizer rate and hill density have positive relationship toward sheath blight (Wu *et al.*, 2012). During off season, high rainfall will cause disease and crop pest outbreaks (Masud, *et al.*, 2012). Weedy rice is one of yield limiting factors and has become a serious threat to direct seeding method since there is no suitable herbicide for controlling this problem. MR 220 CL1 and MR 220 CL2 are new varieties that have been launched with new development technology in order to manage weedy rice. This technology managed to control weedy rice and average yield showed increment by 0.76 ton/ha (Azmi *et al.*, 2012).

1.3 Research Objectives

The main purpose of this study is to use the techniques of regression analysis to investigate paddy production of different paddy lots based on environmental and cultivation characteristics in Muda Agriculture and Development Authority (MADA) area between March to July 2014.

Specifically, the objectives of this study are:

1. To identify the type of relationship between paddy production and paddy lots characteristics.
2. To develop an appropriate regression model in order to determine the relationship between paddy production and paddy lots characteristics.
3. To identify factors affecting paddy production based on a regression model.
4. To estimate and predict the mean paddy production based on paddy lots characteristics.

CHAPTER 2

METHODOLOGY

2.1 General Multiple Linear Regression

Generally, relationship between the dependent variable Y and independent variables $X_1, X_2, \ldots, X_{p-1}$ can be represented through a function as:

$$Y_i = \beta + \beta_1 X_{i1} + \ldots + \beta_{p-1} X_{ip-1} + \varepsilon_i \quad i=1,2,\ldots,n \tag{2.1}$$

Where Y_i is the value of the response variable of ith experiment unit,

$\beta_0, \beta_1, \ldots, \beta_p$ are regression parameter, $X_{i1}, \ldots, X_{ip-1}$ are known constant of the i th unit with ε_i is the random error , $N(0,\sigma^2)$.

The general linear regression model (2.1) with normal error term implies that the observations Y_i are independent normal variables, with mean $E(Y_i)$ and constant variance σ^2 Since $E(\varepsilon)= 0$, the response function for regression model (2.1) is:

$$E(Y_i) = \beta_0 + \beta_1 X_{i1} + \ldots + \beta_p X_{ip} \tag{2.2a}$$

and variance of Y_i is:

$$Var(Y_i) = \sigma^2 I = \sigma^2 \tag{2.2b}$$

Hence, $Y_i \sim N(\beta_0 + \beta_1 X_{i1} + \ldots + \beta_p X_{ip}, \sigma^2)$.

General linear model (2.1) can be written in matrix form as below:

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

$\mathbf{Y}$ is a vector of responses (n x 1), $\mathbf{X}$ is a matrix of independent variables (n x p), β is a vector of parameter (p x 1), $\boldsymbol{\varepsilon}$ is a vector of independent normal random variables with expectation $E(\boldsymbol{\varepsilon}) = 0$ and variance covariance matrix $\sigma^2(\boldsymbol{\varepsilon}) = \sigma^2\mathbf{I}$.

Consequently, the expectation and variance covariance matrix of random vector $\mathbf{Y}$ are;

$$E(Y) = \mathbf{X}\beta \tag{2.4a}$$

$$\text{cov}(Y) = \sigma^2\,\mathbf{I} \tag{2.4b}$$

2.2 Method of Estimation for Regression Parameters

Two methods commonly used to estimate the unknown parameters of linear regression model which are the method of least square and method of maximum likelihood.

2.2.1 The Least Square Method

The method of least squares consists of determining the value for unknown parameters that minimize the overall error. Consider the deviation of Y_i from its expected value which are an error

$$\varepsilon_i = Y_i - E(Y_i) \tag{2.5}$$

The deviations are squared and the sum of n squared deviation is denoted by Q.

$$Q = \sum_{i=1}^{n} (Y_i - \beta_o - \beta_1 X_{i1} - \beta_2 X_{i2} - \ldots - \beta_{p-1} X_{i,p-1})^2 \tag{2.6}$$

Hence, the method of least squares gives the value of $\beta = (\beta_0, \beta_1, \ldots, \beta_{p-1})$ which minimize Q can be derived by differentiating Q with respect to β and setting the result equal to zero.

$$\frac{\delta Q}{\delta \beta} = -2 \sum_{i=1}^{n} (Y_i - \beta_o - \beta_1 X_{i1} - \beta_2 X_{i2} - \ldots - \beta_{p-1} X_{i,p-1}) = 0 \tag{2.7}$$

The result from the differentiation will give the normal equation of least squares for linear regression model:

$$\mathbf{X'Xb = X'Y} \tag{2.8}$$

The least squares estimator derived from normal equation is

$$\mathbf{b = (X'X)^{-1}X'Y} \tag{2.9}$$

2.2.2 Maximum Likelihood Estimation

The Maximum Likelihood Estimation (MLE) is applied when the functional form of the probability distribution of the error term is specified. The estimator of the parameters β and σ^2 can be obtained by this method. Essentially, the method of maximum likelihood chooses as estimates those values of the parameters that are most consistent with the sample data. The values of these parameters that maximize the sample likelihood are known as the Maximum Likelihood Estimator (Cole *et al.* 2014). The likelihood function for n observation $Y_1, Y_2, \ldots, Y_n$ is shown as:

$$L(\beta_0,\beta_1,\ldots,\beta_{p-1}) = \prod_{i=1}^{n} \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left[-\frac{1}{2\sigma^2}(Y_i - \beta_o - \beta_1 X_i - \ldots - \beta_{p-1} X_{p-1})\right]$$

$$= \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(Y_i - \beta_o - \beta_1 X_i - \ldots - \beta_{p-1} X_{p-1})^2\right] \quad (2.10)$$

By taking partial derivatives of L with respect to $\beta = (\beta_0,\beta_1,\ldots,\beta_{p-1})$ and equating to zero, equation (2.10) will give the value of β that maximizes the likelihood function L. Equation (2.10) can be written in function $\log_e L$.

$$\frac{\delta(\log_e L)}{\delta\beta} = -\frac{n}{2}\log_e 2\pi - \frac{n}{2}\log_e \sigma^2 - \frac{1}{2\sigma^2}\sum(Y_i - \beta_o - \beta_1 X_i - \ldots - \beta_{p-1} X_{p-1})^2 = 0 \quad (2.11)$$

Hence, the maximum likelihood estimator of β is:

$$\beta = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} \quad (2.12)$$

The maximum likelihood estimator β is identical to least square normal equation.

2.3 Model Building

Model building includes the process of variable selection and model diagnostics. There are two procedures in variable selection that are widely used in model building which are automatics selection and all-possible regression. Diagnostic tools will be used to ensure the model is fit and adequate.

2.3.1 Selection of Independent Variables

Automated selection procedures consist of three methods; forward selection (bottom up approach), backward elimination (top down approach) and stepwise regression (combination of forward and backward). An automatic selection is the best procedure when potential **X** variables is large. Stepwise selection is most widely used and this method develops a sequence of regression model at each step adding and deleting an **X** variables. The criterion for adding or deleting an **X** variables is stated equivalently in terms of F^* statistic.

$$F_k^* = \frac{MSR\left(x_k\right)}{MSE\left(x_k\right)} \tag{2.13}$$

The forward selection technique begins with no variables in the model. For each of the independent variables, this method calculates F^* statistics that reflect the variable's contribution to the model if it is included. The p-values for these F^* statistics are compared to the value of SLENTRY that is specified in the model statement. If no F^* statistic has a significance level greater than the value of SLENTRY, this procedure is stopped. Otherwise, the variables that has the largest F^* statistic will be added to the model. Forward selection technique then calculates F^* statistics again for the variables still remaining outside the model, and the evaluation process is repeated. Thus, variables are added one by one to the model until no remaining variable produces a significant F^* statistic. For this technique, once a variable is entered in the model it will stay.

The backward elimination technique begins by calculating F^* statistics for a model, including all of the independent variables. Then the variables are deleted from the model one by one until all the variables remaining in the model produce F^* statistics

significant at the level of SLSTAY that is specified in the model statement. At each step, the variable showing the smallest contribution to the model is deleted.

The stepwise method is a modification of the forward-selection technique and differs in that variables already in the model do not necessarily stay. Stepwise procedure check all the variables already included in the model and deletes any variable that does not produce an F^* statistic significant at the level of SLSTAY. After the unnecessary variable is deleted, another variable will be added to the model. The stepwise process ends when none of the variables outside the model has an F^* statistic significant at the level of SLENTRY and every variables in the model is significant at the value of SLSTAY, or when the variables to be added to the model is the one just deleted from it.

All possible regression selection procedure fits all possible models at each step with the suggested independent variables that are associated with the criterion. This procedure is suitable for a small group of potentials **X**. If there are k potential **X** variables in the pool, there would be 2^k possible regression models. A few criteria that will be used in comparing the regression models from this procedure such as R_p^2 and C_p.

The R_p^2 criterion calls for the use of the coefficient of multiple determinations R^2 in order to identify the best subset of **X** variables. Subscript p corresponds to the number of parameters or p - 1 **X** variables in the regression function on which R_p^2 is based. The models with the highest R^2 could be considered "best"(Gortmaker *et al.*, 1994). The coefficient of multiple determinations R_p^2 can be defined as: