CHAPTER 3:

RESEARCH METHODOLOGY

3.0 Introduction

In this chapter, the focus will be on the methods applied in conducting this research. Remenyi et al. (1998) defines research methodology as the procedural framework within which the research is conducted. In general, research methodology describes the overall shape and design of this study and the entire data collection process. Apart from that, this section also discusses the model replicated and adapted in the study. The source of data and the analysis techniques chosen will be discussed in details. The essence is to justify the method applied in this study with rationale and valid explanations.

3.1 Data

Secondary data has been used in the study. The study uses quarterly data over the period of quarter 1, 1998 to quarter 4, 2009. Data for Commercial Banks (Non-performing loan (NPL)) had been obtained from Central Banks of Malaysia website. The investigation is using the aggregated data as the idea is to identify the main risk drivers that could have adverse effect over the entire system. The period in arrears for classifying a loan as non-performing by banking institutions was reduced from six months to three months effective from financial year beginning 1 January 1998.[7] Hence, 1998 is the starting point of the data collection and NPL with 3 months end of period record has been taken for analysis. Besides, data for macroeconomic variables (GDP, CPI, KLCI, UE, and HPI) have been downloaded from Central Bank of Malaysia, Department of Statistics Malaysia, Yahoo! Finance and Bloomberg.

[7] BNM Annual Accounts. 1997. *Chapter 4: The Financial Sector, Banking Measures in 1997,* Prudential Measure.

The whole research had been spent with about six months periods. Before the collection of data, 3 month periods (October 2009 to December 2009) had been spent on literature review. All the relevant topics and researches had been studied in details. After that, two months periods had been spent to collect on all the dependent (NPL) and independent data (GDP, CPI, KLCI, UE, and HPI) which cover period of quarter 1, 1998 to quarter 4, 2009. In addition, the preliminary and regression analysis had spent with another two months from the research period before the conclusion is drawn.

3.2 Models Specifications

Multiple linear regressions have been employed for this analysis by using EViews and SAS applications. The general approach to modelling of this model comprises the Gross Domestic Product (GDP), Consumer Price Index (CPI), FTSE Bursa Malaysia KLCI (KLCI), Unemployment Rate (UE) and House Price Index (HPI) is shown below:

$$Y_i = \beta_0 + \beta_1 GDP + \beta_2 CPI + \beta_4 KLCI + \beta_5 UE + \beta_6 HPI \qquad (1)$$

Where:

Y_i = Credit risk (ΔNPL%) for the period of i

β = regression coefficients for each variable

In this research, the exogenous approach which the relationship between macroeconomic variables and the credit risk is assumed to be the same during periods of economic downturn and expansion or during the sample periods.

As shown above, the dependent variable is the credit risk (NPL) and it proposed as change in NPL rate (ΔNPL%), which with quarter-on-quarter basis. For credit-quality regression models, macroeconomics conditions are better being measured by loan performance. *"Banks' stress testing should focus on each of the balance sheet's sources of risk, including the investment portfolio."*[8] Hence, in bank's balance sheet, NPL is one of the key elements to measure the loan performance. Basically, under BNM rules and regulation, GP3: "Classification of non-performing and Provision of Substandard, Bad and Doubtful Debt. Minimum Standards for Classification of loans as non-performing", a loan will be classified as non-performing when the principal or interest is due and unpaid for six months or more from the first day of default. Blaschke, Jones, Majnoni and Martinez Peria (2001) report an example in which the non-performing loan ratio (NPL) is regressed against the nominal interest rate, the inflation rate, the change in real GDP, and the change in the terms of trade.

According to Kimmo Virolainen (2004), previous studies analysing the macroeconomic determinants of banks' loan losses or non-performing loans include Pesola (2001) for the Nordic countries, Kalirai and Scheicher (2002) for Austria, and Delgado and Saurina (2004) for Spain. Typically, these studies find that loan loss provisions are negatively related to GDP growth and positively related to interest rates.

[8] Thomas G. McKenzie, Director and Chief Examiner Office of Examination. 2010. FCA's Stress Testing Expectations for All FCS Institutions.
http://www.fca.gov/apps/infomemo.nsf/59ab19ff3b724b73852569530076c529/e5bfe888bbc71973852576dc006f5720?OpenDocument

The impact of corporate gratitude turns out to be vaguer, even though some studies find a significant positive effect[9].

On the other hand, the 5 independent variables are macroeconomic variables which the hypotheses may form based on the studies which done earlier:

Hypothesis 1:

There is a negative relationship between Gross Domestic Products (GDP) versus Non-Performing Loan (NPL).

GDP is a primary indicator which can measure the health of economy. It equals the aggregate demand of an economy. Lower GDP growth means lower growth in sales. The lower GDP growth is, the harder it is for firms to generate income through sales. Lower income thus increases the possibility that firms cannot meet their obligations and subsequently caused financial distress in the company and caused the people to default as they are unlikely to get their pay easily.

Besides, we noted that inflation is unavoidable in consumer's daily spending, hence, in this research, nominal GDP, which GDP evaluated at current market prices data is being used. These data include all of the changes in market prices that have occurred during the current year due to inflation or deflation.

Hypothesis 2:

There is a positive relationship between Consumer Price Index (CPI) versus Non-Performing Loan (NPL).

[9] Kimmo Virolainen. 2004. Macro stress testing with a macroeconomic credit risk model for Finland. Bank of Finland, Discussion Papers.

Sometimes, the CPI also called cost of living index. It is an inflationary indicator that measures the change in the cost of a fixed basket of products and services. When the cost of living tends to be high, people's income tend to be lesser as they need to spend more money for the same goods which compare to the past. Hence, lower the loan's repayment capacity and the default rate or credit risk is getting higher.

Hypothesis 3:

There is a negative relationship between Kuala Lumpur Composite Index (KLCI) versus Non-Performing Loan (NPL).

KLCI (currently known as FTSE Bursa Malaysia KLCI) is Malaysia's stock market index and it has been used as a proxy for financial market indicator. This index has similar pattern with cyclical trend of economy. The higher the index is indicated the higher the investor's return and hence, lower the credit risk of the portfolios. Therefore, it is expected a negative relationship between KLCI and the credit risk.

Hypothesis 4:

There will be a positive relationship between Unemployment Rate (UE) versus Non-Performing loan (NPL).

Unemployment rate is being used in this study as it been believed that it has direct impact on portfolio's credit risk. In other words, a fall-off in economic growth, associated with increasing unemployment and subsequently falling incomes for households and/or falling income or revenue for corporates, is likely to suggest a deterioration in the 'collectability' of banks' loan book. A person being unemployed will have lower repayment capacity and hence increase the credit risk.

Hypothesis 5:

There will be a negative relationship between House Property Index (HPI) versus Non-Performing Loan (NPL).

Housing property index is an indication of the house price level. Since from the earlier chapter mentioned that from the lending book, almost half of the household debts were long term secured borrowings, hence, this variable added to account for the variability in housing wealth and for the effect of collateralized loans which might have the probability of falling into default. In addition, collateral can represent quality of customers. Higher collateral value which mean higher quality of customer and hence, we can expect a negative relationship between collateral value and loan default.

3.3 Scenario Analysis

In scenario analysis, I postulate the changes in the underlying independent variables (GDP, CPI, KLCI, UE, and HPI) and revalue the portfolio's NPL with these changes. A strong assumption has been incorporated where future financial distress will strongly reassemble those from the historical event. Hence, the historical adverse market movement will be taken into consideration.

In BIS working paper on Stress-testing financial systems: an overview of current methodologies, Macro Sorge did highlight that:

"One of the key decisions is how to calibrate the size of the shocks to use for stress-testing. Setting the hurdle too low or too high might make the whole exercise meaningless. In general, shocks can be calibrated to the largest past movement in the relevant risk variables over a certain horizon (change from peak to trough or deviation from trend) or be based on historical variance (unconditional or conditional). Alternatively, with sufficient data, one can attempt to estimate the joint empirical distribution of past deviations from trend of the relevant risk variables and use its quantiles for simulating the stress scenario."[10]

Potential major events is identified and deemed to have adverse impact towards the country economic performance. Hence, for this research, three stressed economic scenarios i.e. mild recession, moderate recession and severe recession have been identified for each macroeconomics variables. Macroeconomic indicators are indentified that are affected by the events.

In this research, Malaysia GDP has been used as a primary economy indicator to find out the mild recession, moderate recession and severe recession scenario for Malaysia economy in past 13 years. Changes of the GDP have been calculated in order to identify the major dropped when crisis happened. Figure below shows Malaysia GDP Rate from 1997 to 2009, which it clearly shown 3 blips when crisis took place.

[10] Marco Sorge. 2004. *Stress-testing financial systems: an overview of current methodologies.* BIS Working Paper no 165. Pg 5.

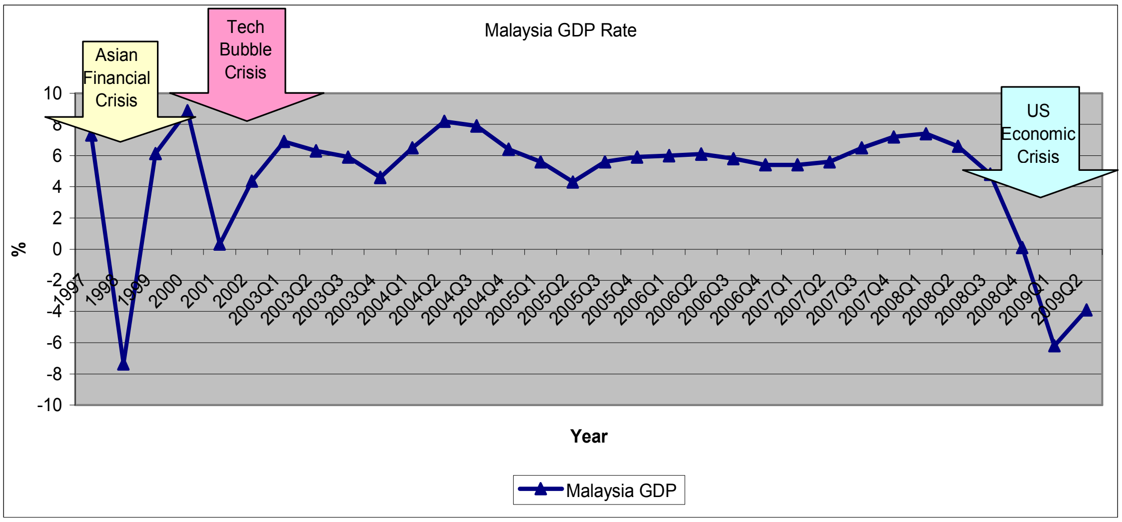

Figure 3.1: Malaysia's GDP from 1997 – 2009.[11]

From figure 3.1 above, it shows that the severe scenario for Malaysia economy was at Asian Crisis, which our GDP has impacted from it at year 1998 and had dropped to -7.36%. Hence, I have taken the following percentage changed from year 1997 for all the independent variables and plug into the regression model that will be tested in the later part in order to find the most severe scenario that the portfolio that will hit in future.

Besides, for moderate recession, recent US economic crisis (2009), which caused by the sub-prime mortgage issue, showed a huge drop in GDP after the Asian Financial Crisis. Hence, all the other independent variables are being taken into account for moderate recession.

---

[11] (i) Swivel Preview. GDP growth rates by country and region 1970-2007. http://www.swivel.com/data_sets/show/1004019.

(ii) Statistic Department, http://www.statistics.gov.my/

Same methods applied to mild scenario, where 2001 Technology Bubble recession being taken into account and the GDP drop was about 8.5%. With the above determination, data for 5 independent indicators of these 3 years shows as follow, these data will be at later stage replace it into the regression to find out the predetermined NPL.

Economic Scenario	GDP	CPI	KLCI	UE	HPI
Severe Recession (1998)	-8.3%	3.0%	21.5%	2.9%	-9.5%
Mid Recession (2009)	-10.4%	2.5%	-13.6%	3.1%	2.5%
Mild Recession (2001)	-9.6%	0.0%	-4.5%	4.0%	2.6%

Figure 3.2: Economic Scenarios for all the independent variables.

3.4 Econometrics Methodology

3.4.1 Unit Root Tests

The above mentioned 5 independent macroeconomic variables are potentially non-stationary. Most real world time series are found to be non-stationary. Hence, this has to be taken into account. Granger and Newbold (1974) argue that a regression analysis using non-stationary variables can easily end up with spurious results and the stationarity or otherwise of a series can strongly influence its behaviour and properties. In fact, even the two variables are totally unrelated, a regression of one on the other could have a high R^2 if two variables are trending over time. So, the natural first step is therefore to investigate the unit root properties of the variables under investigation.

There are many tests available in unit root test such as the Dickey-Fuller (DF) Test, Augmented Dickey-Fuller (ADF) Test, Philips-Perron (PP) Test, Dickey-Fuller Generalised Least Squares (DF-GLS), Ng and Perron (NP) and KPSS or Kwiatkowski, Phillips, Schmidt, Shin (KPSS, 1992) among others in order to test for the data stationarity.

In this research, Augmented Dicky-Fuller (ADF) test has been used to test for unit-root. It has been chosen for its simplicity of hypothesis and ease of understanding.

The normal DF Unit Root Test is based on the following three regression forms:

1. Without Constant and Trend; $\Delta Yt = \delta Yt\text{-}1 + \mu t$ (2)

2. With Constant; $\Delta Yt = \alpha + \delta Yt\text{-}1 + \mu t$ (3)

3. With Constant and Trend $\Delta Yt = \alpha + \beta T + \delta Yt\text{-}1 + \mu t$ (4)

The hypothesis is:

H0: $\delta = 0$ (unit root)

H1: $\delta \neq 0$ (series is stationary)

Decision rule:

If t* > ADF critical value, ==> do not reject null hypothesis, i.e., unit root exists.

If t* < ADF critical value, ==> reject null hypothesis, i.e., unit root does not exist.

Each of the equation will be run separately depending on the data specification.

To overcome the problem of autocorrelation in the basic DF test, the test can be augmented by adding various lagged dependent variables which would produce the following test:

$$\Delta y_t = (\rho - 1) y_{t-1} + \alpha_i \sum_{i=1}^{m} \Delta y_{t-i} + u_t \qquad (5)$$

The correct value for m (number of lags) can be determined by reference to a commonly produced information criteria such as the Akaike criteria or Schwarz-Bayesian criteria. The aim is to maximize the amount of information. The DF and ADF test can also include a drift (constant) and time trend.

3.4.2 Autocorrelation

Autocorrelation is the cross-correlation of a signal with itself and it is a systematic pattern in the errors that can be either attracting or repealing autocorrelation. Informally, it is the correspondence between observations as a function of the time separation between them. James Durbin and Geoffrey Watson (1950) came out a Durbin-Watson statistic which to detect the presence of autocorrelation in the residuals from a regression analysis.

If e_t is the residual associated with the observation at time t, then the test statistic is

$$d = \frac{\sum_{t=2}^{T}(e_t - e_{t-1})^2}{\sum_{t=1}^{T} e_t^2}.$$

Since d is approximately equal to 2(1-r) and it always lie between 0 and 4, where r is the sample autocorrelation of the residuals, d = 2 indicates no autocorrelation. In addition, if the Durbin–Watson statistic is substantially less than 2, there is evidence of positive serial correlation. As a rough rule of thumb, if Durbin–Watson is less than 1.0, there may be cause for alarm. Small values of d $(d<2)$ indicate successive error terms are, on average, close in value to one another, or positively correlated. On the other hand, if $d > 2$, it is a successive error terms, on average which much different in value to one another, i.e., negatively correlated.

3.4.3 Multicollinearity

In this research, multicollinerity issue will be solved before stepwise regression method is performed in order to select the correct model which consists of the useful X variables.

In multiple regressions, independent (X) variables tend to have multicollinearity. Dr. Harvey Motulsky (2002) explained that:

"when two X variables are highly correlated, they both convey essentially the same information. In this case, neither may contribute significantly to the model after the other one is included. But together they contribute a lot. If you removed both variables from the model, the fit would be much worse. So the overall model fits the data well, but neither X variable makes a significant contribution when it is added to your model last. When this happens, the X variables are collinear and the results show multicollinearity."

In other words, the greater the multicollinearity, the greater the standard errors in the data. Confidence intervals for coefficients tend to be very wide and t-statistics tend to be very small once high multicollinearity is present. Coefficients will have to be larger in order to be statistically significant, i.e. it will be harder to reject the null hypothesis when multicollinearity is present.

Besides, when two variables are highly and *positively* correlated, their slope coefficient estimators will tend to be highly and *negatively* correlated. For instance, b1 is less than β1, b2 will tend to be greater than β2. Furthermore, a different sample will likely produce the opposite result. If researcher overestimates the effect of one parameter, he or she will tend to underestimate the effect of the other. Thus, coefficient estimates tend to be very shaky from one sample to the next.

Hence, the nonexistence of multicollinearity is essential to a multiple regression model. In a regression model, there will be a high variance explained (R^2). The higher the R^2, the better the model is. However, parameter estimates are all inflated if collinearity exists, probably the variance, standard error. In this research, Variance inflation factor (VIF), which is common way for detecting multicollinearity, will be used to find out the possible X variables to be thrown out, which a general rule is that the VIF should not exceed 10 (Belsley, Kuh, & Welsch, 1980). Mathematically speaking, the VIF option in the regression procedure can be interpreted in the following ways:

$$VIF = 1/(1\text{-}R\text{-}square) \qquad (6)$$

3.4.4 Stepwise Regression

In this research, stepwise regression will be used to select useful subsets of variables and to evaluate the order of importance of variables. It is known as a common approach to select a subset of variables from a complex model. Basically, it will throw out those X variables that cannot contribute much to the variance explained. Many researchers like June (1997) and Leigh (1996) employed these techniques to determine the order of predictors by its magnitude of influence on the outcome variables.[12]

Marcos Souto et al, had done their research on Brazilian Banks and they have used a structural VAR, which they interested in visualizing the reaction functions to shocks in the main factors. To determine which variables to utilize, they had ran a stepwise OLS regression over the aggregated data and found that short-term and long-term domestic real interest rate are the most significant factors. In addition, they had estimated the VAR with two lags.[13]

In order to investigate the impact of variables over the entire Commercial Bank's NPL, the main risk drivers that could have adverse effect over the entire commercial banking system has to be determined.

Efroymson, MA (1960) mentioned that in stepwise regression technique, three main approaches are forward selection, backward elimination and combination of forward selection and backward elimination. Forward selection involves starting with no variables in the model, trying out the variables one by one and including them if they are 'statistically significant'. On the other hand, backward elimination involves starting with all candidate variables and testing them one by one for statistical significance, deleting any that are not significant. Besides, there are methods which combine the above and testing at each stage for variables to be included or excluded.

---

[12] Dr Chong Ho Yu. 2008. Multi-collineartity, Variance Inflation and Orthogonalization in Regression. The problem of too many variables.
http://www.creative-wisdom.com/computer/sas/collinear_stepwise.html

[13] Marcos Souto, Benjamin M. Tabak, and Francisco Vazquez. *Linking Financial and Macroeconomic Factors to Stress-Test Credit Risk Indicators for Brazilian Banks.*

For this research, third approach which stepwise regression process with combination of forward selection and backward elimination is used to filter out the variables of least significance for the entire Commercial Banks in Malaysia, until a small set of explanatory variables of statistical significance remains in the model. Hence, in order to determine which variables to utilize, stepwise regression over the aggregated data has been run. The SAS application is then used to search through the independent variables in order to build up the regression equation.

Besides, the lag structure of the independent variables takes into account the plausible delay with which macroeconomic shocks affect the banks. Each independent variable (GDP, CPI, KLCI, UE, and HPI) is estimated with two quarter lags.

3.5 Sources of Data

Since all the data is secondary, the central bank sites, department of statistics Malaysia, Yahoo! Finance and Bloomberg will be the primary source of information and statistics for this analysis.

The summary of the sources can be seen from the following table:

Table 3.1: Sources of Data

Database	**Website**
NPL,CPI, HPI	http://www.bnm.gov.my
GDP	Bloomberg
UE	http://www.bnm.gov.my, www.statistic.gov.my
KLCI	http://finance.yahoo.com

3.6 Summary

In summary, this chapter describes the methods to be adopted in conducting this research. From development of hypotheses through various literatures support to the design of the linear regression by adapting the previous studies and the data analysis manner, the mechanism of the study is proven reliable and effective in answering the research questions. In addition, the proposed sample size is justifiable pursuant to Hair et al. (1998) and Coakes and Steed's (2005) recommendations. In the next chapter, the result of the survey will be evaluated in detail.