CHAPTER III

RESEARCH METHODOLOGY

A. Research Object

The object of the research is Java which consists of six provincies, they are DKI Jakarta, West Java, Central Java, DIY, East Java and Banten. This study uses sample data from six provinces with annual data from 2010 until 2016 as a result of the publication of Bappenas (Badan Perencanaan Pembangunan Nasional), DJPK (Direktorat Jenderal Perimbangan Keuangan) and BPS (Badan Pusat Statistik).

B. Types and Sources of Data

1. Data Type

   This type of research is a quantitative research which refers to the systematic empirical investigation of any phenomena via statistical, mathematical or computational techniques. The objective of quantitative research is to develop and employ mathematical model, theories and/or hypotheses pertaining to phenomena. In making this research, researcher use a descriptive quantitative method. Involves collecting and converting data into numerical form so that statistical calculations can be made and conclusions drawn.

2. Data Source

The data used are panel data which is a combination of cross-section data of six provinces and time series data or time series year 2010-2016 of Java. The data used in this research is secondary data. Secondary data is data obtained through publications from other sources such as institutions, research journals, newspapers, magazines, internet, books and other literature. In this study, secondary data were obtained from Bappenas (Badan Perencanaan Pembangunan Nasional), DJPK (Direktorat Jenderal Perimbangan Keuangan) and BPS (Badan Pusat Statistik), research journals, thesis and various literature publications related to this study. The data used are as follows:

Variable	Data Source
Gross Regional Domestic Product	BPS period of 2010-2016
Government Expenditure on Health Sector	DJPK APBD health sector function budget period of 2010-2016
Government Expenditure on Education Sector	DJPK APBD education sector function budget period of 2010-2016
Human Development Index (HDI)	BPS period of 2010-2016
The Number of Working Labor Force	BPS period of 2010-2016

C. Research Variable

The research variables are all things that any form defined by the researchers to be studied in order to obtain information about it, then be

deduced. This study used five variables study that consists of one dependent variable and four independent variables.

1. Dependent Variable

A dependent variable is a variable that is influenced or which become due to the existence of independent variables. This study used economic growth (GRDP) as the dependent variable.

2. Independent Variable

Independent variable is a variable that affects or that causes the change or the incidence of the dependent variable. This study used four independent variables, as follows:

a. Government expenditure on health sector (HEALTH)
b. Government expenditure on education sector (EDUC)
c. Human Development Index (HDI)
d. The number of working labor force (WRKG)

D. Data Collection

Data collection methods used in this study is documentation method, a method which aims to get the data associated with variables research through various sources of literature and institutions. The literature sources used in this study were data publication by Bappenas (Badan Perencanaan Pembangunan Nasional), DJPK (Direktorat Jenderal Perimbangan Keuangan) and BPS (Badan Pusat Statistik), research journals, thesis, articles on the internet, and books. Secondary data was collected through documentation of

data that have been published by various agencies and literature relating to this study.

E. Operational Definition of Research Variable

The research variables used are health government expenditure, education government expenditure, HDI (Human Development Index), the number of working labor force, GRDP (Gross Regional Domestic Product). The operational definition of each variable used in this study are as follows:

1. Economic Growth

   Economic growth is the enhancement of a country's ability to develop economic activities or economic conditions on an ongoing basis through a process of increasing the production capacity of an economy embodied in the form of an increase in national or regional income. The data of economic growth used in this research is GRDP (Gross Regional Domestic Product) at Constant Market Prices by six provinces in Java in the period of 2010-2016. The data of economic growth used is obtained from Badan Pusat Statistik with measurement unit is billion rupiah.

2. Government Expenditure on Health Sector

   Government expenditure on health sector is the allocation of The Indonesian Budget/ Regional Government Budget funds spent by the government in order to fulfill the needs of the health sector. The health sector expenditure data in this study used data on the realization of government health spending in Java including the six provinces on the

period 2010-2016. Government expenditure data in the health sector is obtained from the official website of DJPK (Direktorat Jenderal Perimbangan Keuangan). The measurement unit of government expenditure data on health sector used is in rupiah.

3. Government Expenditure on Education Sector

   Government expenditure on education sector is the allocation of The Indonesian Budget or Regional Government Budget funds issued by the government in education. The education sector expenditure data in this study used data on the realization of government spending in Java, which consists of six provinces in the period 2010-2016. Government expenditure data in the education sector is obtained from the official website of DJPK (Direktorat Jenderal Perimbangan Keuangan). The measurement unit of government expenditure data on education sector used is in rupiah.

4. Human Development Index

   HDI (Human Development Index) is a composite index calculated as a simple average of three basic indexes, which are health, education, and expenditure indexes. The HDI data used in this study is HDI data in Java which consists of six provinces with the period of 2010-2016. The HDI data used is obtained from Badan Pusat Statistik (BPS) with the measurement unit is in index.

5. Working labor force

Working labor force is people who are working within the age of 15-65. The number of labor force employed data was obtained from Badan Pusat Statistik (BPS) in period of 2010-2016 and the unit of measurement used is person.

F. Hypothesis Testing and Data Analysis

1. Analysis Method

For the sake of human resources represented by the government's performance in health and education sectors, Human Development Index and working labor force on economic growth in Java, the authors used panel data analysis. According to Nachrowi (2006) as cited in Atahrim (2013), the analysis using panel data is a combination of time series and cross-section. In accordance with the panel data model, the model equations using cross-section data can be written as follows:

$Yi = \beta_0 + \beta_1 Xi + \varepsilon i$; I = 1,2,...N ..... (3.1)

Where N is the number of cross-section data since panel data is a combination of time-series and cross-section, the model can be written by:

$Yit = \beta_0 + \beta_1 Xit + \varepsilon it$ ..... (3.2)

I = 1,2,...N ; t = 1,2,....T

N = the number of observations

T = period

N x T = the number of panel data

This research concerned the influence of human resource investment which is represented by government expenditure in health and education sector, Human Development Index and labor force who worked on economic growth in Java, used time-series data for 7 years which represented annual data from 2010 to 2016 and cross-section data as much as 6 provincial data in Java that resulted in 42 observation equation function panel data that can be written as follows:

$$LN\_GRDPit = \beta_0 + \beta_1 LN\_HEALTH_{it} + \beta_2 LN\_EDUC_{it} + \beta_3 HDI_{it} + \beta_4 LN\_WRKG_{it} + \mu_{it} \quad \text{..... (3.3)}$$

Where :

Y = LN GRDP at Constant Market Prices by six provinces in Java in the period of 2010-2016

HEALTH = LN Government Expenditure on Health Sector by six provinces in Java in the period of 2010-2016

EDUC = LN Government Expenditure on Education Sector by six provinces in Java in the period of 2010-2016

HDI = Human Development Index by six provinces in Java in the period of 2010-2016

WRKG = LN Working labor force by six provinces in Java period of 2010-2016

β_0 = Intercept

$\beta_1,\beta_2,\beta_3,\beta_4$ = Regression coefficient of independent variable

i = Cross section unit

t = Time series unit

μ = Error t

The existence of unit differences and the value of independent variables in the equation cause the regression equation should be made natural logarithm model or LN. Therefore, a logarithmic function is used to solve the equations that rank is unknown. According to Gujarati (2007: 637) as cited in Atahrim (2013), the advantages of using panel data versus time-series data and cross-section data are:

- Panel data estimation can indicate the existence of heterogeneity in each individual.
- Using panel data, the data is more informative, more varied, reduces the collinearity between variables, increases the degree of freedom and more efficiently.
- The panel data study is more satisfactory for determining dynamic change than with repeated studies of cross-sections.
- Panel data more detect and measure effects that simply can not be measured by time-series or cross section data.
- Panel data helps the study to analyze more complex behaviors.
- Panel data can minimize bias generated by individual or company aggregation due to more data units.

2. Regression Method with Data Panel

A regression model with panel data generally results in difficulty in model specifications. The residual will have three possibilities: residual time series, cross-section or both combination. From three approaches to panel data method, two common approaches used to estimate a regression model with panel data are fixed effect model approach and random effect model approach. In order to determine the method between pooled least square and fixed effect using F test, while Hausman test is used to choose between random effect or fixed effect. In addition, in estimation models of panel data regression models, there are F-test, Chow Test and Hausman test. Below are described three approaches used in panel data:

a) Pooled Least Square (PLS)

This method is also known as Common Effect Model (CEM). In this method, the model assumes that the existing aggregate data indicates the actual conditions in which the intercept value of each variable is the same and the coefficient slope of the variables used is identical for all cross-section units.

In this approach, the unit of cross section and time series are all treated the same, then regressed using ordinary least squares method that will produce the equation with an intercept and the coefficients of the independent variables are constant for each unit.

The weakness in the common effect model is the existence of a mismatch model with the actual state, where the condition of each

object is different from each other even one time will be very different from the condition of the object at another time (Winarno as cited in Atahrim, 2013).

b) Fixed Effect Model

The constraint of pooled least square is the assumption that assumes the same intercept and slope coefficients for each cross-section unit and time series. In order to overcome this, another approach is to use dummy variables to allow for changes in the intercept of each cross-section unit and time series. This approach is called the Fixed Effect Model or Least Square Dummy Variable. The possible assumption of intercept and slope coefficients that occur is as follows:

1. Intercept for each cross-section unit is different, constant slope coefficient.
2. Intercept for each unit of cross-section or time series is different.
3. Intercept and slope coefficients for all individuals or cross-section units vary.

The large use of dummy variables can be a disadvantage for this model because it causes a low degree of freedom, the presence of variables that do not change with time, the possibility of multicollinearity, as well as the assumption of error used, which ultimately affects the coefficients of the estimated parameters.

c) Random Effect Model

This model was formed to overcome the weaknesses in the fixed effect model by entering different parameters between cross-section units and time series into error term. This approach is called the Random Effect Model or Error Component Model and assumes that the component is an error. This model was formed to overcome the weaknesses in the fixed effect model by entering different parameters between cross-section units and time series into error term. This approach is called the Random Effect Model or Error Component Model and assumes that the error components between units of cross section and times series are not correlated with each other.

The main assumption of this random effect model is that the individual error components are not correlated with each other, are not autocorrelated between the cross-section and time series units and also assume that the error individually is not correlated with the combination error. According to Atahrim (2013), this approach tries to improve the efficiency of the Ordinary Least Square modeling process, the interrupts between the unit cross section and time series are taken into account so that the method used is Generalized Least Square (GLS).

3. Selection of Panel Data Method

In a panel of data processing, this mechanism of determining the method for selecting the appropriate panel data that is by comparing the

approach Pooled Least Square method with Fixed Effect Model approach first. If the results show the Pooled Least Square approach model accepted, then Pooled Least Square approach will be analyzed. If the model Fixed Effect model is accepted, then make the comparison again with Random Effect Model approach. In order to carry out which model is more suitable to be used, then tested are as follows:

a. Chow Test

Chow test is a test that will be used to determine whether the Pooled Least Square (PLS) or Fixed Effect Model (FEM) model will be selected for data estimation. This test can be performed with a restricted F-Test or Chow test. In this test carried out by the following hypothesis:

H_0 : PLS Model (Restricted)

H_1 : Fixed Effect Model (Unrestricted)

The basis of rejection of the above hypothesis is to compare the calculation of F-statistic with F-table. The comparison is used if the result of F-statistic is greater (>) than F-table, then H0 is rejected which means the most appropriate model used is Fixed Effect Model. On the other hand, if F-statistic is smaller (<) than F-table, then H0 is accepted and the model used is Common Effect Model (Widarjono as cited in Basuki and Yuliadi, 2015). The calculation of F-statistic is obtained from Chow Test with a formula (Baltagi cited in Basuki and Yuliadi, 2015), as follows :

$$F = \frac{\frac{(\mathrm{SSE}_1 - \mathrm{SSE}_2)}{(n-1)}}{\frac{\mathrm{SSE}_2}{(nt-n-k)}} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots.(3.4)$$

Where:

SSE_1 = Sum Square Error from Common Effect Model

SSE_2 = Sum Square Error from Fixed Effect Model

n = Number of cross-section data

t = Number of time-series data

k = Number of independent variable

While F-table is obtained from:

F-tabel = {α : df (n-1, nt – n – k)}

Where:

α = Level of significance used

n = Number of cross-section data

nt = Number of cross-section data x number of time-series data

k = Number of independent variable

b. Hausman Test

This test is used to determine whether the fixed effect or random effect model will be selected. This test is performed with the following hypothesis:

H_0 : The model follows Random Effect

H_1 : The model follows Fixed Effect

Rejection basic H0 using Chi-Square statistical considerations, if Chi-Square statistic > Chi-Square table, then H0 is rejected (model used is Fixed Effect).

4. Classical Assumption Testing

Before performing data analysis then data is tested according to classical assumption, if there is a deviation from classical assumption used non-parametric statistic test. Conversely, if the classical assumption is fulfilled when using parametric statistics to obtain a good regression model, then the regression model must be free from multicollinearity, autocorrelation, and heteroscedasticity and the resulting data must be normally distributed. Thus, in order to test the classic assumption deviation used several tests as follows:

a. Normality Test

Normality test aims to test whether in the regression model the intruder or residual variable has a normal distribution or not. As it is known that the t-test and F-test assume that the value of the residuals follows a normal distribution. According to Suliyanto (2005) as cited in Atahrim (2013), if this assumption is violated, the statistical test becomes invalid.

There are several methods to determine whether the residual distribution is normal or not, which are Jarque-Bera (J-B) Test and graph method. In this method used J-B Test, if J-B statistic < value of X2 (Chi-Square) table, then the residual value is normally distributed.

b. Multicollinearity test

Multicollinearity test aims to test whether the regression model found a correlation between independent variables. Multicollinearity means there is a significant correlation between two or more independent variables in the regression model. Tests on the presence or absence of multicollinearity are performed by looking at the correlation coefficient between variables. According to Basuki and Yuliadi (2015: 144), multicollinearity is the existence of a linear exact relationship between explanatory variables. Multicollinearity is assumed to occur when the value of R^2 is high, the value of t of all explanatory variables is not significant, and the F value is high.

The consequences of multicollinearity according to Basuki and Yuliadi (2015: 144) are as follows:

- Standard errors tend to increase with increasing levels of correlation between variables.
- Because of the magnitude of standard error, the confidence interval for the relevant population parameters tends to be greater.
- Estimated coefficient and standard error regression become very sensitive to slight changes in the data.

The consequence of multicollinearity is the invalidity of variable significance and the magnitude of variable coefficients and constants. A method to detect the presence or absence of multicollinearity problem can be done by partial correlation method between

independent variables. As a rule of thumb, if the correlation coefficient is high enough above 0,9 then it can be concluded that there is multicollinearity in the model. Conversely, if the correlation coefficient is less than 0,9 then the model does not contain multicollinearity elements.

c. Autocorrelation Test

According to (Suliyanto, 2005: 64) as cited in Atahrim (2013), the autocorrelation test aims to determine whether there is a correlation between a series of observational data described by time series data or cross section data. Autocorrelation arises because consecutive observations over time are related to each other. This problem arises because residuals are not free from one observation to another. It is often found in the time series data.

The consequences of autocorrelation by Basuki and Yuliadi (2015: 149) are as follows:

1. The appraiser is inefficient, the confidence interval widens unnecessarily and the test of significance is less powerful.
2. The estimated residual variation is too low.
3. Testing the meaning of t and F is no longer valid and gives a misleading conclusion about the statistical significance of the estimated regression coefficients.
4. The estimator gives a population description that deviates from the actual population value.

Autocorrelation is the relationship between residuals in one observation with another observation. The consequences of autocorrelation are biased with a smaller variance than the actual value, so the value of R^2 and the F-statistic that is produced tends to be very excessive or overestimated. Method to detect whether or not there is an autocorrelation problem with Durbin-Watson test. The advantage of the D-W test in detecting autocorrelation problems is because the test is based on estimated residuals. According to Basuki and Yuliadi (2015: 153), to detect the presence of serial correlation by comparing the value of Durbin Watson Durbin Watson statistic is calculated by statistical tables, namely:

a. If the probability of F-statistic > 0.05, then the hypothesis is accepted where the model is free from serial correlation.
b. If the probability of F-statistic < 0.05, then the hypothesis is rejected where the model has a serial correlation.

d. Heteroscedasticity Test

Heteroscedasticity test aims to test whether the regression model is formed happened inequality residual variance of the regression model. A good data is a data that homoskedasticity. Homoscedasticity occurs when the variant variables in the regression model have the same or constant value. Heteroscedasticity means variant of non-constant disturbance variable. The problem of heteroscedasticity, therefore,

more often present in the cross-section data than in time-series data. If the variant of a residual observation to another observation remains the same, it is called heteroscedasticity. There are consequences if the residuals are heteroscedasticity:

1. The least squares method estimate does not have a minimum variant (no longer Best), so it only meets the characteristics of LUE (Linear, Unbiased and Estimator). Nevertheless, the least squares method estimator is still linear and unbiased.
2. The standard error calculation can no longer be trusted, because the variant is not minimum. Non-minimum variants lead to inefficient regression estimates.
3. Hypothesis testing based on t-test and F-test can no longer be trusted because the error standard is not reliable.

The method used to detect whether there is a problem of heteroscedasticity can be performed by Park test. Park test is performed by regression of residual functions. If the independent variable is not statistically significant, it can be concluded that the model formed in the regression equation does not contain the problem of heteroscedasticity.

5. Statistics Testing

The statistical test consists of testing the partial regression coefficient (t-test) for each independent variable, testing the coefficient of regression

simultaneously (F-test) and testing the coefficient of determination Goodness of fit test (R^2). According to Gujarati (2003) as cited in Suryanto (2011), states that the significance test is a procedure used to test the truth or error of the null hypothesis of the sample results.

a. Individual Parameter Significance Testing (t-test)

Individual parameter significance test (t-test) is conducted to see the significance of independent variables affect the variable is not bound individually and consider other variables constant. Through the t-test it can also show how far the influence of an individual explanatory or independent variable in explaining the variation of the dependent variable. The null hypothesis (H_0) to be tested is whether a parameter (bi) is equal to zero, or:

$$H_0 : bi = 0$$

It means, if an independent variable is not a significant explanatory on the dependent variable. the alternative hypothesis (H_a) parameter of a variable is not equal to zero, or:

$$H_a : bi > 0$$

According to Imam Gozhali (2009) cited in Suryanto (2011), it means, these variables are significant explanatory dependent variable.

The value of t -statistic can be calculated by the formula:

$$t = \frac{\beta_1 - \beta_i *}{\text{SE}(\beta_i)} \dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots. (3.5)$$

Where:

β_1 = Estimated parameters

$\beta_i *$ = Value of β_i in hypothesis

$SE(\beta_i)$ = Standar error of β_i

At 10 percent level of significance with the test used is as follows:

- If t-statistic > t-table then H_0 is rejected, meaning one of the independent variables influence the dependent variable significantly.
- If t-statistic < t-table then H_0 is accepted, meaning one of the independent variables does not affect the dependent variable significantly.

b. Simultaneous Significance Testing (F-test)

F-test is performed to find out whether the independent variables as a whole are statistically significant in influencing the dependent variable or not. If the value of F-statistic is greater than the value of F-table, then the independent variables as a whole affect the dependent variable.

The hypothesis used is as follows:

$H_0 : \beta_1 = \beta_2 = \beta_3 = ... = \beta_i = 0$

H_1 : At least one regression coefficient is not equal to zero

The value of F-statistic is formulated as follows:

$$x = \frac{\frac{R^2}{(K-1)}}{\frac{1-R^2}{(N-K)}} \dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots (3.6)$$

Where:

K = The estimated number of parameters includes constants

N = Number of observation

At the significance level of 10 percent, the testing criteria used are as follows:

1. H_0 is accepted and H_1 is rejected if F-statistic < F-table, which means explanatory variables simultaneously or together do not affect the variable described significantly.
2. H_0 is rejected and H_1 is accepted if F-statistic > F-table, which means explanatory variables simultaneously or together affect the variable described significantly.

c. Coefficient of Determination Test (Adjusted R^2)

R^2 or the coefficient of determination measures the goodness of fit of the regression equation, ie giving the proportion or percentage of the total variation in the dependent variable described by the independent variable. According to Atahrim (2013), stated that the coefficient of determination (R^2) essentially measures how far the ability of a model in explaining variations in the dependent variable. The value (R^2) is

between zero and one ($0 \leq R^2 \leq 1$). The fit of the model is said to be better if R^2 is almost 1. It means that if the value is close to zero, then the ability of the independent variable to explain the dependent variable is very limited. Whereas, if the model fit value is said to be better if R^2 is closer to 1, it means that the independent variables are increasingly providing almost all the information needed to predict the dependent variable.

The coefficient of determination is formulated as follows:

$$R^2 = \frac{\Sigma(\hat{Y} - \bar{Y})^2}{\Sigma(\bar{Y} - Y)^2} \dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots.\dots (3.7)$$

The fundamental weakness of determination usage is biased towards the number of independent variables included in the model, because each addition of one independent variable R^2 value must increase no matter whether the variable has a significant effect on the dependent variable. So that, many researchers suggest using adjusted R^2 values when evaluating the best regression model.