CHAPTER III

RESEARCH METHODOLOGY

A. Research Object

This study uses data from the Indonesian Family Survey (IFLS). IFLS is a survey conducted to provide data that can be used to study the behaviour of Indonesian households and communities. This longitudinal survey has been carried out continuously since 1993 and has produced five waves to date. RAND Corporation, a non-profit research organisation, conducts the IFLS survey in collaboration with other institutions. Its procedures were well evaluated and approved by IRBs (Institutional Review Boards) in the United States and the University of Indonesia (for IFLS1 and IFLS2) and the University of Gadjah Mada (for IFLS3, IFLS4 and IFLS5) in Indonesia.

Through the stratified sampling scheme, samples from 13 Indonesian provinces were collected, including: four provinces on Sumatra (North Sumatra, West Sumatra, South Sumatra, and Lampung), five provinces on Java (DKI Jakarta, West Java, Central Java, DI Yogyakarta, and East Java), and four provinces on remaining major islands (Bali, West Nusa Tenggara, South Kalimantan, and South Sulawesi) (Strauss et al., 2016).

The number of samples in each waves of IFLS are as follows: IFLS 1 in 1993 is 7,200 households or 22,000 individuals, IFLS 2 in 1997 is 7,500 households or 23,000 individuals, IFLS 3 in 2000 is 10,400 households or 39,000 individuals, IFLS 4 in 2007 is 13,535 households and 44,103 individuals, and

IFLS 5 in 2014 is 16,204 households or 50,148 individuals. The data are provided for the same individuals from time to time. It makes the study of behavioural dynamics of each individual, household or community is feasible to do using the IFLS survey.

B. Data Type

According to the data source, the type of data used in this study is secondary data, where data is obtained indirectly from previously available sources, in this case, the author uses data from the Indonesia Family Life Survey in 1993, 1997, 2000, 2007, and 2014.

As for its characteristics, this study uses a type of quantitative data in which data from observations are produced/converted into measured numbers that can interpret certain phenomena (United States General Accounting Office, 1992).

C. Data Collection Technique

The IFLS data which is obtained from the survey using stratified sampling method, then be adjusted to exclusion and inclusion criteria that have been specified based on the provisions of the literature to collect data from the most representative samples.

D. Operational Definition of Variables

This study consists of one dependent variable, two independent variables, and three control variables. Those variables are explained in more detail on the next page.

1. Dependent Variable
    a. Children's Permanent Income

        'Children' terminology in this study is not limited to the psychological definition or categorisation as children data used are individuals data aged 20-35 years old. 'Children' and 'father' words in this study are used to differentiate the first generation (called father) and the second generation (called children). The permanent income of children is the average income of children in a span of 7 years. The data are taken from IFLS 4 in 2007 and IFLS 5 in 2014. The average income in multiple years is used to minimize bias. According to Solon (1992), if only using one-year income, the estimated IGE value obtained will be reduced. The income referred to includes all annual income from the primary job and side job of children. Not including earnings in the form of social fund assistance such as government funding, zakat and alms, or other similar assistance. Obtained from Book III (Adult Information) Section TK (Employment) of each wave. Data is displayed using the official Indonesian currency unit, rupiah, and has been adjusted to inflation using the Consumer Price Index.

2. Independent Variable
    a. Fathers' Permanent Income

        'Father' terminology in this study is for male parent of the each child. Fathers data used are individuals data aged 40-49 years old. 'Children' and 'father' words in this study are used to differentiate the first

generation (called father) and the second generation (called children). Similar to the children's permanent income variable, fathers' permanent income variable is also the fathers' average income in a span of 7 years from IFLS 1 in 1993, IFLS 2 in 1997, and IFLS 3 in 2000. Including all annual incomes from fathers' primary job and side job without social assistance. Also obtained from Book III (Adult Information) Section TK (Employment) of each wave. This variable also shown uses the official Indonesia currency unit, rupiah, and has been adjusted to inflation using the Consumer Price Index.

b. Education

Educational variables represent the years of schooling taken from the highest level of education attained or completed which is quantified with a value of 1-18 years. To get this data, the author combines two variables which are information about the level of education and the highest class that is being occupied and completed in school. The data are obtained from Book III (Household Roster and Characteristics) Section AR (Householder Roster) of each wave. The grouping of education level based on years of schooling used in this study refers to Law No. 20 Year 2003 about the National Education System of Indonesia.

3. Control Variable

a. Age

Variables of the age of fathers and children are used to control the intergenerational life-cycle bias. As for Grawe (2004), the age limit for

the most appropriate measurement is the forties for father and twenty to mid-thirty for children. Consequently, the data used in this study are only data from fathers whose average age is 40-49 years and children whose average age is 20-35 years. Data are obtained from Book III (Household Roster and Characteristics) Section AR (Householder Roster) of each wave.

b. Children's Gender

Gender variables will participate as a control variable to determine the difference in intergenerational income elasticity between male and female. Data obtained from Book III (Household Roster and Characteristics) Section AR (Householder Roster). The variable value 1 is for male and 3 for female.

c. Fathers' Area of Living

Father's area of living will distinguish between fathers living in urban and rural areas. The father who lives in urban areas is shown as 1 and the father who lives in rural is shown as 2. Data is obtained from Book III (Household Roster and Characteristics) Section AR (Householder Roster). Father's living area will be used as a control variable to analyse differences in elasticity of intergenerational income in rural and urban areas.

E. Data and Instrument Quality Test

1. Classical Assumption Test

Hypothesis testing using a regression model should not diverge

from the classical assumptions (Trinugroho & Lau, 2019). If the classical assumption test is met then the estimate obtained meets the assumptions BLUE (Best Linear Unbiased Estimator) in which the estimate is an unbiased, consistent and efficient (Brebbia, Longhurst, Marco, & Booth, 2017). Some classical assumption tests required for this study are as follows:

a. Normality Test

The normality test aims to test whether in a regression model, the dependent variable, independent, or both has a normal distribution or not. A good regression model is one that has normal or close to normal data distribution (Puspitasari & Santoso, 2013). There are several methods that can be used to do this test. In this study, the author uses the Shapiro Wilk method with the hypothesis Ho is that data is normally distributed and Ha data is not normally distributed. If the results of the probability value are smaller than α (0.00 <0.05) then Ho cannot be rejected, and vice versa.

b. Heteroscedasticity Test

The heteroscedastic test is performed to determine whether the variant of the error is constant or not. As for the classic assumptions test, the variant of the error must be constant. In heteroscedasticity test with Cook-Weisberg or Breusch-Pagan test, Ho is that there is no problem of heteroscedasticity and Ha is for no heteroscedasticity problem. If the chi-square statistic is greater than the chi-square table, then Ho can be rejected and Ha is accepted, and vice versa.

c. Multicollinearity Test

Multicollinearity test is used to determine whether there is a relationship among independent variables. The test can be done using VIF (Variance Inflation Factor). If the VIF value is still less than 10, multicollinearity does not occur (Hair, Anderson, Tatham, & Black, 1998).

1. Endogeneity Test

Endogeneity test is carried out to see if the endogenous variables used have endogeneity. Endogeneity is an econometric terminology that shows the nature of independent variables which can cause problems in interpreting relationships in the regression model because of the selection bias (Susyanty & Pujiyanto, 2013). Endogenous testing needs to be done because improper use of instrumental variables (IV) and two-stage least square (2SLS) will produce an inefficient estimator.

F. Analysis Data and Hypothesis Test

1. Multiple Linear Regression

Multiple linear regression is intended to examine the effect of two or more independent variables (explanatory) on one dependent variable. This model assumes the existence of a straight line or linear relationship between the dependent variable and each predictor. This relationship is usually conveyed in a mathematical model.

In this study, multiple linear regression was carried out to find the estimated value of intergenerational income elasticity (IGE) in several conditions. Referring to previous studies conducted in the United States

(Solon, 1992), Australia (Mendolia, Siminski, & Mendolia, 2015), China (Jin, Bai, Li, & Shi, 2019), Italy (Piraino, 2006) and the Netherlands (Moonen & Van den Brakel, 2011), the mathematical models to estimate the IGE adjusted to life-cycle bias are explained as follows.

a. Regression Model 1: Father-Child

The first mathematical model in this study is for testing the relationship between fathers' and children permanent income in general.

$$\log(\text{kidearn}) = \beta_1 + \beta_2 \log(\text{fatearn}) + \beta_3 \text{kidage} + \beta_4 \text{kidage}^2 + \beta_5 \text{fatage} + \beta_6 \text{fatage}^2 + \varepsilon \quad \text{.....(3.1)}$$

Note:

Log(kidearn) = log of all children's permanent income

Log(fatearn) = log of fathers' permanent income

Kidage = children's age

Fatage = father's age

β_2 = regression coefficient

ε = error term

b. Regression Model 2: Father-Son

The second regression model is for testing the relationship between fathers' and male children only.

$$\log(\text{kidearn})_{\text{mal}} = \beta_1 + \beta_2 \log(\text{fatearn}) + \beta_3 \text{kidage}_{\text{mal}} + \beta_4 \text{kidage}_{\text{mal}}{}^2 + \beta_5 \text{fatage} + \beta_6 \text{fatage}^2 + \varepsilon \quad \text{.....(3.2)}$$

Note:

$\text{Log(kidearn)}_{\text{mal}}$ = log of permanent income of male children

Log(fatearn) = log of fathers' permanent income

Kidage_{mal} = age of male children

Fatage = father's age

β_2 = regression coefficient

ε = error term

c. Regression Model 3: Father-Daughter

The third regression model is for testing the relationship between fathers' and female children only.

$$\log(\text{kidearn})_{\text{fem}} = \beta_1 + \beta_2 \log(\text{fatearn}) + \beta_3 \text{kidage}_{\text{fem}} + \beta_4 \text{kidage}_{\text{fem}}{}^2 + \beta_5 \text{fatage} + \beta_6 \text{fatage}^2 + \varepsilon \quad \text{.....(3.3)}$$

Note:

$\text{Log(kidearn)}_{fem}$ = log of permanent income of female children

Log(fatearn) = log of fathers' permanent income

Kidage_{fem} = age of female children

Fatage = father's age

β_2 = regression coefficient

ε = error term

d. Regression Model 4: Urban

The following is the fourth regression model to test the relationship between fathers' living in urban areas and the children.

$$\log(\text{kidearn}) = \beta_1 + \beta_2 \log(\text{fatearn})_{\text{ur}} + \beta_3 \text{kidage} + \beta_4 \text{kidage}^2 + \beta_5 \text{fatage}_{\text{ur}} + \beta_6 \text{fatage}_{\text{rur}}{}^2 + \varepsilon \text{.....(3.4)}$$

Note:

Log(kidearn) = log of all children's permanent income

$Log(fatearn)_{rur}$ = log of permanent income of fathers living in urban

Kidage = children's age

$Fatage_{rur}$ = age of fathers living in urban area

β_2 = regression coefficient

ε = error term

e. Regression Model 5: Rural

The fifth regression model to test the relationship between fathers' living in rural areas and the children is shown on the next page.

$$\log(\text{kidearn}) = \beta_1 + \beta_2 \log(\text{fatearn})_{\text{rur}} + \beta_3 \text{kidage} + \beta_4 \text{kidage}^2 + \beta_5 \text{fatage}_{\text{rur}} + \beta_6 \text{fatage}_{\text{rur}}{}^2 + \varepsilon \text{.....(3.5)}$$

Note:

Log(kidearn) = log of all children's permanent income

$Log(fatearn)_{rur}$ = log of permanent income of fathers living in rural

Kidage = children's age

$Fatage_{rur}$ = age of fathers living in rural area

β_2 = regression coefficient

ε = error term

2. Transition Probability

The data of both fathers and children are divided into four groups where the first group is for those whose income is in the first quartile (lower than 25th percentile), the second group is for those whose income is in the

second quartile, and so on. It is managed to execute the *xttrans* command in Stata to get the report of transition probabilities. Calculating transition probabilities to measure the mobility of income across generation was also used by Moonen & Van den Brakel in 2011.

3. Explanatory Power

The first step to determine explanatory power from education towards IGE, is to find the influence of education as an intermediate variable for the permanent income of father and children. To estimate this, author uses the Two-stages Least Square (2SLS) method. The 2SLS method is a single method for solving simultaneous equations where there is a correlation between the error variable and its endogenous variables. In the 2SLS method, endogenous variables that are correlated with an error variable are replaced with the estimation of its own values.

In this study, the education variable acts as an endogenous variable instrumented by the log of father's permanent income with the log child's permanent income as the dependent (endogenous) variable. A set of simultaneous equations used in this study, also used by Jin et al. (2019), are as follows.

$$\text{education} = \theta_1 + \theta_2 \log(\text{fatearn}) + \varepsilon_1 \quad \text{.....(3.6)}$$

$$\log(\text{kidearn}) = \omega_1 + \omega_2 \text{ education} + \varepsilon_2 \quad \text{.....(3.7)}$$

Note:

Education = children's education

Log(fatearn) = log of fathers' permanent income

Log(kidearn) = log of children's permanent income

After getting the results from 2SLS, the second step is to calculate the explanatory power value using the following formula (Jin et al., 2019).

$$\text{Explanatory Power} = \frac{\omega_1}{\beta_2} \quad \text{.....(3.8)}$$

Note:

ω_1 = two-stages least square coefficient

β_2 = linear regression coefficient (IGE Indonesia