Chapter Four
Empirical Specification and Data

This chapter specifies the empirical model used in the study. The chapter consists of four parts. The earnings function and the associated OLS statistical model is presented in Section 4.1, followed by the workforce selection model in Section 4.2. Section 4.3 discusses the variables selected, and Section 4.4 presents the data and descriptive statistics. The OLS approach relates the earnings (log earnings) of individuals to their educational levels, age and other socio-economic characteristics. The OLS model was fitted using CPS and VT initial earnings data. The CPS and VT earnings function estimates are then combined to create the earnings profiles of VT graduates. Once the earnings profiles of VT graduates are created, the earnings of VT graduates without the VT degree are imputed as the earnings difference between college and high school graduates at every age, for the nation as a whole, to the sample of VT graduates. The discounted benefits of a VT education are then calculated. The discounted costs are the foregone earnings of the graduates, and tuition. The discounted costs were deducted from the discounted benefits to get estimates of NPVs of a VT education.

As discussed earlier, the standard OLS approach does not consider selectivity issues. Self-selectivity arises due to the fact that some individuals choose to work while others do not, and the earnings are likely to be correlated with the decision to participate in the laborforce. The OLS estimates, which do not consider this fact, are thus biased and do not reflect the true structural parameter estimates. The returns to education could, therefore, be overestimated. Heckman (1976a, 1979) proposed a model that adjusts the estimates for self-selection. The estimates in this case are unbiased and asymptotically efficient. A model that corrects for the endogenous workforce participation decision is used in the second part of the study to estimate the earnings functions for the CPS and VT samples, again. The CPS and VT earnings function estimates are combined to create the earnings profiles of VT graduates. Once the earnings profiles of VT graduates were created, the earnings of VT graduates without the VT degree are imputed as the earnings difference between college and high school graduates, for the nation as a whole, applied to the VT graduates’ age-earnings profiles. The earnings of the graduates are also multiplied by their probabilities of participating in the workforce, to adjust the benefits for the changes education creates in the probabilities of workforce participation. For example, college graduates are generally more likely to participate in the workforce than high school graduates. The probabilities of workforce participation of VT graduates with and without the VT degree are based on the probabilities of workforce participation of college and high school graduates in the CPS sample, respectively. The predicted earnings of the graduates are multiplied by the probabilities to get the probability-adjusted earnings. The probability-adjusted earnings are used to calculate the discounted benefits and the discounted values of foregone earnings. The sum of the discounted values of foregone earnings and tuition gives the discounted costs. The discounted costs were deducted from the discounted benefits to get the NPVs.

4.1 The OLS Approach (model without the workforce participation decision correction)

4.1.1 CPS Earnings Function

In the first part of the study, Current Population Survey (CPS 1999) data are used to estimate an earnings function. CPS data gives a snapshot of individuals with different ages and their earnings at a point in time. Since the purpose is to see the earnings differential between high school and college graduates, a sample of individuals between 18 and 65 years, either with a high school or college degree, and not currently in school, was selected from the CPS. The estimated equation has the following functional form.

$$LnY = \beta_0 + \beta_1 GEN + \beta_2 HS + \beta_3 AFAM + \beta_4 AMIND + \beta_5 ASIAN + \beta_6 AGE + \beta_7 AGESQ + \beta_8 AGEHS + e \qquad (4.1)$$

where, Y is the earnings after taxes; GEN is the gender (=1 if female,=0 if male); HS is the level of education (=1 if high school graduate, =0 if college graduate); AFAM is the African American dummy (=1 if African American, otherwise 0); AMIND is the

American Indian dummy (=1 if American Indian, otherwise 0); ASIAN is the Asian dummy (=1 if Asian, otherwise 0); AGE is the age of the individual (actually age-22 is used to shift the base to 22 years because the salaries of VT graduates are observed at 22 years, by assumption); AGESQ is AGE squared; and AGEHS is the interaction between AGE and HS.

4.1.2 VT Earnings Function

Another earnings function was estimated for the Virginia Tech (VT) sample. The VT sample consists of the starting salaries and personal characteristics of students receiving an undergraduate degree in 1998-99. This data were gathered from the records of the Career Services Office and from Institutional Planning and Research at VT. Future earnings of the graduates were imputed by combining initial earnings estimates with earnings growth implied by the coefficients of age and age-squared from the CPS earnings function. It is therefore assumed that a typical VT graduate will experience the same earnings growth (with age) as that of a person in the nationally representative CPS sample. The VT earnings function is specified as:

$$\begin{aligned} LnY = {} & \gamma_0 + \gamma_1 GEN + \gamma_2 AFAM + \gamma_3 ASIAN + \gamma_4 FORN + \gamma_5 HISP + \gamma_6 ARCH \\ & + \gamma_7 HRE + \gamma_8 BUS + \gamma_9 AGRIC + \gamma_{10} ENGG + \gamma_{11} NAT + e \end{aligned} \tag{4.2}$$

where Y is the starting salary of the graduate after taxes; GEN is the gender (=1 if female,=0 if male); AFAM is the African American dummy (=1 if African American, otherwise 0); ASIAN is the Asian dummy (=1 if Asian, otherwise 0); FORN is the Foreign dummy (=1 if Foreigner,=0 otherwise); and, HISP is the Hispanic dummy (=1 if Hispanic, =0 otherwise). Whites was the base for the race dummies. ARCH, HRE, BUS, AGRIC, ENGG, and NAT are the college dummies, and they refer to the Colleges of Architecture, Human Resources, Business, Agriculture, Engineering and Natural Resources, respectively. The College of Arts & Sciences was considered the base category.

The coefficients of age and age-squared (β_6 and β_7 respectively) from the CPS earnings function (Equation 4.1) are then plugged into the above equation to create the earnings profiles. The equation that was used to create the earnings profiles is presented below (Equation 4.3). It is assumed that a VT graduate remains in the workforce from age 22 to 65, while the graduate without the VT degree remains in the workforce from age 18 to 65.

$$\begin{aligned} Ln\hat{Y} = {} & \gamma_0 + \gamma_1 GEN + \gamma_2 AFAM + \gamma_3 ASIAN + \gamma_4 FORN + \gamma_5 HISP + \gamma_6 ARCH \\ & + \gamma_7 HRE + \gamma_8 BUS + \gamma_9 AGRIC + \gamma_{10} ENGG + \gamma_{11} NAT \\ & + \beta_6 AGE + \beta_7 AGESQ \end{aligned} \tag{4.3}$$

4.1.3 Calculation of NPVs

Once the earnings profile of a VT Graduate is created, the earnings without the VT degree is found by adding the estimated loss in earnings from having only a high school degree (β_2) and the coefficient of the interaction term for high school degree and age (β_8). The interaction term (β_8) was used to capture the earnings difference between a college and a high school graduate with age. The equation that is used to create the earnings of the graduates without the VT degree is presented below, assuming HS =1.

$$\begin{aligned} Ln\hat{Y} = {} & \gamma_0 + \gamma_1 GEN + \gamma_2 AFAM + \gamma_3 ASIAN + \gamma_4 FORN + \gamma_5 HISP + \gamma_6 ARCH \\ & + \gamma_7 HRE + \gamma_8 BUS + \gamma_9 AGRIC + \gamma_{10} ENGG + \gamma_{11} NAT \\ & + \beta_6 AGE + \beta_7 AGESQ + \beta_2 HS + \beta_8 AGEHS \end{aligned} \tag{4.4}$$

The salary difference between the VT graduate and what he/she would have earned without the degree was discounted at the rate of 5 per cent. This gives the discounted benefits of a VT degree. The foregone earnings (of a VT Graduate) and the cost of study (tuition) are discounted to get the discounted costs. The foregone earnings are the earnings of the graduates without the VT degree, from 18 to 21 years. The foregone earnings are based on the earnings for nine months. In other words, it is assumed that the VT graduate works for 3 months in a year while attending college. This is a standard assumption in the literature. The discounted costs are subtracted from the discounted benefits to get the NPVs. The NPVs are estimated using the following equation:

$$NPV = \sum \frac{(\mathrm{Y_{VT}} - \mathrm{Y_{HS}})}{(1+\mathrm{r})^{\mathrm{t}}} - \sum \frac{\mathrm{Y_{CS}}}{(1+\mathrm{r})^{\mathrm{t}}} \tag{4.5}$$

where Y_{VT} are the earnings of a VT graduate, Y_{HS} are the earnings of the graduate without the VT degree, Y_{CS} is the cost (sum of foregone earnings and tuition) of getting the VT degree, r is the discount rate and t the time period in question.

4.2 Workforce selection model (model with the endogenous workforce participation decision correction)

The model that corrects for the endogenous workforce participation decision is now presented.

4.2.1 CPS Earnings Function

The model jointly estimates the earnings equation and the workforce participation equation. The first equation relates the natural logarithm of earnings to the level of education, age and race. The equation is specified as follows:

$$\begin{aligned} LnY = {} & \beta_0 + \beta_1 GEN + \beta_2 HS + \beta_3 AFAM + \beta_4 AMIND + \beta_5 ASIAN + \beta_6 AGE + \beta_7 AGESQ \\ & + \beta_8 AGEHS + u_1 \end{aligned} \tag{4.6}$$

where Y is the earnings after taxes; GEN is the gender (=1 if female,=0 if male); HS is the level of education (=1 if high school graduate, =0 if college graduate); AFAM is the African American dummy (=1 if African American, otherwise 0); AMIND is the American Indian dummy (=1 if American Indian, otherwise 0); ASIAN is the Asian dummy (=1 if Asian, otherwise 0); AGE is the age of the individual (actually age-22 is used to shift the base to 22 years); AGESQ is AGE squared; AGEHS is the interaction between AGE and HS; and, u_1 is the error term.

The selection equation models the decision to participate in the laborforce as a function of the above characteristics plus two other variables (number of children under the age of six in the family and the interaction between gender and the number of children under six in the family) not included in the earnings function. The equation is specified as follows:

$$\begin{aligned} WFPART = {} & \gamma_0 + \gamma_1 GEN + \gamma_2 HS + \gamma_3 AFAM + \gamma_4 AMIND + \gamma_5 ASIAN + \gamma_6 AGE \\ & + \gamma_7 AGESQ + \gamma_8 CHI + \gamma_9 CHIGEN + u_2 \end{aligned} \tag{4.7}$$

where WFPART is the decision to participate in the laborforce (=1 if wages are observed,=0 otherwise); GEN is the gender (=1 if female,=0 if male); HS is the level of education (=1 if high school graduate, =0 if college graduate); AFAM is the African American dummy (=1 if African American, otherwise 0); AMIND is the American Indian dummy (=1 if American Indian, otherwise 0); ASIAN is the Asian dummy (=1 if Asian, otherwise 0); AGE is the age of the individual (actually age-22 is used to shift the base to 22 years); AGESQ is AGE squared; AGEHS is the interaction between AGE and HS; CHI is the number of children under six in the family; CHIGEN is the interaction between CHI and GEN; and, u_2 is the error term.

The individual works if,

$$\begin{aligned} & \beta_0 + \beta_1 GEN + \beta_2 HS + \beta_3 AFAM + \beta_4 AMIND + \beta_5 ASIAN + \beta_6 AGE + \beta_7 AGESQ \\ & + \beta_8 AGEHS + u_1 \geq \gamma_0 + \gamma_1 GEN + \gamma_2 HS + \gamma_3 AFAM + \gamma_4 AMIND + \gamma_5 ASIAN \\ & + \gamma_6 AGE + \gamma_7 AGESQ + \gamma_8 AGEHS + \gamma_9 CHI + \gamma_{10} CHIGEN + u_2 \end{aligned} \tag{4.8}$$

Therefore, the error terms in the two equations are likely to be correlated, that is, Cov (u_1, u_2) is not equal to zero. Estimation of the earnings equation with only individuals that are working could lead to biased estimates. The estimation procedure involves either the two-step error correction method (developed by Heckman) or the maximum likelihood method. The two-step error correction procedure (used in this study) involves estimating the parameters of Equation 4.7 by the probit method, that models whether the individual is in the laborforce or not. The earnings equation (Equation 4.6) is then estimated by

OLS, using the estimates of Equation 4.7 obtained by the probit method. The earnings equation now has consistent estimates of the β's.

4.2.2 VT Earnings Function

The Heckman two-step procedure was again used to estimate the VT earnings function. In this case the earnings function relates the natural logarithm of the initial earnings of VT graduates to gender, race and college. The equation is specified as follows:

$$\begin{aligned} LnY &= \gamma_0 + \gamma_1 GEN + \gamma_2 AFAM + \gamma_3 ASIAN + \gamma_4 FORN + \gamma_5 HISP + \gamma_6 ARCH \\ &\quad + \gamma_7 HRE + \gamma_8 BUS + \gamma_9 AGRIC + \gamma_{10} ENGG + \gamma_{11} NAT + e_1 \end{aligned} \qquad (4.9)$$

where Y is the starting salary of the graduate after taxes; GEN is the gender (=1 if female,=0 if male); AFAM is the African American dummy (=1 if African American, otherwise 0); ASIAN is the Asian dummy (=1 if Asian, otherwise 0); FORN is the Foreign dummy (=1 if Foreigner,=0 otherwise); and, HISP is the Hispanic dummy (=1 if Hispanic, =0 otherwise). Whites was the base for the race dummies. ARCH, HRE, BUS, AGRIC, ENGG, and NAT are the college dummies, and they refer to the colleges of Architecture, Human Resources, Business, Agriculture, Engineering and Natural Resources, respectively. The College of Arts & Sciences was considered the base.

The selection equation for the VT graduates is specified as follows:

$$\begin{aligned} WFPART &= \gamma_0 + \gamma_1 GEN + \gamma_2 AFAM + \gamma_3 ASIAN + \gamma_4 FORN + \gamma_5 HISP + \gamma_6 ARCH \\ &\quad + \gamma_7 HRE + \gamma_8 BUS + \gamma_9 AGRIC + \gamma_{10} ENGG + \gamma_{11} NAT + \gamma_{12} QCA + e_2 \end{aligned} \qquad (4.10)$$

where WFPART is the decision to participate in the laborforce (=1 if starting salary is observed,=0 otherwise); GEN is the gender (=1 if female,=0 if male); AFAM is the African American dummy (=1 if African American, otherwise 0); ASIAN is the Asian dummy (=1 if Asian, otherwise 0); FORN is the Foreign dummy (=1 if Foreigner,=0 otherwise); HISP is the Hispanic dummy (=1 if Hispanic, =0 otherwise); ARCH, HRE, BUS, AGRIC, ENGG, NAT are the college dummies and they refer to the colleges of

Architecture, Human Resources, Business, Agriculture, Engineering and Natural Resources, respectively; and QCA, is the Quality Credit Average of the graduate.

The coefficients of age and age-squared (β_6 and β_7) from the CPS earnings function (Equation 4.6) are again plugged into the VT earnings equation (Equation 4.9) to create the lifetime earnings profiles of the graduates. Once the earnings profiles are created, the earnings loss for only having a high school degree (β_2) and the interaction term (β_8) are added as before to the VT earnings equation (Equation 4.9) to create the earnings of the graduates without the VT degree.

4.2.3 Adjustment for workforce participation

The earnings of the graduates are dependent on the probability of participating in the laborforce. This study attempts to adjust the earnings of the graduates (with and without VT degree) for laborforce participation. The CPS selection (participation) equation was used to predict the probabilities of workforce participation of VT graduates, with and without the VT degree. For this purpose, a dataset was created that consisted of the variables in the CPS selection equation - gender, race, age, age-squared, number of children under six in the family and the interaction of gender and the number of children under six in the family. The number of children under six in the family was assumed to be equal to the average number of children under the age of six in the family, for different age-groups (18-22, 23-27, 28-32, 33-37, 38-42, 43-47, 48-52, 53-57, 58-62, 63 and above) in the CPS sample. The predicted probabilities were multiplied by the predicted earnings of VT graduates to get the expected earnings of graduates over lifetime.

4.2.4 Calculation of NPVs

The probability-adjusted earnings are used to calculate the discounted benefits, costs and the NPVs.

4.3 Discussion of variables

4.3.1 CPS Earnings Function

Most of the explanatory variables that are included in the earnings function have been proposed in the human capital literature. The earnings of individuals have been hypothesized to depend on the educational level, age and personal characteristics. Education has often been measured by the number of years of schooling. However, in this study a dummy variable has been used to denote the educational level of the individual (a high school degree) since we are interested in the earnings premium associated with a college education. Numerous studies have found a positive relationship between earnings and education. Age (or experience) has also been found to have a positive impact on earnings. In this study age has been used as a proxy for experience. According to the human capital theory, earnings increase with age at a diminishing rate. Earnings peak at a point (usually around mid-life) and then fall (see Chapter 3 for the theoretical model). The age-squared term is used to capture the diminishing returns to age. Socio-economic characteristics also play an important role in determining earnings. The most common socio-economic characteristics that have been used in the literature are gender and race. This study also uses them as explanatory variables. The earnings of individuals with different levels of education is likely to vary with age. An interaction term (age*high school) is used to capture the earnings difference between the high school dummy and age. The same variables are used in both the OLS earnings equation and the selectivity corrected earnings equation.

4.3.2 CPS Selection Equation

The explanatory variables used in the selection equation are similar to those used in the earnings function, except two of them, which were hypothesized to determine the decision to participate in the laborforce, but not to determine earnings. For identification of the earnings equation the selection equation must have at least one variable that is not present in the earnings equation. The variables in the selection equation that are not present in the earnings equation are the number of children under six in the family, and the interaction between gender and the number of children under six in the family. The first term was used to see the working habits of families with young children. The presence of young children at home is likely to affect the workforce participation decision since they (children) require the presence of parents to take care of them. The interaction term (number of children under six in the family and gender) was used to take into account the likely stronger effect of the presence of young children in the family on the workforce participation decision of mothers.

4.3.3 VT Earnings Function

The explanatory variables used in the estimation of the VT earnings function significantly differ from those used in the CPS earnings function. The common variables are gender and race. Age could not be used as an explanatory variable in this case because most VT graduates was assumed to be of the same age (that is 22 years). Although human capital theory does not say anything about the importance of school quality and the choice of major, recent studies have focused on the importance of these two factors in determining earnings. The choice of major is particularly important in the skills-driven labor market of today. This study considers the college of the graduate as an explanatory variable. The same variables are used in both the OLS earnings equation and the workforce selection earnings equation.

4.3.4 VT Selection Equation

The explanatory variables used in the selection equation in this case are similar to those that were used in the VT earnings function equation, except one new variable that was used only in this equation. This variable is the Quality Credit Average (QCA) of the graduate. It is hypothesized that the QCA might influence the decision to participate in the laborforce, but not earnings. Particularly, students with higher QCA are more likely to go for higher studies and thus postpone entry into the laborforce.

4.4 Data and Descriptive Statistics

4.4.1 CPS Sample – OLS Model

Current Population Survey data (CPS 1999) was used to estimate the CPS earnings function. The Current Population Survey is a nationally representative household survey conducted monthly by the Census Bureau to provide estimates of employment, unemployment, and other characteristics of the labor force, estimates of the population as a whole, and estimates of various sub-groups in the population. Data are collected on several socio-economic characteristics of the population including work experience, income, noncash benefits, migration, employment status, occupation, health insurance, Medicaid, Medicare etc. Information on demographic characteristics such as age, sex, race, household relationships is also available for each person in the survey. The universe consists of civilian non-institutional population of the US living in housing units and members of the Armed Forces living in civilian housing units on a military base or in households not on a military base.

The CPS is a hierarchical dataset with 3 record types. The first record type is Household, with 124 variables for 65,337 records. The Family record type has 76 variables for 57,325 records, and Person record type has 430 variables for 132,324 records. From the Person record type, individuals with a high school or college degree, between 18 and 65 years, currently not in school and in the laborforce were selected. The sample thus consisted of 30,157 individuals. However, it must be noted that salaries were not reported by all individuals. The mean age for the sample was about 40 years and the mean salary $32,312. Out of the 30,157 individuals positive salaries were reported for 27,706 individuals – 17,469 being high school graduates and 10,237 being college degree holders. The mean salary for the high school group was $25,351, and the average for the college group was $44,191. The high school group consists of 9076 males and 8393 females, while the college group has 5286 males and 4951 females. Salary statistics by race and educational level are given in Table 4.1.

Table 4.1: CPS Sample Statistics by educational level and race

Educational Level	Race	N	Mean ($)	Standard Deviation($)
High School	African American	1,775	21,278	14,320
	American Indian	201	21,828	15,824
	Asian	389	22,308	20,775
	White	15,104	25,955	22,755
	ALL	17,469	25,351	21,984
College	African American	647	37,667	30,252
	American Indian	54	42,641	42,758
	Asian	522	39,544	30,198
	White	9,014	44,938	41,915
	ALL	10,237	44,191	40,803
Both Groups		27,706	32,312	31,662

4.4.2 CPS Sample – Workforce Selection Model

For the analysis using the workforce selection model, a broader sample was selected. In this case the sample consisted of individuals between 18 and 65, with a high school or college degree and not in school. The sample thus included individuals both in and out of the laborforce. The sample consisted of 38,742 observations. Salaries were observed for 30,278 individuals and unobserved for 8464 individuals. Salary statistics by race and educational level are given in Table 4.2. There is sampling error in this case too.

Table 4.2: CPS Sample Statistics for workers by educational level and race

Educational Level	Race	N	Mean ($)	Standard Deviation($)
High School	African American	2042	20,031	14,412
	American Indian	257	20,089	16,228
	Asian	446	21,257	20,102
	White	16,649	24,958	23,178
	ALL	19,394	24,289	22,328
College	African American	692	36,830	29,817
	American Indian	61	40,162	41,172
	Asian	565	37,824	29,810
	White	9566	43,995	42,601
	ALL	10,884	43,198	41,372
Both Groups		30,278	31,086	31,888

4.4.3 VT Sample – OLS Model

The Virginia Tech sample consists of data on earnings (starting salaries) and personal characteristics (race, gender, college) of undergraduates that graduated in the academic year 1998-99. Data were gathered from the records of the Career Services Office and from Institutional Planning and Research. The Career Services office conducts a survey each year to collect data on employment after graduation, earnings, location of employment etc. of the graduating students. The earnings and college data were obtained from the Career Services Office, while the Office of Planning and Institutional Research provided data on personal characteristics. Out of the 3993 graduates in 1998-99, usable information was obtained for 1761 students (Table 4.3). The mean salary for the sample was $32,664 with a standard deviation of $10,334. The mean salary was highest for engineering majors ($40,955) and lowest for natural resources graduates ($24,377). In the sample, the number of graduating students was highest for the College of Arts & Sciences (446), closely followed by the College of Engineering (442). The mean salary for all males in the sample was $35,051, while the figure was $29,349 for females.

Table 4.3: VT Starting Salaries by College

College	N	Mean ($)	Standard Deviation ($)
Arts & Sciences	446	28,510	10,099
Agriculture	122	24,796	6,663
Architecture	65	30,300	7,359
Business	414	35,681	8,310
Engineering	442	40,955	8,152
Human Resources	207	25,862	6,408
Natural Resources	65	24,377	7,557
ALL	1,761	32,664	10,334

4.4.4 VT Sample – Workforce Selection Model

In the estimation of the VT earnings function using the Heckman model, the sample consisted of individuals both with, and without observed values of salaries. Salaries were observed for 1761 individuals and unobserved for 1235 individuals. The sample size was thus 2996 observations.