

# Appendix to: “Accounting for the Gender Gap in College Attainment”\*

Suqin Ge<sup>†</sup>

Fang Yang<sup>‡</sup>

Virginia Tech

SUNY-Albany

December 9, 2009

## Abstract

In this appendix we provide a detailed description of the data processing procedure on PSID. We also describe the estimation results, along with a full description of our two-stage estimation of wage.

## 1 Appendix

### 1.1 PSID sample

The PSID is a longitudinal survey of U.S. families and the individuals who make up those families. Approximately 4,800 U.S. families were sampled in 1968, and these families were reinterviewed annually until 1997. From 1997 onwards, PSID was changed to a biennial data collection and two major changes were made: a reduction of the core sample and the addition of a new sample of post-1968 immigrant families and their adult children.

We first find parents’ education for the selected sample by linking parents and children from Individual Files (1968–2005). The PSID facilitates the intergenerational linkage by

---

\*We thank Michele Boldrin, Betty Daniel, Mariacristina De Nardi, Zvi Eckstein, Raquel Fernandez, John Jones, Michael Sattinger, and seminar participants at Virginia Tech, 2008 AEA meetings, University of Virginia, SUNY-Albany, 2008 North American Summer Meeting of the Econometric Society, Xiamen University, W.E. Upjohn Institute for Employment Research, the Philadelphia Fed, 2008 Annual Meeting of the Southern Economic Association, Université Laval, and 2009 Midwest Macro Meetings for helpful comments and suggestions. Ge acknowledges the AEA/CSWEP summer fellowship from W.E. Upjohn Institute for Employment Research. All remaining errors are our own.

<sup>†</sup>Ge: Department of Economics, Virginia Tech, Blacksburg, VA 24061 (email: ges@vt.edu).

<sup>‡</sup>Yang: Department of Economics, University at Albany, Albany, NY 12222 (email: fyang@albany.edu).

providing the parent’s ID in the Individual Files. If a linkage cannot be found in Individual Files, we use 2003 Parent Identification Files to link an individual with his or her parents. If the above procedure fails to provide parents’ education information, we find parents’ education by using parents’ and parents-in-law’s education as reported by the head in Family Files. In 1974, questions were asked about how much education had been completed by the household head’s parents and by the spouse’s parents. In the later waves, these parental education questions were asked for new heads and spouses. By merging Individual Files with Family Files, we are able to find parents’ education for those who were heads or spouses or siblings of the heads.

## 1.2 Estimation of wage

The model is estimated on the March CPS from 1964 to 2007. We restrict the sample to individuals who are between the ages of 18 and 65 who are not in the armed forces and not self-employed. To be consistent with the decision model, we restrict our attention to individuals who are either married or single (never married). Hourly wage is deflated to 2006 dollars using the CPI. Definitions of variables are given in Appendix section 1.2.2. We run separate probit wage selection and log wage regression for each gender in each year. The reduced-form probit selection results and estimated coefficients of the wage equations in 2007 are provided in Appendix sections 1.2.3 and 1.2.4.

### 1.2.1 Estimation procedure of wages

Consider the following wage function on a sample of working men and women:

$$\log w_i = X_i\beta + \mu_i,$$

where  $\log w_i$  is the logarithm of hourly wage, and  $X$  is a vector of characteristics such as schooling and work experience. It is argued, however, that the sample of employed workers is not a random sample and that this selectivity might bias the coefficients. Formally, we can write down a participation equation

$$\begin{aligned} E_i &= 1 \text{ if } Z_i\gamma + \varepsilon_i \geq 0, \\ E_i &= 0 \text{ if } Z_i\gamma + \varepsilon_i < 0, \end{aligned}$$

where  $Z$  includes variables that predict whether or not a person works. Therefore, the probability of an individual working is

$$(1) \quad \Pr(E_i = 1) = \Pr(\varepsilon_i \geq -Z_i\gamma) = \Phi\left(\frac{Z_i\gamma}{\sigma}\right),$$

where  $\sigma^2$  is the variance of  $\varepsilon_i$ , and  $\Phi(\cdot)$  is cumulative distribution function of the standard normal.

The selectivity problem is apparent by taking expectations of the wage function over

the sample of employed workers:

$$E(\log w_i | E_i = 1, X_i) = X_i\beta + E(\mu_i | \varepsilon_i \geq -Z_i\gamma).$$

Supposing  $\mu_i$  and  $\varepsilon_i$  are jointly normally distributed, let  $\sigma_{\mu,\varepsilon}$  be the covariance between  $\mu_i$  and  $\varepsilon_i$ . We can now write

$$E(\mu_i | \varepsilon_i \geq -Z_i\gamma) = \frac{\sigma_{\mu,\varepsilon}}{\sigma_\varepsilon} \frac{\phi(Z_i\gamma/\sigma)}{\Phi(Z_i\gamma/\sigma)},$$

where  $\phi(\cdot)$  is the standard normal density. When  $\sigma_{\mu,\varepsilon}$  is not zero, selectivity bias occurs. To estimate the potential wage consistently, we need to add the selection term (the inverse Mills ratio)

$$(2) \quad \frac{\phi(Z_i\gamma/\sigma)}{\Phi(Z_i\gamma/\sigma)} \equiv V_i$$

in the OLS regression as

$$\log w_i = X_i\beta + \alpha V_i + \eta_i.$$

### 1.2.2 Definitions of variables in $X$ and $Z$

Age	Respondent's age
Age <sup>2</sup>	Square of variable "Age"
HI	Dummy variable: 1 if respondent is a high school dropout
HG	Dummy variable: 1 if respondent is a high school graduate
SC	Dummy variable: 1 if respondent has some college education
CG	Dummy variable: 1 if respondent is a college graduate
Exp	Respondent's years of work experience
Exp <sup>2</sup>	Square of variable Exp
Black	Dummy variable: 1 if respondent is black
Married	Dummy variable: 1 if respondent is married
Nchild	Number of own children in household
Nchlt5	Number of own children under age 5 in household
Northeast	Dummy variable: 1 if household is located in Northeast area
Midwest	Dummy variable: 1 if household is located in Midwest region
South	Dummy variable: 1 if household is located in South region
West	Dummy variable: 1 if household is located in West region
Metro	Dummy variable: 1 if household is located in a metropolitan area
Manager	Dummy variable: 1 if respondent is a manager or professional
Whitecollar	Dummy variable: 1 if respondent has white-collar occupation other than those in management
Bluecollar	Dummy variable: 1 if respondent has blue-collar occupation
V	See Equation (2)

Variable	Males		Females	
	Coefficient	<i>t</i>	Coefficient	<i>t</i>
Constant	-2.5929	-41.75	-2.6571	-43.10
HG	0.3134	15.67	0.5029	24.77
SC	0.3882	18.68	0.6520	32.05
CG	0.7044	31.09	0.8212	38.86
Age	0.1627	46.82	0.1448	41.89
Age <sup>2</sup>	-0.0022	-52.07	-0.0018	-44.29
Black	-0.3328	-16.14	-0.0018	-0.10
Marry	0.4641	22.41	-0.0499	-2.91
Nchild	0.0396	4.82	-0.0800	-12.86
Nchlt5	0.0315	1.79	-0.2708	-22.21
No. of obs.	48,145		51,315	
-2 ln(likelihood ratio)	8285.05		5252.96	
$\chi^2$ degree of freedom	9		9	

Table 1: Participation Selection Rules: Probit Analysis (CPS 2007)

### 1.2.3 Estimation results: probit selection

The reduced-form probit selection rule in equation (1) is estimated in each year for men and women. We estimate these probits year by year because some evidence shows that how individuals select themselves into the workforce has shifted over time (Mulligan and Rubinstein 2007). Table 1 presents estimated coefficients and asymptotic *t*-statistics of the reduced form participation probit for 2007.<sup>1</sup> Our findings are generally in accord with previous research. Specifically, we find that educational attainment has a positive and statistically significant impact on the probability of participation for both men and women. The probability of working increases in age at a decreasing rate for both men and women. Black men are less likely to participate than nonblacks. Men who are married or have children are more likely to participate than other men, even though the effect of the number of children is not statistically significant. Married women and women with children are less likely to participate.

### 1.2.4 Estimation results: wage equations

Estimated coefficients and asymptotic *t*-statistics of the wage equations in 2007 corrected for selections are found in Table 2. Estimated coefficients on education, experience, occupation dummies, race, and region dummies are similar to estimates from typical wage equations found in the literature. College education attainments are generally more important for women's wage than for men's. Experience has more of a positive impact on men's wage than on women's.

Selectivity biases are particularly interesting. One would expect that individuals with higher wage potential should be more likely to participate in the labor force. The estimation

<sup>1</sup>Estimates for other years are available from the authors.

Variable	Males		Females	
	Coefficient	<i>t</i>	Coefficient	<i>t</i>
Constant	1.5958	29.79	1.4525	31.64
HG	0.3278	22.01	0.2822	13.43
SC	0.4650	28.34	0.4704	20.27
CG	0.7743	36.96	0.8072	31.61
exp	0.0462	19.04	0.0336	19.29
exp <sup>2</sup>	-0.0009	-13.81	-0.0006	-15.43
manager	0.3618	36.55	0.4974	38.16
white-collar	0.0099	1.09	0.1966	19.19
Midwest	-0.0713	-6.83	-0.0746	-6.61
South	-0.0829	-8.39	-0.0766	-7.09
West	-0.0445	-4.38	-0.0453	-4.01
metro	0.1150	12.74	0.1476	15.27
black	-0.1424	-9.38	-0.0298	-2.41
married	0.2748	16.84	0.0425	4.01
<i>V</i>	0.3131	5.08	0.2283	6.51
<i>R</i> <sup>2</sup>	0.3026		0.2362	

Table 2: Estimates of wage equation: CPS 2007

results confirm that individuals who expect to earn more are more likely to participate in the labor force. The coefficients of *V* (defined in equation (2) in Appendix 1.2.1) are positive and statistically significant for both men and women. Therefore, observed wage patterns of men and women are higher than the population mean pattern would have been.