Factors Affecting the Wage of Adult Civilians in the United States ECO 521: Econometrics

Kai Li — kai.li@stonybrook.edu

Stony Brook University — May 16, 2022

1 Introduction

The effect of relevant economic factors on wages has been a long-lasting interest in the field of labor economics. Based on classical economic theory, for example, age can have a strong impact on an individual's wage because one is more likely to accumulate additional experience and skills as an experienced worker compared to beginners; geographic location may also come to play as local demand for work can vary across the geographic factors affecting wages; finally, existing studies have also shown the strong causation between educational attainment and wages (e.g., Card [5] and Lazear [14]). In this paper, we investigate a series of important economic variables relevant to wages, selected based on economic theory and empirical research, to further analyze their relationships. We use a reliable open-source dataset from the United States government as the foundation for this analysis. For the regression model, we prefer taking the natural log of wage as the dependent variable. We also notice the potential problem of sample selection bias in estimating the log wage based on the previous literature. A popular solution to sample selection, called Heckman's two-step selection method, is formulated and discussed in this paper. We implement the method in statistical software R to analyze the dataset. More specifically, we have a probit equation with labor force participation regressed on some variables as well as an ordinary least squares (OLS) equation with log wage regressed on some variables. Regression results including coefficient estimates, standard errors, and t-tests of inference are provided for analysis and interpretation.

The rest of the paper is organized as follows. A literature review is given in Section 2. Section 3 further discusses the source of the data, data cleaning procedures, summary statistics of the data, and some exploratory visual analysis. A detailed methodology formulation with statistical asymptotic properties is given in Section 4. Section 5 provides our regression results based on the methodology introduced. Interesting findings observed from the results are further analyzed and illustrated. Finally, we conclude the paper in Section 6.

2 Literature Review

The first few steps of analysis involve the selection of a dataset, variables, and a sample. We build our variable selection process according to existing empirical research and relevant economic theories. For example, there is an observed wage difference between college graduates and high school graduates [18]. Therefore, it is reasonable and necessary to include at least a variable recording respondents' educational attainment. Additionally, wage differences are significant both within and among various experience, education, and gender groups [4, 18]. For a similar reason, we look for relevant variables that describe experience and gender. We also borrow ideas from classic econometric and labor economics studies, such as Mroz [19], and surveys from econometricians to obtain insights into deciding the essential variables to minimize the risk of omitted variable bias.

In labor theory, theoretical and applied econometrics, sample selection has been a recent literature topic with a wide range of discussions. Initially, the model of labor supply consisting of two equations, namely the wage equation and the labor-hours equation, has raised researchers' attention in this field. The discussion of the model begins with Gronau [10] and Lewis [15]. Then, Heckman [12] proposes an estimator for sample selection, truncation, and limited dependent variables statistical models to estimate the model of labor supply. The labor supply model can be generalized as a sample selection model in econometrics. The initial parameterized model was proposed by Heckman [13] using the two-step estimation method. In practice, Heckman's method is enormously used, though the method of maximum likelihood (ML) is also used sometimes. Unfortunately, the modeling result between Heckman's two-step estimates and the maximum likelihood estimates may vary considerably. For instance, Greene [9] compares the estimation performance of Heckman's method, ML, and OLS to the Mroz [19] study and observes a significant variation in the estimation coefficients across the methods.

Note that Heckman's two-step method is a limited information maximum likelihood (LIML) method. There is some existing research comparing the performance of the LIML estimator, the full information maximum likelihood (FIML) estimator, the subsample OLS estimator, etc. on sample selection models using Monte Carlo simulation methods. Puhani [25] investigates Heckman's two-step estimator and recommends the FIML estimator instead of Heckman's LIML estimator if no collinearity problem arises. Nelson [22] suggests the subsample OLS estimator when collinearity is small, the LIML if collinearity is intermediate, and the FIML otherwise. What is more, FIML may be difficult to implement computationally [25]. Critiques also exist in literature. For example, Nelson [24] critiques Heckman's estimator regarding predictive power, collinearity, and the sensitivity of the estimated coefficients. However, Heckman's LIML estimator is consistent because of the large-sample property, and that may be the potential reason for its wide applicability in practice.

An important assumption in the selection model is the normality condition, which is also debatable, as appears in some previous literature, such as Goldberger [7]. In particular, the estimator may not be consistent if the normality condition is violated or the given sample is too small [7]. On the other hand, the existence of conjectures does not mean that the normality assumption in a selection model is wrong [9]. We believe there is a trade-off between the actual case and the hypothesized scenario. More generalized models to deal with selectivity issues using nonparametric, semiparametric and robust estimators are proposed. For example, Newey et al. [23] and Vella [26] give a detailed overview of Heckman's classical method as well as semiparametric methods as relaxations of distributional assumptions in sample selection. Moreover, Martins [16] constructs an alternative normality assumption in selection models. However, these estimators may not be appropriate for a broad variety of models and the models are much less operational [9]. Therefore, given our sample size is large, we assume the error terms in the sample selection model are normally distributed, and employing Heckman's two-step estimation is appropriate here.

3 Data

The Current Population Survey (CPS) is the main survey of labor force statistics for the United States population, sponsored jointly by the U.S. Census Bureau and the U.S. Bureau of Labor Statistics (BLS) [6]. In particular, the samples and the variables used in this paper are extracted from the Annual Social and Economic Supplement (ASEC) of the Current Population Survey.

We have obtained the most recent sample from March 2021 with 15 variables that form a dataset with 163,543 observations. Not all variables are useful for analysis until we have applied the necessary cleaning procedures to the observations in the data. For instance, observations that are encoded as blank, missing, or not in universe (NIU)[1] are deleted. Dummy variables are created to break down categorical variables with multiple unordered choices. For example, *Region*[2] is divided into four specific regions with a binary variable for each region. Outliers are treated appropriately after a careful analysis of the potential reason as well as the goodness of fit of regression models. For instance, we consider the number of outliers and decide whether a listwise deletion or an imputation is required. Transformations and derived variables are considered based on the given problem.

We intend to focus on individuals who are adult civilians aged from 16 to 64 in the United States. *Popstat* reports the person's status in the population, such as adult civilian, armed forces, or child. After removing blank, missing, or NIU observations, everyone remaining in the sample is an adult civilian. We do not have a variable matching the hourly wage exactly. So, we construct a new variable *Wage* given each respondent's money received as an employee for the preceding calendar year, the usual hours worked per week last year, and the number of weeks that the individual worked for profit, pay, or as an unpaid family worker during the previous calendar year. More precisely, we construct *Wage* for every individual as

$$\text{hourly wage} = \text{annual income}/(\text{hours/week} \times \text{weeks/year}).$$

For estimation purposes, we remove wages that are below the minimum wage level. The federal minimum wage for covered nonexempt employees is $7.25 per hour in 2021.

We notice that *Ftotval*, the total family income for each individual, and *Wage* include a very small proportion of outliers. Based on our empirical survey, it is often the case that the two variables can have some large outliers in the sample. We have also compared the performance of elementary regression models with and without outliers. The models without outliers perform considerably better than the models with outliers. Hence, we believe it is plausible to remove the unusual values using the interquartile range (IQR) technique. However, we are still reminded to follow the best-practice guidelines for outliers from Aguinis et al. [1].

In the empirical literature, researchers take the log wage as the dependent variable for two reasons. One is that the distribution of wages is severely skewed to the right. In other words, we may observe a much larger number of people who have low wages compared to individuals who are on the right tails of the distribution. This is still a major concern after removing outliers as discussed above. Therefore, considering log wages as opposed to wages centers the distribution so that preferred assumptions for modeling are more likely to be satisfied, such as normality conditions. Furthermore, from a regression interpretation perspective, a difference in log wages can be interpreted as a percentage difference in wages, while a difference in wages is the absolute magnitude difference in wages. We are more interested in wage percentage differences. Thus, we transform all wages into log wages in the dataset.

Figures 1 and 2 provide immediate histogram visualizations illustrating and verifying the skewness between variables *Wage* and *ln(Wage)* after removing the outliers. We observe that the wage distribution skews to the right as expected and that the log transformation smooths the distribution so that the normality condition can be properly assumed. Table 1 provides a summary statistics table for all relevant variables of interest. Specific variable labels and descriptions are outputted in Table 4 in Appendix A.

[1] Cases that are outside of the universe for a variable are labeled "NIU".

[2] Variable names, such as *Region*, are displayed in an italic font.

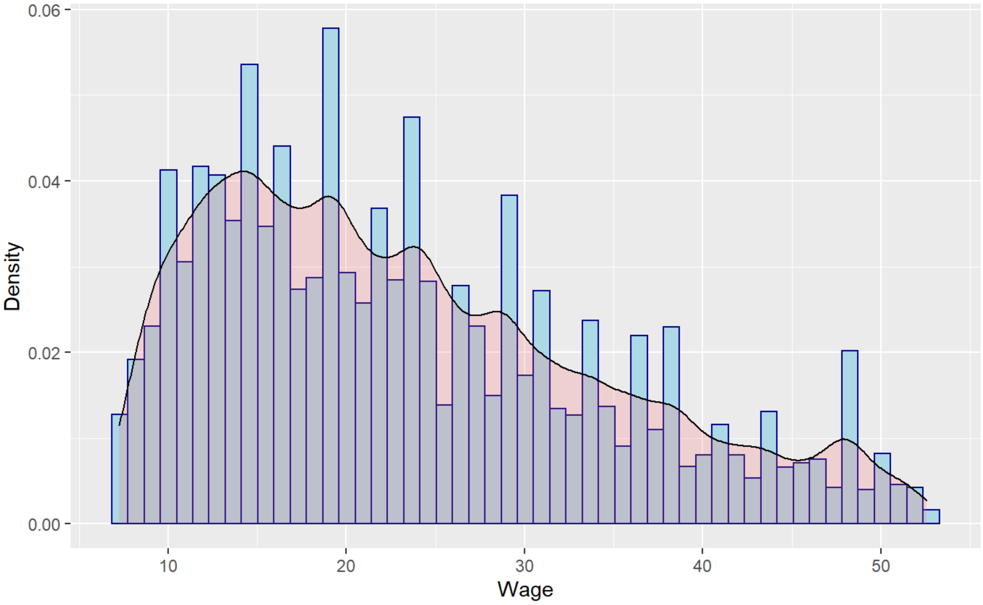

Figure 1: Histogram of Monthly Wage

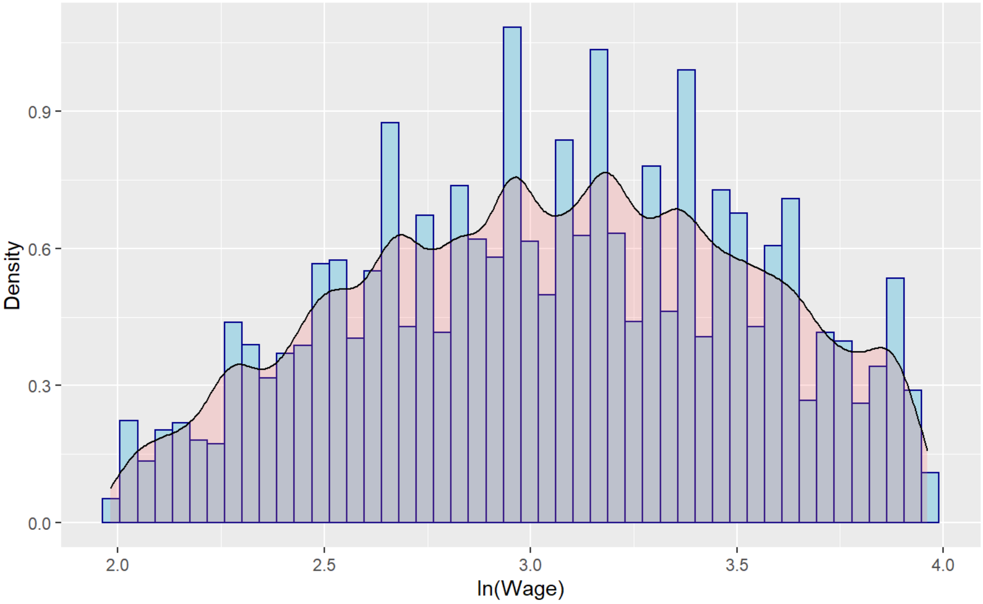

Figure 2: Histogram of Monthly ln(Wage)

4 Empirical Methodology

After an appropriate cleaning of the dataset, we can move on to select the best methodology to investigate our research question. As introduced in Section 2, our procedure is based on the empirical background literature in a combination of labor theory and econometrics. We first discuss the sample selection model in a mathematical sense in Section 4.1 and then go to the formulations of Heckman's LIML estimation methods in Section 4.2.

Table 1: Summary Statistics

No.	Name	Obs	Mean	Std. Dev.	Min	Max
1	Northeast	46247	0.149	0.356	0	1
2	Midwest	46247	0.19	0.392	0	1
3	South	46247	0.362	0.481	0	1
4	West	46247	0.299	0.458	0	1
5	Metro	46247	0.793	0.405	0	1
6	Age	46247	40.216	12.612	16	64
7	Male	46247	0.503	0.5	0	1
8	White	46247	0.76	0.427	0	1
9	Married	46247	0.522	0.5	0	1
10	Famsize	46247	3.111	1.591	1	15
11	Chlt5	46247	0.141	0.348	0	1
12	Labforce	46247	0.946	0.225	0	1
13	< High_School	46247	0.084	0.278	0	1
14	High_School	46247	0.28	0.449	0	1
15	Some_College	46247	0.172	0.377	0	1
16	College	46247	0.351	0.477	0	1
17	> College	46247	0.113	0.317	0	1
18	Diffany	46247	0.035	0.185	0	1
19	Ftotval	46247	90000.387	54912.6	−4966	254002
20	ln(Wage)	46247	3.048	0.479	1.981	3.962

4.1 Sample Selection Model

Sample selection is also sometimes called incidental truncation, which is a form of truncation in which an individual attempts to infer about a bigger population from a sample that is taken from a different subpopulation [9]. In our case, wages in the sample selected may only be observed for the ones that are in the labor force, and therefore, sample selection bias comes to play for the unobservable group of individuals who are not in the labor force. In static labor theory, wages are closely related to the chance that ones can be employed and therefore the selection problem may bias estimates [11]. One potential solution is to simply estimate wages using existing working individuals to predict the ones missing from the sample. On the other hand, this can lead to biased estimates for the parameters of wage functions and labor supply functions [2, 10, 13].

Hence, the sample selection model proposed by Heckman [12] has been used in research to tackle this issue. The sample selection model is also called the Tobit 2 model or the Type 3 Tobit model [3]. We begin the mathematical discussion of the methodology by the following theorem without proof.

Theorem. *If x and y have a bivariate normal distribution with means μ_x and μ_y, standard deviations σ_x and σ_y, correlation ρ, and a is a constant, then*

$$\mathrm{E}(x|\text{truncation}) = \mu_x + \rho\sigma_x\lambda(z),$$
$$\mathrm{Var}(x|\text{truncation}) = \sigma_x^2[1 - \rho^2\delta(z)],$$

where $z = (a - \mu_y)/\sigma_y$, $\phi(\cdot)$ is the standard normal density, $\Phi(\cdot)$ is the standard normal distribution *function and*

$$\begin{aligned} \lambda(z) &= \phi(z)/[1-\Phi(z)] \quad \text{if truncation is } z > a, \\ \lambda(z) &= -\phi(z)/\Phi(z) \qquad \text{if truncation is } z < a, \end{aligned}$$

and $\delta(z) = \lambda(z)[\lambda(z) - z]$.

Recall that sample selection is a kind of truncation. Based on the theorem, if the truncation is from below, the truncated mean moves in the direction of the correlation. If the truncation is from above, the truncated mean moves in the opposite direction of the correlation. Also, the sample selection problem decreases the variance since both ρ^2 and $\delta(z)$ are between zero and one.

Now we can illustrate the sample selection problem more specifically. Suppose the equation that determines sample selection (e.g., labor force participation), called the selection equation, is[3]

$$z_i^* = \boldsymbol{w_i}'\boldsymbol{\gamma} + u_i. \tag{1}$$

Furthermore, suppose the equation of our primary interest (e.g., log of wage), also called the output equation, is

$$y_i = \boldsymbol{x_i}'\boldsymbol{\beta} + \epsilon_i. \tag{2}$$

We observe the outcome y_i only if $z_i^* > 0$. Additionally, we suppose that u_i and ϵ_i have a bivariate normal distribution with zero means, standard deviations $\sigma_u = 1$ and σ_ϵ, and correlation ρ. Now, by the above theorem, the observed dependence between y and $\boldsymbol{x}$ can be written as

$$\begin{aligned} \mathrm{E}(y_i|y_i \text{ is observed}) &= \mathrm{E}(y_i|z_i^* > 0) \\ &= \mathrm{E}(\boldsymbol{x_i}'\boldsymbol{\beta} + \epsilon_i|\boldsymbol{w_i}'\boldsymbol{\gamma} + u_i > 0) \\ &= \mathrm{E}(\boldsymbol{x_i}'\boldsymbol{\beta} + \epsilon_i|u_i > -\boldsymbol{w_i}'\boldsymbol{\gamma}) \\ &= \boldsymbol{x_i}'\boldsymbol{\beta} + \mathrm{E}(\epsilon_i|u_i > -\boldsymbol{w_i}'\boldsymbol{\gamma}) \\ &= \boldsymbol{x_i}'\boldsymbol{\beta} + \rho\sigma_\epsilon\phi(-\boldsymbol{w_i}'\boldsymbol{\gamma})/[1-\Phi(-\boldsymbol{w_i}'\boldsymbol{\gamma})] \\ &= \boldsymbol{x_i}'\boldsymbol{\beta} + \beta_\lambda\lambda_i(\alpha_u), \end{aligned}$$

where we define $\beta_\lambda = \rho\sigma_\epsilon$, $\alpha_u = -\boldsymbol{w_i}'\boldsymbol{\gamma}$ and $\lambda_i(\alpha_u) = \phi(-\boldsymbol{w_i}'\boldsymbol{\gamma})/[1-\Phi(-\boldsymbol{w_i}'\boldsymbol{\gamma})] = \phi(\boldsymbol{w_i}'\boldsymbol{\gamma})/\Phi(\boldsymbol{w_i}'\boldsymbol{\gamma})$ to simplify the notation. λ is called the inverse Mills ratio evaluated at α_u, also called the hazard function for the standard normal distribution. Therefore, we have the sample selection model as

$$y_i|z_i^* > 0 = \mathrm{E}(y_i|z_i^* > 0) + v_i = \boldsymbol{x_i}'\boldsymbol{\beta} + \beta_\lambda\lambda_i(\alpha_u) + v_i.$$

To summarize, assuming that z_i and $\boldsymbol{w_i}$ are observed for a random sample of respondents, but y_i is observed only when $z_i = 1$, we have the model

$$\mathrm{E}(y_i|z_i = 1, \boldsymbol{x_i}, \boldsymbol{w_i}) = \boldsymbol{x_i}'\boldsymbol{\beta} + \beta_\lambda\lambda_i(\alpha_u). \tag{3}$$

In empirical research, Greene [9] states that z_i^* cannot be observed, but only the sign of z_i^* is observed. In other words, the disturbance variance u_i in the selection equation, Equation (1), cannot be estimated. That is why researchers assume that σ_u^2 to be 1 in the previous general

[3] Henceforth, we use a bold font $\boldsymbol{x}$ to denote a column of $\boldsymbol{X}$. Especially, $\boldsymbol{x_i}$ is the ith column of $\boldsymbol{X}$.

framework. Now, we restate the model as follows:

Probit selection equation: If $z_i^* > 0$, we have $z_i^* = \boldsymbol{w_i}'\boldsymbol{\gamma} + u_i$ and $z_i = 1$; otherwise, z_i^* and z_i are 0;
$\mathrm{P}(z_i = 1|\boldsymbol{w}_i) = 1 - \Phi(-\boldsymbol{w_i}'\boldsymbol{\gamma}) = \Phi(\boldsymbol{w_i}'\boldsymbol{\gamma})$; and
$\mathrm{P}(z_i = 0|\boldsymbol{w}_i) = \Phi(-\boldsymbol{w_i}'\boldsymbol{\gamma}) = 1 - \Phi(\boldsymbol{w_i}'\boldsymbol{\gamma})$.
Outcome equation: If $z_i = 1$, $y_i = \boldsymbol{x_i}'\boldsymbol{\beta} + \epsilon_i$ is observed; and

$$\begin{bmatrix} u_i \\ \epsilon_i \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & \rho \\ \rho & \sigma_\epsilon^2 \end{bmatrix}\right).$$

4.2 Heckman's Estimation Procedure

Intuitively, maximum likelihood estimators can be used to estimate the parameters in the sample selection model discussed in Section 4.1. However, as Heckman's [13] two-step estimation is broadly used in practice nowadays, we consider Heckman's method to estimate the parameters of interest here. Heckman's procedure is separated into two steps:

1. Using maximum likelihood, we can obtain parameter estimates of $\boldsymbol{\gamma}$, denoted by $\widehat{\boldsymbol{\gamma}}$, by estimating the probit selection equation. For each observation in the sample, compute $\widehat{\lambda}_i = \phi(\boldsymbol{w_i}'\widehat{\boldsymbol{\gamma}})/\Phi(\boldsymbol{w_i}'\widehat{\boldsymbol{\gamma}})$ and $\widehat{\delta}_i = \widehat{\lambda}_i\left(\widehat{\lambda}_i + \boldsymbol{w_i}'\widehat{\boldsymbol{\gamma}}\right)$. $\widehat{\delta}_i$ will be useful in constructing consistent estimators of ρ^2 and σ_ϵ^2.

2. Estimate $\boldsymbol{\beta}$ and $\beta_\lambda = \rho\sigma_\epsilon$, call them $\widehat{\boldsymbol{\beta}}$ and $\widehat{\beta}_\lambda$ respectively, by OLS of y on $\boldsymbol{x}$ and $\widehat{\lambda}$.

Consistent estimators of the variance σ_ϵ^2 and the correlation ρ can be obtained by

$$\widehat{\sigma}_\epsilon^2 = \frac{\boldsymbol{e}'\boldsymbol{e}}{n} + \frac{\sum_{i=1}^n \widehat{\delta}_i}{n}\widehat{\beta}_\lambda^2, \quad \widehat{\rho}^2 = \frac{\widehat{\beta}_\lambda^2}{\widehat{\sigma}_\epsilon^2},$$

where $\boldsymbol{e}$ is the vector of residuals from the OLS estimation of the second step in Heckman's method. Remark that $\widehat{\rho}^2$ is not a sample correlation and therefore can be outside of the interval from 0 to 1. Lastly, let $\boldsymbol{x}_i^* = [\boldsymbol{x}_i, \lambda_i]$, and Greene [8] derived that

$$\text{Est.Asy.Var}\left(\widehat{\boldsymbol{\beta}}, \widehat{\beta}_\lambda\right) = \widehat{\sigma}_\epsilon^2 \left(\boldsymbol{X}^{*\prime}\boldsymbol{X}^*\right)^{-1} \left[\boldsymbol{X}^{*\prime}\left(\boldsymbol{I} - \widehat{\rho}^2\widehat{\boldsymbol{\Delta}}\right)\boldsymbol{X}^* + \boldsymbol{Q}\right] \left(\boldsymbol{X}^{*\prime}\boldsymbol{X}^*\right)^{-1},$$

where $\boldsymbol{I} - \rho^2\boldsymbol{\Delta}$ is a diagonal matrix with $1 - \rho^2\delta_i$ on the diagonal and

$$\boldsymbol{Q} = \widehat{\rho}^2\left(\boldsymbol{X}^{*\prime}\widehat{\boldsymbol{\Delta}}\boldsymbol{W}\right)\text{Est.Asy.Var}\left(\widehat{\boldsymbol{\gamma}}\right)\left(\boldsymbol{W}'\widehat{\boldsymbol{\Delta}}\boldsymbol{X}^*\right).$$

For hypothesis testing, regular t-test procedures can be applied to individual coefficient estimates in both the probit selection equation and the outcome equation. We can also test whether sample selection bias is a problem in the model because, under the usual OLS setting, we have

$$\mathrm{E}\left(y_i|\boldsymbol{x}_i\right) = \boldsymbol{x_i}'\boldsymbol{\beta}.$$

Compared to the sample selection model in Equation (3), we have an extra term $\beta_\lambda\lambda_i(\alpha_u)$ in Equation (3) that can be used to test for selectivity bias. Since $\sigma_\epsilon > 0$, the coefficient of $\lambda(\alpha_u)$, $\beta_\lambda = \rho\sigma_\epsilon$, is zero when ρ is zero. That is, consider testing the null hypothesis H_0: No Sample Selection Bias vs. the alternative hypothesis H_A: Sample Selection Bias. Mathematically speaking, it is equivalent to testing H_0: $\rho = 0$ vs. H_A: $\rho \neq 0$ with a t-test statistic. Melino [17] showed that the square of Heckman's t-statistic is the Lagrange multiplier statistic.

5 Results

In the results section, we present our results using the empirical methodology discussed in Section 4. Based on Heckman's two-step procedure, we first provide the probit regression results on labor force participation in Section 5.1. Then, we present the OLS regression model with the inverse Mills ratio as an explanatory variable to predict log wages in Section 5.2.

5.1 Labor Force Participation Estimates

Table 2 reports the probit coefficient estimates of *Labforce* on a group of selected variables defined in Equation (1) using maximum likelihood estimation. We round the coefficient estimates and standard errors to three decimal places. Note that we use the centering technique on *Age* in the ML estimation. We are aware that we should not have multicollinearity in this case because Age^2 is a deterministic nonlinear function of *Age*. However, the coefficient estimates and their interpretation on *Age* and Age^2 are the same whether one uses centering while reducing the multicollinearity problem. Therefore, we apply centering here to our model. Additionally, we scale *Age* and Age^2 after centering so that the coefficient estimates will not be on a small order of magnitude. In particular, we divide *Age* and Age^2 after centering by a factor of 100. Besides the ease of reading and formulating the regression models, scaling *Ftotval* converts one's total family income into a unit of thousands, which provides a more intuitive and practical interpretation. Given the summary statistics of *Ftotval* as shown in Table 1, we scale *Ftotval* by dividing by a factor of 10,000. From a statistical analysis perspective, the centering and scaling techniques do not affect statistical inference in the regression models, which is an advantage for our analysis and interpretation.

Most results are consistent with what we expect from the empirical analysis as well as the classical labor theory briefly discussed in Section 1. For example, we anticipate that the shape of the age distribution on labor force participation would be a downward quadratic curve. A positive intercept with a positive linear term and a negative quadratic term shows the desired distribution. Also, we expect whites to be more likely to have a higher labor force participation rate than other races. The negative sign of the coefficient estimate of *White* demonstrates the relationship. Moreover, we believe that the more members in a family, the less likely that an individual will be in the labor force. Furthermore, households with higher labor force participation have a higher total family income.

On the other hand, it is surprising to see that the coefficient estimate for *Male* is not significant. We believe that men have a higher chance of participating in the labor force than women. From another perspective, the coefficient estimate for the interaction term $Male \times Married$ is significant at the level of 1%. This result is not unexpected because, in labor theory, the relationship of marital status to labor force participation behaves differently for men and women. Hence, the inclusion of the interacted term provides better goodness of fit. A natural question to ask is whether *Male* should be dropped from the regression model. Based on the hierarchical principle, we prefer not to exclude the variable *Male* in the probit outcome equation because the interaction term $Male \times Married$ is significant.

5.2 Log Wage Estimates

The log wage estimates using the least squares method in Equation (2) are outputted in Table 3. Again, we agree with most of the estimation results in the outcome model. For instance, the region where people live is very likely to influence the hourly wages they get. Also, respondents who live in a metropolitan area have a higher chance of getting higher wages due to a higher living cost.

Table 2: Probit Selection Equation

Tobit 2 model (sample selection model)
2-step Heckman / heckit estimation
46247 observations (2481 censored and 43766 observed)
32 free parameters (*df* = 46216)
Dependent variable: Labforce
Probit selection equation:

Intercept	2.053*** (0.052)
[Age-mean(Age)]/100	0.740*** (0.083)
$[(\text{Age-mean(Age)})^2]/100$	−0.144*** (0.006)
Male	0.014 (0.027)
White	0.103*** (0.023)
Married	−0.185*** (0.031)
Famsize	−0.046*** (0.007)
Chlt5	−0.076* (0.032)
< High_School	−0.403*** (0.053)
High_School	−0.260*** (0.046)
Some_College	−0.379*** (0.047)
College	−0.167*** (0.045)
Diffany	−0.616*** (0.040)
Ftotval/100,000	0.246*** (0.022)
Male×Married	0.371*** (0.042)

* The regression is estimated using R. Standard errors are given in parentheses under coefficients. Individual coefficients are statistically significant at the *10%, **5%, or ***1% significance level.

The effect of education on wages matches our expectations, as explained in Section 2. For a similar reason, as described in Section 5.1, we use the centered and scaled age as an explanatory variable in the outcome equation.

We also have an interesting observation related to *Male* and its interaction term. We conjecture that whites may have higher wages compared to other races, but the sign of the coefficient estimate

Table 3: Outcome Equation

Tobit 2 model (sample selection model)
2-step Heckman / heckit estimation
46247 observations (2481 censored and 43766 observed)
32 free parameters ($df = 46216$)
Dependent variable: ln(Wage)
Outcome equation:

Intercept	3.386***
	(0.017)
Northeast	0.012*
	(0.007)
South	−0.057***
	(0.005)
Midwest	−0.016***
	(0.006)
Metro	0.077***
	(0.005)
[Age-mean(Age)]/100	0.242***
	(0.036)
[(Age-mean(Age))^2]/100	0.027***
	(0.003)
Male	0.049***
	(0.011)
White	−0.025**
	(0.011)
< High_School	−0.390***
	(0.021)
High_School	−0.326***
	(0.016)
Some_College	−0.206***
	(0.018)
College	−0.127***
	(0.015)
Male×White	0.039***
	(0.008)
Inverse Mills Ratio ($\beta_\lambda = \rho\sigma_\epsilon$)	−2.132***
	(0.071)
ρ	−2.163
σ_ϵ	0.985

[*] The regression is estimated using R. Standard errors are given in parentheses under coefficients. Individual coefficients are statistically significant at the *10%, **5%, or ***1% significance level. $R^2 = 0.2507$.

of *White* is negative. Again, we have an interaction term capturing the dependence effect between *Male* and *White* in the outcome equation. Additionally, the interaction term *Male* × *White* is significant at the level of 1%. Therefore, it makes sense to us that the coefficient estimate of *White* is negative in this case.

Note that we have the coefficient estimate of the inverse Mills ratio significant at the 1% level. In other words, we reject the null hypothesis that we have no selectivity bias at the 1% level. This confirms our initial expectation that the selection problem arises under such research questions based on the empirical literature discussed in Section 2. Additionally, the use of Heckman's two-step method for the current problem is verified to be appropriate in dealing with the sample selection problem. On the other hand, it may appear counterintuitive to have a negative coefficient estimate on β_λ. In empirical studies, it is possibly difficult to predict whether the coefficient estimate of the inverse Mills ratio in this model is positive or negative. Previous literature, such as Mulligan and Rubinstein [20, 21], has found both negative and positive coefficients of the inverse Mills ratio under a similar framework in labor economics. Future work is needed to investigate the reason for the negative coefficient estimate in our case.

6 Conclusion

This paper concerns the application of Heckman's two-step method in log wage estimation by using the most up-to-date data from the Current Population Survey. In particular, a wide range of existing literature introduces the labor supply model and Heckman's method for correcting bias from incident truncated dependent variables. Before performing an econometric analysis, we clean the given data so that the maximum amount of information can be retrieved from regression analysis. We restate the sample selection model and Heckman's method in the literature to comprehensively present the mechanism and the central ideas in Heckman's work. The probit selection equation results provide us with positive feedback for our expectations before performing the analysis. In particular, we observe the importance of adding interaction terms to the model to improve model fitting and decrease the chance of omitted variable bias. Similarly, for the outcome equation, we have expected trends of the variable relationships and key interaction terms. Noted that the significance of the inverse Mills ratio, we conclude that sample selectivity bias exists in the data.

We have some future work that can be tried at this stage. First, other methods to estimate the sample selection model can be applied. For example, as discussed in Sections 2 and 4, we can try the methods of ML, FIML, and subsample OLS. Furthermore, it may be worthwhile to check the normality assumptions more formally, such as through hypothesis testing or more advanced graphical visualizations. If it is not satisfied, semiparametric, and nonparametric methods should be considered. More importantly, we can compare the model performance using different models and select the best model for the given problem and the dataset. Finally, additional investigation is required to interpret the negative coefficient estimate of the inverse Mills ratio.

A Variable Description

Table 4 shows the variables used in this paper with specific labels and detailed descriptions.

B Acknowledgment

The paper serves as a project for a graduate econometrics course for Spring 2022 in the Department of Economics at Stony Brook University. The author thanks Prof. Steven Stern for his experienced, professional, and helpful empirical and theoretical suggestions and comments in finishing the paper (and for providing such a great econometrics class).

Table 4: Variable Definition Table

No.	**Name**	**Label**	**Description**
1	Northeast	Northeast Region	Northeast gives if the housing unit was located in Northeast Region. If yes, it equals 1 and 0 otherwise.
2	Midwest	Midwest (formerly North Central) Region	Midwest gives if the housing unit was located in Midwest Region. If yes, it equals 1 and 0 otherwise.
3	South	South Region	South identifies if the housing unit was located in South Region. If yes, it equals 1 and 0 otherwise.
4	West	West Region	West identifies if the housing unit was located in West Region. If yes, it equals 1 and 0 otherwise.
5	Metro	Metropolitan central city status	Metro indicates whether a household was located in a metropolitan area. If yes, it equals 1 and 0 otherwise.
6	Age	Age	Age gives each person's age at last birthday.
7	Male	Male	Male gives whether each person is a male. If yes, it equals 1 and 0 otherwise.
8	White	White	White gives whether each person is white. If yes, it equals 1 and 0 otherwise.
9	Married	Married	Married gives whether each person is married. If yes, it equals 1 and 0 otherwise.
10	Famsize	Number of own family members in household	Famsize counts the number of own family members residing with each individual, including the person her/himself.
11	Chlt5	Own children under age 5 in household	Chlt5 identifies if there is own children age 4 and under residing with each individual. If yes, it equals 1 and 0 otherwise.
12	Labforce	Labor force status	Labforce is a dichotomous variable indicating whether the respondent participated in the labor force during the preceding week.
13	< High_School	Not completed high school	< High_School indicates if a respondent has not completed high school. If yes, it equals 1 and 0 otherwise.
14	High_School	High school diploma	High_School indicates if a respondent has obtained a high school degree. If yes, it equals 1 and 0 otherwise.
15	Some_College	Some college but no degree	Some_College indicates if a respondent has went to college but not completed a college degree. If yes, it equals 1 and 0 otherwise.
16	College	College diploma	College indicates if a respondent has obtained a college degree, including an asssociate's degree and a bachelor's degree. If yes, it equals 1 and 0 otherwise.
17	> College	A degree higher than college completed	> College indicates if a respondent has obtained a degree higher than college degree, including a master's degree, a professional school degree, and a doctorate degree. If yes, it equals 1 and 0 otherwise.
18	Diffany	Any difficulty	Diffany indicates whether the respondent has any physical or cognitive difficulty, as measured by an affirmative response to at least one of the CPS' six physical or cognitive difficulties (hearing difficulty, vision difficulty, difficulty remembering, physical difficulty, disability limiting mobility, and personal care limitation).
19	Ftotval	Total family income	Ftotval reports the total income for the respondent's family.
20	Wage	Wage	Wage indicates each respondent's hourly pre-tax wage for the previous calendar year.

References

[1] H. Aguinis, R. K. Gottfredson, and H. Joo. Best-practice recommendations for defining, identifying, and handling outliers. *Organizational Research Methods*, 16(2):270–301, 2013.

[2] D. Aigner. An appropriate econometric framework for estimating a labor supply function from the seo file. *International Economic Review*, 15(1):59–68, 1974.

[3] T. Amemiya. Tobit models: A survey. *Journal of Econometrics*, 24(1):3–61, 1984.

[4] F. D. Blau and L. M. Kahn. Rising wage inequality and the u.s. gender gap. *The American Economic Review*, 84(2):23–28, 2013.

[5] D. Card. Chapter 30 - the causal effect of education on earnings. *Handbook of Labor Economics*, 3:1801–1863, 1999.

[6] S. Flood, M. King, R. Rodgers, S. Ruggles, J. R. Warren, and M. Westberry. Integrated public use microdata series, current population survey: Version 9.0 [dataset]. Minneapolis, MN: IPUMS, 2021. `https://doi.org/10.18128/D030.V9.0`.

[7] A. S. Goldberger. Abnormal selection bias. *Studies in Econometrics, Time Series, and Multivariate Statistics*, pages 67–84, 1983.

[8] W. H. Greene. Sample selection bias as a specification error: A comment. *Econometrica*, 49(3):795–798, 1981.

[9] W. H. Greene. *Econometric Analysis.* Pearson, 8th edition, 2018.

[10] R. Gronau. The effect of children on the housewife's value of time. *Journal of Political Economy*, 81(2):S168–S199, 1973.

[11] J. J. Heckman. Shadow prices, market wages, and labor supply. *Econometrica*, 42(4):679–694, 1974.

[12] J. J. Heckman. The common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimator for such models. *Annals of economic and social measurement*, 5(4):475–492, 1976.

[13] J. J. Heckman. Sample selection bias as a specification error. *Econometrica*, 47(1):153–161, 1979.

[14] E. Lazear. Age, experience, and wage growth. *The American Economic Review*, 66(4):548–558, 1976.

[15] H. G. Lewis. Comments on selectivity biases in wage comparisons. *Journal of Political Economy*, 82(6):1145–1155, 1974.

[16] M. F. O. Martins. Parametric and semiparametric estimation of sample selection models: An empirical application to the female labour force in portugal. *Journal of Applied Econometrics*, 16(1):23–39, 2001.

[17] A. Melino. Testing for sample selection bias. *The Review of Economic Studies*, 49(1):151–153, 1982.

[18] E. Moretti. Real wage inequality. *American Economic Journal: Applied Economics*, 5(1):65–103, 2013.

[19] T. A. Mroz. The sensitivity of an empirical model of married women's hours of work to economic and statistical assumptions. *Econometrica*, 55(4):765–799, 1987.

[20] C. B. Mulligan and Y. Rubinstein. Selection, investment, and women's relative wages since 1975. *National Bureau of Economic Research*, (11159), 2005.

[21] C. B. Mulligan and Y. Rubinstein. Selection, investment, and women's relative wages over time. *The Quarterly Journal of Economics*, 123(3):1061–1110, 2008.

[22] F. D. Nelson. Efficiency of the two-step estimator for models with endogenous sample selection. *Journal of Econometrics*, 24(1):181–196, 1984.

[23] W. K. Newey, J. L. Powell, and J. R. Walker. Semiparametric estimation of selection models: Some empirical results. *The American Economic Review*, 80(2):324–328, 1990.

[24] P. A. Puhani. Foul or fair? the heckman correction for sample selection and its critique. a short survey. *ZEW Discussion Papers*, 97(7), 1997.

[25] P. A. Puhani. The heckman correction for sample selection and its critique. *Journal of Economic Surveys*, 14(1):53–68, 2002.

[26] F. Vella. Estimating models with sample selection bias: A survey. *The Journal of Human Resources*, 33(1):127–169, 1998.