ARE ITEM NON-RESPONSE PROBLEMS IGNORABLE IN SURVEY DATA?

AN ANALYSIS USING CURRENT POPULATION SURVEY DATA

by

Jie (Jay) Yang

(5488536)

An Essay Submitted to the Department of Economics
in Partial Fulfillment of the Requirement for
the Degree of Master of Arts

Supervisor: Professor Pierre Brochu

ECO 7997
University of Ottawa
Ottawa, Ontario, Canada
December 2012

Abstract

This paper analyzes the wage non-response problem using the 2011 Current Population Survey. I find that the increasing wage non-response cannot be ignored, as including the imputed wages in a OLS regression leads to an imputation bias. In addition, the imperfect match criteria, such as not using union status to find a hot donor, is a main source of the imputation bias. I conclude that dropping the imputed values provides better estimates than simply including the imputed wage values in a wage equation.

Keywords: hourly wage, item non-response, hot deck imputation, non-ignorable, bias, confidentiality, selection model.

1 Introduction

High quality survey data is very important for decision makers in this information society, and missing data due to non-response can negatively affect data quality and related economic inference. More and more respondents now refuse to answer some survey questions, like those on income and wages (e.g. high non-response rate in the Current Population Survey (CPS) for hourly wage and working hour questions). Moreover, standard econometric theories are usually developed to analyze full samples, but not well developed for dealing with a sample with missing values (Little & Rubin 2002). To create a full sample, statistical agencies will allocate values for missing data (i.e. use imputation) based on the available information.[1] The approach used towards missing and imputed values in a dataset can make a big difference empirically. Without careful analysis, using imputed values could be worse than simply dropping the non-response.

Imputation methods are widely used in many large scale datasets, but they may not actually solve the non-response bias problem. Andridge & Little (2010) argue that the theories behind the most popular hot deck imputation methods are not well developed.[2] Researchers should be careful to use imputed values. Hirsch & Schumacher (2004) show that the current imputation method in the CPS is far from perfect, and may lead to bias. Bollinger & Hirsch (2006, 2010) argue that the coefficient bias from using dependent variable imputation (i.e. imputed wage) can be worse than simply dropping the observations with missing data.

[1] In US datasets, such as the CPS, they refer to imputation as "allocation". In this paper, I will use the imputation terminology.

[2] Hot deck imputation is a commonly used method for dealing with missing values in surveys. The basic idea is to replace the missing values by using the data from an observed response that has the similar characteristics as the non-response. In section 4.4, I will provide more detailed procedures and examples for the method.

In this paper, I update the evidence of Bollinger & Hirsch (2010). More precisely, I use CPS data for 2011 to explore the impact of missing earnings information, and the various approaches for dealing with this issue, when estimating human capital models. I believe that wage non-response can lead to serious biases, and the current imputation methods may not solve the problem, so it is worthwhile to systematically study the wage non-response issues.

My main finding is that the wage item non-response problem in the CPS is not ignorable, and dropping the individuals who do not report their wages will lead to a selectivity bias. In addition, the use of selection models suggests that the imputation bias from using imputed wage in the CPS is larger than the original non-response bias from dropping wage non-response observations.

The rest of the paper is organized as followed. Section 2 provides some background information of non-response issues. Section 3 reviews some theoretical and empirical literature on missing data. Section 4 discusses the CPS data used in this paper, and section 5 introduces the econometric models. Section 6 provides an analysis of the empirical results. In section 7, I carry out robustness checks, and the section 8 concludes.

2 Non-response Issues

2.1 Types of Non-response

There are two types of non-response in household surveys. The first type is called unit (total) non-response. This is the case where the individual does not respond to the survey at all. The second type of non-response, which I focus in this paper, is called item (partial) non-response.

The respondent answers some of the survey questions, but not others. Wage non-response is a common example of item non-response, and will be the focus of this study. Many survey respondents are not willing to provide wage (or earnings) information.[3] In such cases, the statistical agencies will not drop the individual from the survey, as partial information is better than no information. It does mean, however, that the researchers will be faced with an incomplete dataset (or one with imputed data). If non-response is not random, and there is no reason to believe it is, it may bias the researcher's empirical results.[4]

2.2 Reasons for Wage Non-response

The main reason for wage non-response is probably concerns of confidentiality, as well as the fear of sharing very personal information (Bollinger & Hirsch 2006; Card, Chetty, Feldstein & Saez 2011; Huisman 1999; Little & Rubin 2002). On the one side, researchers want more detailed data; they want the surveys to ask more specific and detailed questions. The more detailed questions raise confidentiality concerns, which leads to a lower response rate. In addition, asking more questions increases the response burden because respondents have to spend more time to answer the survey. There is an increasing non-response problem in many surveys because of increased response burden (Bethlehem, Cobben & Schouten 2011). For example, the CPS non-response rate increased dramatically with the introduction of the new survey questionnaire in 1994. Before 1994, the non-response rates were around 15%. Following the introduction of the new questionnaire, non-response rates approximately doubled in size. Moreover, Figure 1 shows that the CPS non-response rate problem is getting worse. By 2011, the non-response rate reaches almost 36%.

[3] Non-response could also occur because the individual does not know the answer to the question.
[4] In the literature review section, I provide a more detailed explanation of why it may lead to biased results.

2.3 Practices to Reduce Survey Burden

The best way to deal with the non-response problem is to prevent non-response from happening at the start, which means a better survey design, and a thorough attempt to get answers from each interviewee at the operational level.[5] To reduce the burden of the survey and to reduce non-response, statistical agencies have tried different approaches.

Besides re-contacting the interviewees, the statistical agencies sometimes try to get access to external datasets. For example, the Survey of Labour Income Dynamic (SLID), a major source of Canadian income data, asks interviewees for permission to access their tax file to get their income information (Burbidge, Magee & Robb 2003). As a results, the interviewees do not have to answer 13 income related questions, which reduces the response burden. However, because the individuals must provide consent, some will still have confidentiality concerns, which in turn will lead to non-response.

US statistical agencies also have tried to link survey data to administrative data. At present, only administrative data that relates to Medicare and Medicaid services are widely accessible to researchers. Card, Chetty, Feldstein, & Saez (2011) suggest expanding access to other administrative datasets in the US. They believe that secure access to administrative data is very important for research because it has significantly less non-response problems (or measurement error) than the traditional survey data. However, the request of accessing (and linking) to the administrative data raises important confidentiality concerns, and as such most economic researchers still cannot get the access to the tax record or social security record. Moreover, administrative data also has some limitations. For example, it may not have

[5] The operational level means interviewers' practical effort during the data collecting process, such as to be friendly and professional during the interviews, or re-contacting the interviewees again when they are not responding on time.

enough details, and may also have some measurement errors. Thus, these efforts are not the most effective ways to solve non-response problems.

Statistical agencies have also tried other approaches to allay these fears of sharing personal information. Some confidential data only can be accessed at a secured location, and only by the qualified researchers. For example, Canadian master files survey data can only be accessed at the local Research Data Centre (RDC). In some cases, the statistical agencies have kept the public access to the data file, but to ensure confidentiality, they have switched entries of some individuals that are statistically similar; an approach favored in the US. Alternatively, the statistical agencies may collapse certain information into categories. For example, income is sometime collapsed into intervals of $10,000 or $20,000. As such, the researchers do not get the actual income values. These efforts may allay certain fears of sharing information, but there will always remain a significant minority of respondents who refuse to answer certain questions.

2.4 Imputation Practices

Using a dataset with missing values can lead to serious problems because regression analysis will require that individuals with missing observations be dropped from the sample. In such a case, one will only be keeping employees that have reported wages, and the employees without wage values will be out of the sample. As a result, much information is lost, and the remaining sample may not be representative of the population of interest because of non-response (Green & Milligan 2010).

Alternatively, the statistical agencies may use some imputation methods to impute values to the non-response, and hot deck imputation is the most popular method. Historically, the term "hot deck" refers to the deck of computer punch cards for donors that are for non-response, and "hot" means it use the current processing data rather than the pre-processed or other source "cold" data as donors (Andridge & Little 2010). The donor, or reference household for each non-respondent, is the most recent person surveyed in the survey with the similar characteristics (e.g. gender, age, occupation, and education). These characteristics that are used to find the hot donor are called imputation criteria or match criteria, and these criteria are quite different in different surveys. If a wage non-respondent has the same imputation criteria values as a donor (e.g. the same gender, occupation, and education), the donor's wage value will be the imputed wage value for the non-respondent. The old donor will be replaced when a new matched donor appears in the survey.

US statistical agencies use hot deck imputation methods for American surveys, like the CPS, and most importantly they also include flags to identify which values were imputed. The SLID also uses a hot deck imputation method for wage non-response in Canada, but it does not provide imputation flags. It means researchers cannot identify which observations have imputed wages. As a result, researchers who use this Canadian dataset cannot drop the non-response in their samples; they are forced to rely on imputed values.[6] For the other two major Canadian wage data sources, the Labour Force Survey (LFS) and the Public Use Census Micro File (Census-PUMF), they do not provide imputation flags in their public use files either. Given that I am interested in exploring the implication of using imputed data versus dropping these observation, I must rely on US data. More specifically, I will rely on CPS data for 2011.

[6] Without imputation flags, researchers cannot identify which wage was reported and which one was imputed. This is a limitation of the SLID and other Canadian datasets.

3 Literature Review

The literature review consists of four parts. In the first part, I review the theoretical literature that analyzes the non-response problem. In the second part, I examine the wage non-response literature that focuses on the CPS. The third part looks at selection models which I later use to address the wage non-response problem. In the final part, I cover papers that focus on measurement error related to construction of the hourly wage.

3.1 General Non-response Literature

Bethlehem, Cobben, & Schouten (2011) provide a theoretical overview of non-response issues. They argue that non-response is positively associated with the size of the estimated bias; the higher the non-response rate, the larger is the bias. Larger surveys also increase the response burden (so more than small surveys), which may increase the non-response rate, and finally lead to more bias. If the missing data is random, a model without the imputed variables will not influence the quality of the estimated coefficients in regression analysis, but the variances may be larger. However, if the missing data is not random, which is usually the case in survey data, the estimator will be biased. There is no guarantee that an imputation approach will reduce the problem. The authors believe that the bias reduction depends on the characteristics of the non-respondents, and the specific imputation procedure used.

Little & Rubin (2002) summarize the literature addressing the missing data issue, and briefly review the statistical analysis of non-response problems since 1970s. They find that many statistical questions can be viewed as missing data problems, and the related non-response research is a useful way to study statistics in general. In addition, they point out that it is more difficult to solve the missing data problems in small samples. They also provide precise definitions and detailed proofs of different methods to deal with missing data problems (e.g. maximum likelihood approach, imputation, re-weighting, and modeling). However, those methodologies are not widely used by empirical researchers because the theories are often too complicated to implement in practice, and every dataset has its own specific non-response pattern.

Andridge & Little (2010) review different forms of hot deck imputations that attempt to resolve non-response bias, using both the CPS and the third National Health and Nutrition Examination Survey (NHANES III). Although hot deck imputation methods for dealing with item non-response are widely used in many large scale surveys, the theory behind these methods is not well developed. For example, some important variables (e.g. union status) are not in the imputation criteria, and the choice of these criteria is based on the preference of statistical agencies rather than the economic theory. The conceptual literature that support the hot deck imputation is very limited. It is widely used simply because the methods are relatively easier to implement than other alternatives. If a researcher simply uses the imputed data without looking at the characteristics of the non-respondents and the imputation procedures, this may lead to a more serious bias.

Durrant (2009) reviews many imputation methods in social science studies, and points out the difficulty in making the right choice among various alternatives. He suggests that statistical agencies or primary data users should consider the estimator of interest, the missing data pattern, the properties of different imputation methods, and the type of data available before using any particular imputation methods. It means the imputation procedure should be flexible for specific missing data and particular datasets. He finds that a multiple imputation approach, which provides a choice among different imputed values, is better than the single imputation system. In addition, he argues that statisticians should make more development on various imputation methods, and make the imputation information more easily accessible to researchers. If researchers know the specific procedure of a imputation method, they can make a better choice between using or dropping the imputed values.

3.2 CPS Imputation Literature

Hirsch & Schumacher (2004) find an increasing non-response rate in the CPS, and briefly summarize the historical changes of different CPS data files and their imputation methods. They combine the May CPS from 1973 to 1981 and the CPS Outgoing Rotation Group files (CPS-ORG) from 1983 to 2001 to analyze the wage gap between union members and non-union members. They also use the March CPS from 1996 to 2001 carrying out the analysis by industry, and other sector wage.

They criticize the current hot deck imputation procedure in the CPS because it does not include all relevant variables when finding the hot donor. For example, a non-respondent need the same gender, occupation to find a hot donor, but does not need the same union status to find the donor. They find that the wage gap estimates are negatively biased when union status is not a match criterion. The size of the bias depends on the size of union membership in the non-response group. They provide a framework to analyze the match bias when a variable is not used in the imputation criteria. Bias of other non-match criteria such as public sector, industry are also discussed in the paper.

Bollinger & Hirsch (2006) extend the analysis of Hirsch & Schumacher (2004), and find that even if non-response is random, it will still lead to some bias. In addition, not only the choice of match criteria (e.g. age, race, occupation, and education vs. union status), but also the measurement of criteria (e.g. values vs. categories) make a big difference on the imputed values. They use non-student wage and salaries of workers 18 years and over from the January 1998 to December 2002 CPS-ORG file. In their full sample, 28.7% of men's earnings are imputed, and 26.8% of women's earnings are imputed. They find a larger match bias for men because men have a higher non-response rate than women.

They use age and education as examples, and find that even if a particular variable is used in the imputation process as a match criterion, the estimated coefficient may still have some match bias. Age and education both are used in the imputation process to find the hot donor. However, they are not in absolute value form, and the CPS hot deck cells only include six age categories and three education categories. The imperfect measurement will generate a match bias in the imputed wage data. Therefore, simply using the imputed values in the CPS will lead to some bias when a researcher estimates the wage gap. This paper also provides some theoretical framework to solve the non-response problem. For example, dropping the imputed values, and reweighting the respondent-only sample is a good approach. However, these approach are very complicated, and not very practical in nature.

Bollinger & Hirsch (2010) expand on their previous works by using more current CPS data (up to 2008), and by introducing some selection models to deal with the non-response issues. They treat wage non-response as a self-selection problem, and find that there is a negative selection into non-response among men for the wage question, but there is nothing significant for women. Because of the negative non-response bias, the average predicted wage value for men is understated by 9% if compared to the results from the sample that drops wage non-respondents, and is understated by 2% for women. In short, they believe that non-response or the imputation is not ignorable in the CPS wage data.

They also find proxy reporting (i.e. reporting by a reference person, such as a family member) and interview time (i.e. interview during tax months) have impacts on the wage response rate, but not directly influence the wage values. Thus, they use the two variables as the exclusive restrictions for the selection equations (e.g. will not appear in the wage equations).[7] They find that including the imputed wages in a regression leads to a larger bias than just dropping the non-respondents in the model.

3.3 Selection Model Literature

Because selection model is not a straightforward econometric model, and it is not original designed for wage non-response problem. To get a general idea of the selection model, I also review some other papers.

Tobin (1958) is the first paper to deal with limited or unobserved dependent variables. Although he does not directly deal with non-response problem, but wage non-response can be treated as a limited dependent variable problem. He examines these issues in the context of the relationship between household income and luxury spending. He wants to include the observations with zero spending (i.e. unobserved self-selection process), and so introduces a latent equation to solve the problem, which is the main idea of the Tobit model. The original model has some limitations, and cannot be used to solve the wage non-response problem directly, but the basic idea is similar. His idea of introducing a unobserved variable in the econometric model is a very important starting point for the later selection models. That is why the selection model (also called Type II Tobit model) is a more general form of the original Tobit model.

[7] The exclusive restriction is related to identification problem in the selection model, which I will discuss more in the Econometric Model section.

Heckman (1976) is the original paper for the actual selection model, and the model can be used to solve wage non-response problem. He uses women labour supply as an example to introduce the selection model. In this model, there are two latent or unobserved equations. One is for the women's decision to participate in the labour market, and the other is the latent wage equation for every women, which include the unobserved wages from non-participants. He proposes a maximum likelihood approach for dealing with such issues. In a following study, Heckman (1979) introduces a two step solution for the same selection issue, which is less computational demanding and easier to use. Although researchers have made some progresses on the basic selection model, the basic idea from the two original paper does not change much.

Lee & Marsh (2000) introduce a new estimation method to solve the selection issue, and correct the non-response bias in the Panel Study of Income Dynamics (PSID). They find that people who just quit a job are more likely to not report their job status than other people, and dropping the non-respondents can lead to a selectivity bias. In addition, they argue that different distribution assumptions of the error terms can materially affect the results. They propose a simpler solution by implementing a new maximum likelihood estimation function. However, their empirical results do not clearly support their theoretical prediction. The selection model provides very similar results with the OLS model that drops all the non-respondents.

3.4 Hourly Wage Literature

Bound, Brown, & Mathiowetz (2001) provide an overview of the measurement error literature. They believe that economic researchers need to be more concerned about the measurement error bias given that they are relying more and more on micro data. Income and earnings measurement error is a very important issue, and can result from many sources such as reported error, non-response, and imputation. Wage non-response bias is a important part of the wage measurement error because the missing wages lead to an incomplete sample, which cannot fully represent the population of interest.

They find the reported annual earnings have less measurement error than the weekly or hourly wage rates. In addition, they mention that the hourly wage usually needs to be calculated from earnings and working hours in many datasets. Errors from both reported earnings and reported hours will result in a higher level of response errors for the hourly wage. Therefore, directly reported hourly wages will have less measurement errors than the calculated ones. I will account for this issue when carrying out robustness checks.

Card, Lemieux & Riddell (2004) use three datasets (US, Canada, and UK) to analyze the union impact on wage inequality. Because they rely on quite different datasets, they provide a detailed discussion of data issues. For the Canadian LFS, they criticize Statistics Canada for not providing imputation flags, and for not using union status when finding an imputation donor. They find a downward bias in the estimated effect of unions, when union status is used as a match criteria in the imputation process. It is a similar negative non-response bias to what Hirsch & Schumacher (2004) find.

Lemieux (2006) provides more detailed methods to deal with the wage measurement and censoring issues in the CPS. He examines the determinants of increasing residual wage inequality in the US from 1973 to 2003. He finds that the choice of CPS files can dramatically alter the findings. He argues that using the supplement data (CPS-May/ORG) is better than using the regular March data (March CPS) to analyze the wage differential questions. Moreover, he believes that the hourly wage is the best measure for wage analysis because it can be used as a standard price for labour. Throughout his analysis, Lemieux (2006) drops individuals with imputed values. As such, he does not compare the benefits or costs of relying on imputed data, unlike what I do in my study.

4 Data

4.1 Data Source

This study uses data from the 2011 Merged CPS-ORG File (CPS-MORG). The CPS is an American monthly household survey with public micro-data available since 1962, and is the most important government source of US labour market information. The CPS interviews approximately 47,000 households every month, and provides detailed labour status, income, employment, spending, wealth, and education among others.

Every household in the CPS is interviewed once a month for four consecutive months, followed by eight months out of the survey, and the household is then re-interviewed for an additional four months. After a total of eight months in the CPS, the household will be permanently out of the sample. Therefore, there are always eight groups of households in a regular monthly CPS sample. The observations in the 4th and 8th months, which are referred to as the "outgoing rotation groups" (ORG), have additional information.

Since 1979, only households in their 4th and 8th months (or the ORG) are interviewed for the usual working hours and wage rates, and the Bureau of Labor Statistics (BLS) puts them in a single yearly file called the CPS-MORG. Therefore, a household will enter twice in the

ORG files (in different years), but the annual CPS-MORG will not include the same household twice in the file.[8] Before the 1980s, there was no CPS-MORG, so researchers had to rely on the May CPS if they want longer data series.

Another widely used CPS file is the March CPS, which is also named the Annual Demographic File (CPS-ADF). It provides annual demographic information as the monthly CPS does, but provides more information on working experience, income, social benefit, and migration, which is similar to the SLID in Canada. Many research papers include both the March and ORG files in their analysis, so they can make some comparison between the results from the two files at the same period. Unlike the CPS-ORG which collects information from the last week (reference week), the CPS-ADF collects data for the previous year. Another difference is that the CPS-ADF does not contain union status. Lemieux (2006) suggests that the CPS-ORG has more advantages to deal with wage issues than the CPS-ADF, and as such this study will use the most recent CPS-MORG file.

4.2 Population of Interest

My population of interest consists of individuals that worked in the reference year, even if they are not currently working in the reference week. I drop all the self-employed workers. Although some of the self-employed workers report their wage rates in the CPS, they are not the same as wage rate of normal employees due to the special characteristics of self-employed workers. These restrictions generate a sample of 166,259 observations. However, not everyone has an hourly wage data available, so the final sample will be smaller (165,744

[8] For example, a household that enters the CPS in January 2011 will provide wage data in April 2011 (the 4th month in the survey), and in April 2012 (the 8th month in the survey). As a result, this household will enter the annual CPS-MORG both in 2011 and 2012, but each year CPS-MORG file will only include the same household once.

observations). In Appendix 1, I provide a flowchart that provides a detailed exposition of how I generate the hourly wage, and how I define the imputation rate.

4.3 Hourly Wage Rate

I chose the hourly wage rate as the income measure in my econometric model because it is the price of labour. The potential reported bias from using the hourly wage is not the key concern in my study, as I focus on the non-response bias. Not everyone reports hourly wage in the CPS. For those individuals for whom I only have their weekly earnings, I have to use working hours to get their hourly wages. In my final sample, every observation should have a valid hourly wage values, either reported from the respondents or imputed from the BLS.

4.3.1 Paid Hourly

In the flowchart of Appendix 1, there are 98,900 individuals that are paid by the hour, and only 33 individuals that do not have any reported or imputed wage in the CPS. 1,358 individuals who reported being paid hourly declared that their wage rate is below the lowest minimum wage (i.e. $5.15) in the US.[9] I treat the 33 special cases and the 1,670 individuals who report lower than the minimum wage, as if they were not paid by the hour, so I only have 97,509 households paid hourly. For this group, I can use the imputation flag (called "I25c" in the CPS) to identify who are wage non-respondents (see Appendix 2).

4.3.2 Not Paid by Hours

For the rest of the sample, (which includes the 67,359 not paid by the hour, the 33 special cases, and the 1,358 who reported a wage lower than the minimum wage), I use the weekly

[9] US federal minimum wage is $7.25 in 2011, but every state has its own minimum wage, and I use the lowest state level minimum wage ($5.15).

earnings divided by the working hours to get an hourly wage rate. For usual earnings, everyone has a value in the data, but some of the values are zero.

For working hours, there are two variables in the CPS: usual working hours, and last week's working hours. If a household has information on usual working hours per week, I use his or her reported weekly earning divided by their usual working hours to get the hourly wage (64,858 individuals). For the others, I divide the reported weekly earnings by their last week's working hours (3,665 individuals). There are also 227 individuals who do not have positive working hours, so they will not have a hourly wage, and as such, I drop these observations. I also drop 288 individuals who have a zero hourly wage. I am left with 68,523 new wage data in this group. In the end, I have 165,744 observations in my final sample.

4.3.3 Low End of Hourly Wage

In my econometric model, I use the log wage because Cameron & Trivedi (2009) points out that a log-normal distribution is more appropriate, and many empirical researchers including the pioneering work of Mincer (1974) use the log wage. Figure 2 shows that the log wage is more like a normal distribution than the regular wage. However, all wage values below $1 will have a negative natural log value, and a $1 wage will have a zero log value. Thus, I change all the wage below or equal to $1 to $1.0001 to get all the log value above 0. Adding a small value to get positive values for the dependent variable is common in applied econometrics (Cameron & Trivedi 2009).

4.3.4 Top-coding Wage Rate

The top coding in the 2011 CPS is $99.99 for the hourly wage and $2,885 for weekly earnings. Top coding is another source of measurement bias for wage equation. For example, if someone's actual hourly wage is $200 (over the top-coding), but it only can be recorded as $99.99, and the lower recorded values can lead to a negative bias for estimated values in a wage equation. Therefore, I need to make adjustments to the final hourly wage data. Like Bollinger & Hirsch (2006), I assume that the top coded wages are from a Pareto distribution, and the estimated mean values can be accessed from a published index.[10] According to the index, for those males whose hourly wages are equal or over $99.99, I trim the values to $182.98 (99.99 times 1.83); for the females, the top coded wages become $164.98 (99.99 time 1.65).[11]

4.4 Wage Imputation

4.4.1 Imputation in the CPS

Figure 1a shows that the CPS is experiencing an increasing non-response rate for the wage question, with the non-response rate now standing around 30%. The classical hot deck imputation method in the CPS was introduced at 1947 for the income non-response in the Income Supplement file. Hirsch & Schumacher (2004) provide details of the historical changes in the CPS imputation methods. The current procedures have changed significantly overtime, and the various CPS files have quite different hot deck imputation methods.

For the nonresponsive values for the wage questions, the Census and the March CPS use "sequential hot deck" imputation methods, and the CPS-ORG uses a "cell hot deck" procedure. The most recent "cell hot deck" procedure creates cells are based on seven criteria: gender (2 categories, i.e. male or female), age (6 categories), race (2 categories), education (3

[10] The index values can be accessed at: http://www.unionstats.com/.

[11] For more detailed information on issues related to wage top-coding in the CPS, Lemieux (2006) has a appendix file in his paper.

categories), occupation (13 categories), hour worked (8 categories), and receipt of tips, commissions or overtime (2). This generates a combination of 14,976 cells with assigned donors which have reported wage values. In addition, since the categories may changes over time, the number of assigned donors changes over time. To identify the household with imputed earning information, the CPS uses four allocation flags, which I give a brief introduction in the Appendix 2. To find the final non-response or imputation rate, I have to use the four flags after I finish the data generating process.

4.4.2 Imputation Rate

There are two major parts to the imputed wage rate. For the people who are paid hourly, I use one imputation flag (called "I25c" for hourly wage) to identify non-response rate. 63.05% of them report their hourly wages, and for the remaining 36.95%, the CPS imputes their wage. For those not paid hourly, I used another two flags (called "I25d" for weekly earnings, and "I25a" for usual hours) to identify the imputed values. Moreover, for those that I change their wages because of the low values (less or equal to $1) or the top-coding (equal or more than $99.99), I will not treat the changes as imputation. 33.89% individuals who are not paid by hours, have imputed wage rates. As a result, I have a total 59,191 (35.69%) observations with imputed wage rate in my final sample.

4.5 Other Variables

For the experience variable, there is no direct data available in the CPS. Therefore, I use Mincerian experience which consist of age minus year of schooling minus 6. However, not every household has positive experience data (although they should have as I am only dealing with those currently working). Instead of dropping these individuals, as some researchers do, I use the following procedure: For the 1,229 individuals with negative (-1 to -6) experience, I assign them 0.1 years of experience. For the remaining 5,849 individuals with 0 experience, I assign them 0.5 years of experience. Given that they represent a small portion of the total sample, the choices to drop or to keep them does not materially affect my findings.

I create dummy variables for gender (female), race (non-white), age (being young), working in the public sector, and working part-time.[12] In addition, I also create categorical variables for occupation and highest education attainment. Table 1 provides the detailed description of these variables.

4.6 Summary of Statistics

Columns (1) and (2) of Table 2 show summary statistics for those individuals who provided wage information and those that did not, respectively. 106,593 individuals reported wages, and 59,191 are non-respondents (35.69%), i.e. have imputed wage rates. For ease of exposition, I will focus on the percentage difference which is presented in column (3) of Table 2. Because the variables are dummies (values are less than 1), and the sample size are very large, so their variance are very small. Therefore, even though the differences are small, they are all statistical significant.

There is no clear non-response pattern for the gender, marital status, public sector, union status, and occupation, and their percentage differences between non-respondents and respondents are less than 10%, which is relatively smaller than the 20% or 30% for other

[12] Being young implies an age between 15 to 29.

variables such as race and education. In addition, married female working in the public sector with union membership are slightly more likely to answer the wage question in the CPS.

There are some economically significant differences when it comes to education, age, race, and part time working status. For the five education categories, high school diploma holders are the only group with higher non-response rate (i.e. 11.4%). Since high school diploma holders is the base group, the estimated coefficients of education returns may be biased. In addition, young people are much more likely to response the wage question (10.4% less non-response), but the older individuals are not (38.5% more non-response). Moreover, white individuals are more likely to response than the non-white. As a group of the non-whites, 34.5% more black people are in the non-response group. Moreover, part-time workers are much more likely to provide their earnings information than full time workers, and there are 21% less part-time workers in the non-response group.

There is 25.8% difference for the proxy response, which means people whose survey answers are provided by other family members (i.e. are not directly provided by themselves) are more likely to not answer the income questions. Because answering earnings questions requires more private information than gender, occupation, and other simple questions, it is not surprising that proxy report households have a high wage non-response rate. Moreover, if the income interviews are in the tax months (February and March), people are more likely (17.7% less non-response) to answer income questions. In addition, people who are in their 4th interview (e.g. the first time to answer the earnings and working hour questions) have a higher non-response rate. These three variables will be important variable in my selection models because they may not have much influence on the real wage rates, but they do affect people's willingness to respond to the wage questions.

Table 2 shows that the mean values of many key variables are quite different, so the non-respondents do not have the same pattern as the respondents in the CPS, and their wage distribution will not be the same. When the characteristics of the non-respondents are different from the observed respondents, the non-response is not random, and I have to use some econometric models to correct the selectivity bias.

5 Econometric Model

This paper includes a simple ordinary least square (OLS) model, and a selection model. The OLS model will be estimated in two ways: First using only individuals with reported wages, which means I drop the imputed values. The second approach includes individuals with imputed wage values. If non-response has some special pattern and bias, I need a selection model to consider the bias. For the selection model, I will provide two solutions. I will also compare the predicted wage rates from both OLS and selection models.

5.1 Simple OLS Model

My econometric model of the wage equation takes the following form:

$$lnwage_{\mathrm{i}} = \beta_0 + edu_i^{'}\beta_1 + \beta_2 exp_{\mathrm{i}} + \beta_3 exp_i^2 + per_i^{'}\beta_4 + job_i^{'}\beta_5 + \boldsymbol{\varepsilon}_{\mathrm{i}} \qquad (1)$$

where $lnwage_{\mathrm{i}}$ is the log wage of individual i. The vector edu_{i} represents educational attainment.[13] exp_{i} is Mincerian experience as described in section 4.5, and exp_i^2 is the square

[13] The category includes below high school, high school, college graduate, bachelor degree, and graduate schooling. As is commonly done in the literature, the base group is high school graduates.

of experience. Mincer (1974) uses this structure to capture the concavity of the experience effect on the wage rate.

Besides the basic education and experience structure of the Mincer equation, I add two additional sets of controls: person specific controls and job specific controls. The person specific controls (per_{i}) include dummy variables for age (being young), race (non-white), and marital status (married). The job specific controls (job_{i}) include dummy variables for occupation, public sector, part time status, and union status.[14] More details on these variables can be found in Table 1. For ease of exposition, when discussing self-selection issues, I rewrite equation (1) as:

$$lnwage_{\mathrm{i}}=\mathbf{X}_{\mathbf{i}}'\boldsymbol{\beta}+\varepsilon_i \qquad (2)$$

where $\mathbf{X}_{\mathbf{i}}'$ just represents the vector of all explanatory variables in the equation (1).

5.2 Selection Model

5.2.1 Selection Equations

The selection model follows the structure of Heckman (1976, 1979), and has two parts. In the first part, I assume there is latent utility variable ($\mathrm{U}_{\mathrm{i}}^*$) to determine whether the interviewees answer the wage question or not in the CPS.

$$\text{Selection Part:}\begin{cases}\mathrm{U}_{\mathrm{i}}^* = \mathbf{Z}_{\mathbf{i}}'\boldsymbol{\gamma} + v_i^* & \text{(3a)}\\ resp_i = 1 \;\; if \; \mathrm{U}_{\mathrm{i}}^* > 0 & (3b)\\ resp_i = 0 \;\; if \; \mathrm{U}_{\mathrm{i}}^* \leq 0 & (3c)\end{cases} \qquad (3)$$

[14] The occupation categories include managers, business workers, professional workers, basic service and sales, agricultural workers, manufacturing workers. The base group consist of workers in the manufacturing sector.

where if U_i^* is above zero, individuals will respond to the wage question (i.e. $resp_i$ =1). If the utility value is below 0, I cannot find the reported wage. Because the latent variable is an assumed variable, it is not easy to understand, and a figure may be helpful here. In Figure 3a, the latent symbols represent the unobserved U_i^*. The observed symbols are the *resp*$_i$. For every observed *resp*$_i$, it is determined by the latent value U_i^*.

$\mathbf{Z}_\mathbf{i}'$ is a vector of variables that affect the utility of response (U_i^*), and it includes all explanatory variables found in equation (1). As to avoid identification by functional form, researchers should include other variables in $\mathbf{Z}_\mathbf{i}'$ that not in $\mathbf{X}_\mathbf{i}'$.[15] For this study, these include three dummy variables: the proxy response (i.e. reported by family members), the tax month (i.e. interviewed on February or March), and the 4th interview month (i.e. first time to answer income questions). Bollinger & Hirsch (2010) just use the same first two variables (i.e. not including the 4th interview month) in their selection model.

5.2.2 Outcome Equations

The second part of the selection model is the outcome equations.

$$\text{Outcome Part:} \begin{cases} lnwage_i^* \ = \mathbf{X}_i'\boldsymbol{\beta} \ + \varepsilon_i^* & & \text{(4a)} \\ lnwage_i \ = lnwage_i^* & if\ \mathrm{resp_i} = 1 & \text{(4b)} \\ lnwage_i \ = not\ observed & if\ \mathrm{resp_i} = 0 & (4c) \end{cases} \qquad (4)$$

where (4a) is another latent equation, and $lnwage_i^*$ is the log wage value, which is not always observable in the dataset. I only observe the wage when the interviewees choose to answer the wage questions. In Figure 3b, I can see the unobserved wages is not always below

[15] If there is no additional variables, it means that one is identifying the effect due to non-linearity of inverse Mills ratio.

or above the observed wages, so Tobit model does not work here, and only selection model is suitable. In addition, $\mathbf{X}_\mathbf{i}'$ is the same vector as in equation (2).

Equation (3) and (4) together is the basic structure of a selection model, and Figure 3 gives a whole picture of the model. In the second part, it is clear to see that the OLS prediction line using only reported wage is different from the $\mathrm{E}(lnwage_i^*)$, and it means there is a selectivity bias due to non-response. Selection model can be used to find a better prediction of the full model.

5.2.3 Joint Distribution Assumptions

To complete the selection model, the error term ε_i and v_i should be correlated, and jointly normal distributed.

$$\text{Assumptions: } \begin{cases} \varepsilon_\mathrm{i} \sim \mathrm{N}\ (0, \sigma_\varepsilon^2) \\ v_i \sim \mathrm{N}\ (0, 1) \\ \mathrm{corr}\ (\varepsilon_i\ , v_i) = \rho \end{cases} \qquad (5)$$

where the variance of the error term v_i is unitary as it is not identified in the estimation (i.e. only the sign matters). The two error terms will have a correlation ρ. The assumption of the distributions of the two errors terms will highly affect the final results. Equations (3), (4), and (5) together are a full selection model.

5.2.4 Omitted Variable Equation

By taking the conditional expectation values of $lnwage_i$, I can put the whole selection model in one simple equation, and the details are presented in Appendix 3. The new equation presents the selectivity bias as an omitted variable problem, and the new equation is

$$lnwage_i \; = \; \mathbf{X}_i'\boldsymbol{\beta}^* \; + \beta_\lambda \hat{\lambda}_i(\mathbf{Z}_\mathbf{i}'\boldsymbol{\gamma}^*) + \; \varepsilon_i^{**} \qquad (6)$$

where $\hat{\lambda}_i(\mathbf{Z}_\mathbf{i}'\boldsymbol{\gamma}^*) = \frac{\phi\,(\mathbf{Z}_\mathbf{i}'\gamma^*)}{\Phi\,(\mathbf{Z}_\mathbf{i}'\gamma^*)}$, which is the inverse Mills ratio. ϕ (.) is the probability density function (pdf), and Φ (.) is the cumulative normal distribution function (cdf). The new coefficient β_λ (β_λ =$\rho\sigma_\varepsilon$)in the model can be treated as the selectivity bias.

5.3 Solutions of Selection Model

For a selection model, I provide two solutions. The maximum likelihood approach is easy to understand in theory, but difficult to follow in practice. The two-step approach is a easier method to solve the selection model in practice.

5.3.1 Maximum Likelihood Estimation

According to equations (3) and (4), the basic maximum likelihood function for the selection model is very straightforward. It takes the form:

$$\mathrm{L}{=}\prod_{i=1}^{n}\{\,\mathrm{Pr}(U_i^* \leq 0)\}^{\,1-resp_i}\,\{\mathrm{f}(lnwage_i|U_i^* > 0) * \mathrm{Pr}(U_i^* > 0)\}^{\,resp_i} \qquad (7)$$

where the first part is for the non-respondents, and the second part is for the respondents.

5.3.2 Two Step Estimation

Alternatively, Heckman (1979) suggests a two step solution for the selection model. The first step uses a probit model and the full set of observation to get a consistent estimator of $\boldsymbol{\gamma}^*$ from equation (3a). Thus, the inverse Mills ratio ($\hat{\lambda}_i(\mathbf{Z}_\mathbf{i}'\boldsymbol{\gamma}^*)$) also can be obtained. The second step is to use the result from step one, and put it into equation (6) to get the final estimates.

Greene (2012) and Cameron & Trivedi (2005) argue that two step estimates are considered to be more robust than the ML estimators because two step method only requires normality of the first step equation rather than the bivariate normality. In practice, the two-step model seems more popular because it is easier to understand and to find the relationship between the selection equation and outcome equations. I will use the two step model as the true model to compare with the other models.

5.4 Predicted Wages

I will also compare the predicted wage rates from different models. Cameron and Trivedi (2009) suggest comparing the predictions of the different models as another means of evaluation their suitability. Therefore, I will compare the existing mean values and the predicted mean values from each model after the regression.[16] All the models are described, and the results will be discussed in the next section.

6 Empirical Results

6.1 Results from OLS Models

Columns (1) and (2) of Table 3 present the OLS results of estimating equation (1). All the results are statistically significant, which is not surprising given that I am dealing with very large samples. As a result, I will focus on economic significance. Generally speaking, the estimated coefficients difference between the response model (dropping the imputation) in column (1) and the full model (with imputed values) in column (2) are not very large (e.g. lower than 0.01), and all the coefficients have the same expected signs.

[16] Note that when the dependent variable is in the log-normal form, the conditional means for the models are not simply *exp [lnwage]*. See the note of Table 6.

The largest difference occurs for the union status variable, a difference of 0.05. Union status is not a hot imputation criterion to find the donor, so it might be a source of the difference. In addition, the standard errors in the column (1) are generally larger than in the column (2), which implies that with the imputed wage, the estimated coefficients are more significant in the wage equation. However, without imputation, the response model has a selectivity bias. With the imputation wages, the model has a imputation bias. Since both of the OLS models have bias, I cannot give a conclusion here. I have to compare the OLS results with the selection models.

6.2 Results from Selection Models

Columns (3) and (4) of Table 3 present the results from the selection model. Instead of the coefficients, I present the marginal effects (as commonly done in the literature). For the maximum likelihood and two step models, their marginal effects not only have the same signs, but also have very similar estimates. Even the standard errors are very close to each others.

The main difference is for the inverse Mills ratio, and it is because the maximum likelihood and two step have different theoretical structures. The ML estimated coefficient for the inverse Mills ratio is -0.326, and is -0.173 for the two step. Both of them are significant, and it means a negative selection bias for wage non-response. It means that if I ignore the wage non-response model in the sample, I miss an important variable in the wage equation, which is negatively correlated with the wage values. The negative sign of the selection bias result is the same as Bollinger & Hirsch (2010), but the absolute value is larger. With a higher absolute value and a smaller standard error, the ML estimator produces stronger evidence of negative selection bias in the CPS than the two step estimator.

Another difference is the standard errors for the two step estimators are generally larger than for the ML estimators. Cameron & Trivedi (2009) also find larger standard errors in the two step model, and they believe it is because the inverse Mills ratio is collinear with other variables in the outcome equations. Wooldridge (2002) suggests to use exclusive restrictions in the selection model which will reduce the potential collinearity problem, and especially for small samples.

6.3 Comparison

Columns (1) to (3) of Table 5 present the percentage differences of the estimators among different models. I find the OLS model without imputed values (column 1) is close to the selection models, which probably means that the selective bias from wage non-response is relatively small. However, the model with imputed values is far from the selection model, which suggests that the imputation bias is relatively large. These results indicate that dropping the non-response is a better approach. Deleting the wage non-response in the CPS is a better choice than using the wage imputation when someone runs a simple OLS wage equation. The non-response bias is still there, and relatively smaller than the imputation bias.

6.4 Specific Estimated Coefficients

Table 5 shows that the OLS estimates for the coefficient of the variable *non-white* have a large bias. There is a 36.6% difference between the response model and the two step model. This may be because non-white have a larger proportion of non-response (see Table 2), and large non-response rate can lead to a large non-response bias. In addition, there is a 53.7%

difference between full model and the two step model. If the CPS imputation method can correct the non-response bias, the estimated coefficients with imputed values should have been closer to the selection model findings. However, the difference between them is larger, i.e. it is 53.7%. I have to mention that race is a criterion for the CPS hot deck imputation to find the hot donor. It means that the bias from imputed values are larger than the original non-response bias, and imputation cannot solve non-response problem.

Another main finding is that the non-response bias (1.6%) for union status coefficient is much lower than the imputation bias (-29.6%). The union membership between the response and the non-response is not large according to the summary of statistics, so the large bias is not simply from the non-response. The main reason may be the same as the one Hirsch and Schumacher (2004) mentioned, that union status is not a hot deck imputation criterion in the CPS. Lemieux (2006) also points out the negative imputation bias associated with the union status. The difference is that Hirsch and Schumacher (2004) only use a simulation model predict the potential bias from union, and Lemieux (2006) just give a trend of the bias, but here I can provide an actual magnitude.

6.5 Wage Predictions

Columns (1) to (3) of Table 6 present the predicted mean wages from the different models. First, for the OLS model, the wage prediction for the respondent-only is slightly higher than the full model, but the pair of predicted wages are very close. In addition, the two selection models have higher wage predictions than the OLS model. It means that the predicted mean wage is understated in a simple OLS model because it does not include the wage non-response, and the selection models correct some of the negative selectivity bias. In addition, because the absolute coefficient value for the inverse Mills ratio in the ML model is larger than the value in the two step model, the ML model corrects a larger negative non-response bias. Therefore, the predicted wage from the ML selection model is the highest.

7 Robustness Check

As a robustness check, I re-estimated the models restricting my sample to individuals who are paid hourly. Individuals who are paid hourly are asked to report their hourly wage directly, which result in less measurement errors.

7.1 Summary of Statistics

Columns (4) and (5) of Table 2 show summary statistics for the robustness check sample. There are still 98,661 individuals, of which 36,513 (36.93%) are wage non-respondents. In the new sample, the differences between response and non-response are still statistically significant due to the large sample. The characteristics and pattern of the non-respondents in the restricted sample are different from the non-respondents in the full sample.

The means are similar when it comes to gender, age, race, marital status, proxy status, and interview time. The more important differences are for education and occupation. The non-response rate is higher for more educated people (i.e. holding at least a bachelor degree). For those who are working in management positions or the business sector, they also have a lower percentage in the response group. For the union status, there are more than half non-response in this restricted sample.

7.2 Results from Models

Table 4 presents the results for my restricted sample (i.e. only individuals paid hourly). All the estimated coefficients and marginal effects have the same signs as compare to the full sample results (Table 3). However, most absolute values are smaller in the robustness check sample.

Columns (4), (5), and (6) of Table 5 present the percentage differences between different models in Table 4. I find larger percentage differences in the robustness check models. The values in the column (6) are smaller than the column (3). It means that in the restricted sample (the one with less measurement errors), the two selection models results are closer. In addition, the values in the column (5) are smaller than the column (4), so the error from imputation values are still larger than the non-response bias in the robustness check.

For some specific variables, such as the non-white and the union status dummies, I also find the same results. For the non-white, the differences between models are still large because of the high non-response rate. For union status, because it is not in the imputation procedure, the OLS model with imputed values gives a very biased estimate.

The estimated coefficients for the inverse Mills ratio of the ML method and the two step method are very close now. However, columns (3) and (4) of Table 6 show different mean values of predicted wages. The two step model's prediction is closer to the observed sample mean, but the ML model predicts a higher wage.

The robustness check strengthens my finding in the full sample with the smaller, more restricted, and less measurement error sample.

8 Conclusion

In conclusion, wage non-response problem in the CPS is not ignorable, and simply using the imputed values may lead to a larger bias than dropping the imputed values in the wage equation. The widely used hot deck imputation does not have strong conceptual support in the literature, so using imputed wages in the CPS cannot correct the non-response bias, and may lead to a even worse imputation bias. I believe that the CPS and other similar labour economic datasets should provide more detailed imputation procedures and methodologies to the public, and accept different criticisms and suggestions from researchers. As a common problem in the economic research, missing data and non-response is not only the responsibility of few statistical agencies working in the data centres, but also the concern of every data users. If empirical researchers have a more proactive attitude towards non-response and imputation related issues, I believe they will make better use of the data, and draw more reasonable inference.

References

Andridge, R. R. & Little, R. J. A. (2010), "A Review of Hot Deck Imputation for Survey Non-response", *International Statistical Institute* **78**(1): 46-64.

Bethlehem, J., Cobben, F. & Schouten, B. (2011), *Handbook of Nonresponse in Household Survey*, Hoboken, N. J. : John Wiley & Sons.

Bollinger, C. R. & Hirsch, B. T. (2006), "Match Bias from Earnings Imputation in the Current Population Survey: The Case of Imperfect Matching", *Journal of Labor Economics* **24**(3): 689-722.

_____(2010), "Is Earning Nonresponse Ignorable?", Working Paper: Society of Labour Economists (SOLE) Meeting, and NBER Labour Studies Program Meetings.

Bound, J., Brown, C. & Mathiowetz, N. (2001), "Measurement Error in Survey Data", in *Handbook of Economics*, ed. by Heckman, J. and Leamer, E. vol. 5, chap. 59, pp. 3705-3843. Elsevier.

Burbidge, J. B., Magee, L. & Robb, A. L. (2003), "Wages in Canada: SCF, SLID, LFS and the Skill Premium", Hamilton, Ontario: McMaster University.

Cameron, A. C. & Trivedi, P. K. (2005), *Microeconometrics*, Cambridge: Cambridge University Press.

_____ (2009) *Microeconometrics Using Stata*, College Station, Texas: Stata Press.

Card, D., Chetty, R., Feldstein, M. & Saez, Z. (2011), "Expanding Access to Administrative Data for Research in the United States", white papers on *Future Research in the Social, Behavioral & Economic Science*. National Science Foundation.

Card, D., Lemieux, T. & Riddell, W. C. (2004), "Union and Wage Inequality", *Journal of Labor Research* **25**(4): 519-559.

Durrant, G. B. (2009), "Imputation Methods for Handling Item-nonresponse in Practice: Methodological Issues and Recent Debates", *International Journal of Social Research Methodology* **12**(4): 293-304.

Green, D. A. & Milligan, K. (2010), "The Importance of the Long Form Census to Canada", *Canadian Public Policy* **6**(3): 383-388.

Greene, W. H. (2012), *Econometric Analysis*, 7th ed., Prentice Hall: Pearson Education.

Heckman, J. J. (1976), “The Common Structure of Simultaneous 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.

_____(1979), "Sample Selection Bias as a Specification Error", *Econometrica* **47**(1): 153-162.

Hirsch, B. T. & Schumacher, E. J. (2004), "Match Bias in Wage Gap Estimates Due to Earnings Imputation", *Journal of Labor Economics* **22**(3): 689-722.

Huisman, M. (1999), *Item Nonresponse: Occurrence, Causes, and Imputation of Missing Answers to Test Items,* Leiden University, The Netherlands: DSWO Press.

Lee, B. J. & Marsh, L. C. (2000), "Sample Selection Bias Correlation for Missing Response Observations", *Oxford Bulletin of Economics and Statistics* **62**(2): 305-322.

Lemieux, T. (2006), "Increasing Residual Wage Inequality: Composition Effects, Noisy Data, or Rising Demand for Skill?", *American Economic Review* **96**(3): 461-498.

Little, R. J. A. & Rubin, D. B. (2002), *Statistical Analysis with Missing Data*, 2nd ed., New York: John Wiley & Sons.

Mincer, J. (1974), *Schooling, Experience and Earnings*, Columbia University Press: New York.

Tobin, J. (1958), "Estimation of Relationships for Limited Dependent Variables", *Econometrica* **26**(1): 24-36.

Wooldridge, J. M. (2002), *Econometric Analysis of Cross Section and Panel Data*, Cambridge, Massachusetts: MIT Press.

Table 1: Key Variable Description

Variables	**Variable Description**
wage	Hourly wage rate in dollar value
lnwage	Log value of wage
imputation	1 if the hourly wage value is imputed, 0 otherwise
resp	1 if the hourly wage value is reported (not imputed), 0 otherwise
below high school	1 if the highest degree is lower than high school, 0 otherwise
high school	1 if the highest degree is high school diploma or equivalent, 0 otherwise
college	1 if some college or a college diploma, 0 otherwise
bachelor	1 if the highest degree is a bachelor degree, 0 otherwise
graduate	1 if some graduate school or graduate degree, 0 otherwise
experience	working experience in year (age-education year-6)
experience squared	Square value of experience
female	1 if female, 0 otherwise
young	1 if age between 15 and 25 , 0 otherwise
non-white	1 if not white people, 0 if white
married	1 if married, 0 otherwise
manager	1 if management occupation, 0 otherwise
business	1 if business related occupation, 0 otherwise
profession	1 if science, art, legal, medical, or other professional occupation, 0 otherwise
service	1 if low end service sector or sales occupation, 0 otherwise
agriculture	1 if agricultural or fishing occupation, 0 otherwise
manufacturing	1 if manufacturing, transportation, and factory workers occupation, 0 otherwise
public	1 if work in public sector as an employee, 0 otherwise
part-time	1 if part time job, 0 otherwise
union	1 if a union member, 0 otherwise
proxy	1 if the survey is proxy reported, 0 otherwise
tax month	1 if wage question is asked during the tax months (i.e. February or March), 0 otherwise
4^{th} *interview month*	1 if in the 4th month interview in the CPS (i.e. the first time to answer earnings questions), 0 otherwise

Table 2: Summary of Statistics (*Weighted Mean*)

	Main Sample			**Robustness Check**		
Variables	(1) Response	(2) Non Response	(3) % Difference	(4) Response	(5) Non Response	(6) % Difference
A. Gender						
female	0.486	0.478	-1.6%	0.511	0.501	-1.9%
B. Age						
age 15 to 25	0.164	0.147	-10.3%	0.236	0.199	-15.5%
age 26 to 45	0.460	0.422	-8.3%	0.425	0.406	-4.6%
age 46 to 65	0.347	0.391	12.7%	0.310	0.359	15.7%
age 65+	0.029	0.040	38.5%	0.029	0.037	26.6%
C. Education						
below high school	0.098	0.087	-10.4%	0.139	0.115	-17.4%
high school	0.268	0.299	11.4%	0.342	0.357	4.5%
college	0.299	0.296	-0.8%	0.356	0.340	-4.4%
bachelor	0.216	0.212	-1.9%	0.131	0.141	8.0%
graduate school	0.119	0.106	-11.6%	0.033	0.047	42.2%
D. Race						
white	0.826	0.789	-4.5%	0.818	0.777	-4.9%
non-white	0.174	0.211	21.2%	0.182	0.223	22.2%
black	0.101	0.136	34.5%	0.113	0.154	35.6%
D. Occupation						
manager	0.094	0.089	-5.9%	0.032	0.039	22.1%
business	0.044	0.045	1.8%	0.022	0.026	20.0%
profession	0.258	0.243	-5.8%	0.186	0.196	4.9%
service	0.387	0.408	5.6%	0.471	0.468	-0.6%
agriculture	0.008	0.006	-20.0%	0.010	0.008	-24.0%
manufacturing	0.209	0.209	0.0%	0.279	0.264	-5.5%
E. Other Dummies						
married	0.535	0.533	-0.3%	0.460	0.480	4.4%
public sector	0.166	0.159	-4.1%	0.117	0.143	22.2%
part-time	0.279	0.221	-21.0%	0.372	0.280	-24.6%
union	0.120	0.115	-3.8%	0.118	0.124	5.0%
proxy response	0.450	0.566	25.8%	0.467	0.594	27.3%
tax month	0.176	0.145	-17.7%	0.174	0.145	-16.6%
4th interview month	0.491	0.509	3.6%	0.496	0.510	2.9%
Number of Observations	106,593	59,151	___	62,348	36,513	___

NOTE.- Columns (1) and (4) only include the people who directly report their hourly wages.

Table 3: Results from the Main Model

	OLS		**Selection Models** *(marginal effects)*	
Variables	(1) No Imputation	(2) With Imputation	(3) Maximum Likelihood	(4) Two Step
below high school	-0.173 (*.005*)***	-0.196 (*.005*) ***	-0.192 (*.005*) ***	-0.182 (*.006*) ***
college	0.136 (*.004*) ***	0.094 (*.003*) ***	0.128 (*.004*) ***	0.131 (*.004*) ***
bachelor	0.378 (*.004*) ***	0.363 (*.004*) ***	0.376 (*.004*) ***	0.374 (*.005*) ***
graduate	0.575 (*.006*) ***	0.517 (*.005*) ***	0.562 (*.006*) ***	0.565 (*.006*) ***
experience	0.020 (*.005*) ***	0.019 (*.000*) ***	0.022 (*.005*) ***	0.021 (*.005*) ***
experience squared	-0.000 (*.000*) ***	-0.000 (*.000*) ***	-0.000 (*.000*) ***	-0.000 (*.000*) ***
female	-0.162 (*.003*) ***	-0.164 (*.002*) ***	-0.167 (*.003*) ***	-0.164 (*.003*) ***
young	-0.081 (*.006*) ***	-0.099 (*.005*) ***	-0.063 (*.006*) ***	-0.071 (*.006*) ***
non-white	-0.056 (*.004*) ***	-0.063 (*.003*) ***	-0.027 (*.004*) ***	-0.041 (*.004*) ***
married	0.092 (*.003*) ***	0.069 (*.003*) ***	0.089 (*.003*) ***	0.090 (*.003*) ***
manager	0.299 (*.006*) ***	0.291 (*.005*) ***	0.300 (*.006*) ***	0.298 (*.006*) ***
business	0.274 (*.008*) ***	0.272 (*.006*) ***	0.279 (*.008*) ***	0.277 (*.008*) ***
profession	0.156 (*.005*) ***	0.162 (*.004*) ***	0.161 (*.005*) ***	0.158 (*.005*) ***
service	-0.080 (*.004*) ***	-0.080 (*.003*) ***	-0.066 (*.004*) ***	-0.073 (*.004*) ***
agriculture	-0.265 (*.016*) ***	-0.257 (*.013*) ***	-0.269 (*.016*) ***	-0.267 (*.016*) ***
public	-0.037 (*.003*) ***	-0.022 (*.003*) ***	-0.044 (*.003*) ***	-0.040 (*.004*) ***
part-time	-0.144 (*.005*) ***	-0.146 (*.003*) ***	-0.177 (*.005*) ***	-0.162 (*.004*) ***
union	0.189 (*.007*) ***	0.131 (*.004*) ***	0.182 (*.007*) ***	0.186 (*.004*) ***
constant	2.453 (*.007*) ***	2.507 (*.006*) ***	2.614 (*.008*) ***	2.541 (*.011*) ***
inverse Mill's ratio	___	___	-0.326 (*.005*) ***	-0.173 (*.015*) ***
Number of Observations	106,593	165,744	165,744	165,744

NOTE.-Standard errors are in the parentheses. * significant at 10%; ** significant at 5%; *** significant at 1%.

Table 4: Robustness Check

	OLS		**Selection Models** *(marginal effects)*	
Variables	(5) No Imputation	(6) With Imputation	(7) Maximum Likelihood	(8) Two Step
below high school	-0.130 (*.005*) ***	-0.155 (*.004*) ***	-0.143 (*.005*) ***	-0.139 (*.005*) ***
college	0.098 (*.004*) ***	0.069 (*.003*) ***	0.094 (*.004*) ***	0.094 (*.004*) ***
bachelor	0.250 (*.005*) ***	0.244 (*.004*) ***	0.254 (*.004*) ***	0.251 (*.005*) ***
graduate	0.476 (*.009*) ***	0.398 (*.007*) ***	0.507 (*.010*) ***	0.495 (*.009*) ***
experience	0.015 (*.001*) ***	0.015 (*.000*) ***	0.016 (*.001*) ***	0.016 (*.001*) ***
experience squared	-0.000 (*.000*) ***	-0.000 (*.000*) ***	-0.000 (*.000*) ***	-0.000 (*.000*) ***
female	-0.105 (*.003*) ***	-0.108 (*.003*) ***	-0.112 (*.004*) ***	-0.110 (*.004*) ***
young	-0.066 (*.007*) ***	-0.076 (*.005*) ***	-0.052 (*.007*) ***	-0.055 (*.007*) ***
non-white	-0.057 (*.004*) ***	-0.058 (*.003*) ***	-0.035 (*.004*) ***	-0.039 (*.004*) ***
married	0.088 (*.003*) ***	0.067 (*.003*) ***	0.090 (*.004*) ***	0.089 (*.004*) ***
manager	0.120 (*.009*) ***	0.128 (*.007*) ***	0.143 (*.010*) ***	0.137 (*.010*) ***
business	0.144 (*.011*) ***	0.143 (*.009*) ***	0.160 (*.011*) ***	0.157 (*.011*) ***
profession	0.148 (*.005*) ***	0.143 (*.004*) ***	0.159 (*.006*) ***	0.155 (*.005*) ***
service	-0.176 (*.004*) ***	-0.173 (*.003*) ***	-0.163 (*.004*) ***	-0.166 (*.004*) ***
agriculture	-0.240 (*.016*) ***	-0.234 (*.013*) ***	-0.242 (*.017*) ***	-0.241 (*.016*) ***
public	-0.037 (*.005*) ***	-0.035 (*.004*) ***	-0.048 (*.005*) ***	-0.046 (*.005*) ***
part-time	-0.149 (*.003*) ***	-0.156 (*.003*) ***	-0.183 (*.004*) ***	-0.177 (*.004*) ***
union	0.296 (*.005*) ***	0.203 (*.004*) ***	0.291 (*.005*) ***	0.292 (*.005*) ***
constant	2.467 (*.008*) ***	2.506 (*.006*) ***	2.598 (*.009*) ***	2.574 (*.011*) ***
inverse Mills ratio	___	___	-0.259 (*.006*) ***	-0.206 (*.015*) ***
Number of Observations	62,348	98,861	98,861	98,861

NOTE.- Standard errors are in the parentheses. * significant at 10%; ** significant at 5%; *** significant at 1%.

Table 5: Results Percentage Difference between Different Models

Variables	**Full Samples** (% difference with two step)			**Robustness Checks** (% difference with two step)		
	(1) **OLS** **Response**	**(2)** **OLS with** **Imputation**	**(3)** **ML** **Selection**	**(4)** **OLS** **Response**	**(5)** **OLS with** **Imputation**	**(6)** **ML** **Selection**
below high school	-4.9%	7.7%	5.5%	-6.5%	11.5%	2.9%
college	3.8%	-28.2%	-2.3%	4.3%	-26.6%	0.0%
bachelor	1.1%	-2.9%	0.5%	-0.4%	-2.8%	1.2%
graduate school	1.8%	-8.5%	-0.5%	-3.8%	-19.6%	2.4%
experience	-4.8%	-9.5%	4.8%	-6.3%	-6.3%	0.0%
experience squared	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
female	-1.2%	0.0%	1.8%	-4.5%	-1.8%	1.8%
young	14.1%	39.4%	-11.3%	20.0%	38.2%	-5.5%
non-white	36.6%	53.7%	-34.1%	46.2%	48.7%	-10.3%
married	2.2%	-23.3%	-1.1%	-1.1%	-24.7%	1.1%
manager	0.3%	-2.3%	0.7%	-12.4%	-6.6%	4.4%
business	-1.1%	-1.8%	0.7%	-8.3%	-8.9%	1.9%
profession	-1.3%	2.5%	1.9%	-4.5%	-7.7%	2.6%
service	9.6%	9.6%	-9.6%	6.0%	4.2%	-1.8%
agriculture	-0.7%	-3.7%	0.7%	-0.4%	-2.9%	0.4%
public	-7.5%	-45.0%	10.0%	-19.6%	-23.9%	4.3%
part-time	-11.1%	-9.9%	9.3%	-15.8%	-11.9%	3.4%
union	1.6%	-29.6%	-2.2%	1.4%	-30.5%	-0.3%
constant	-3.5%	-1.3%	2.9%	-4.2%	-2.6%	0.9%
Number of Observations	106,593	165,744	165,744	62,348	98,861	98,861

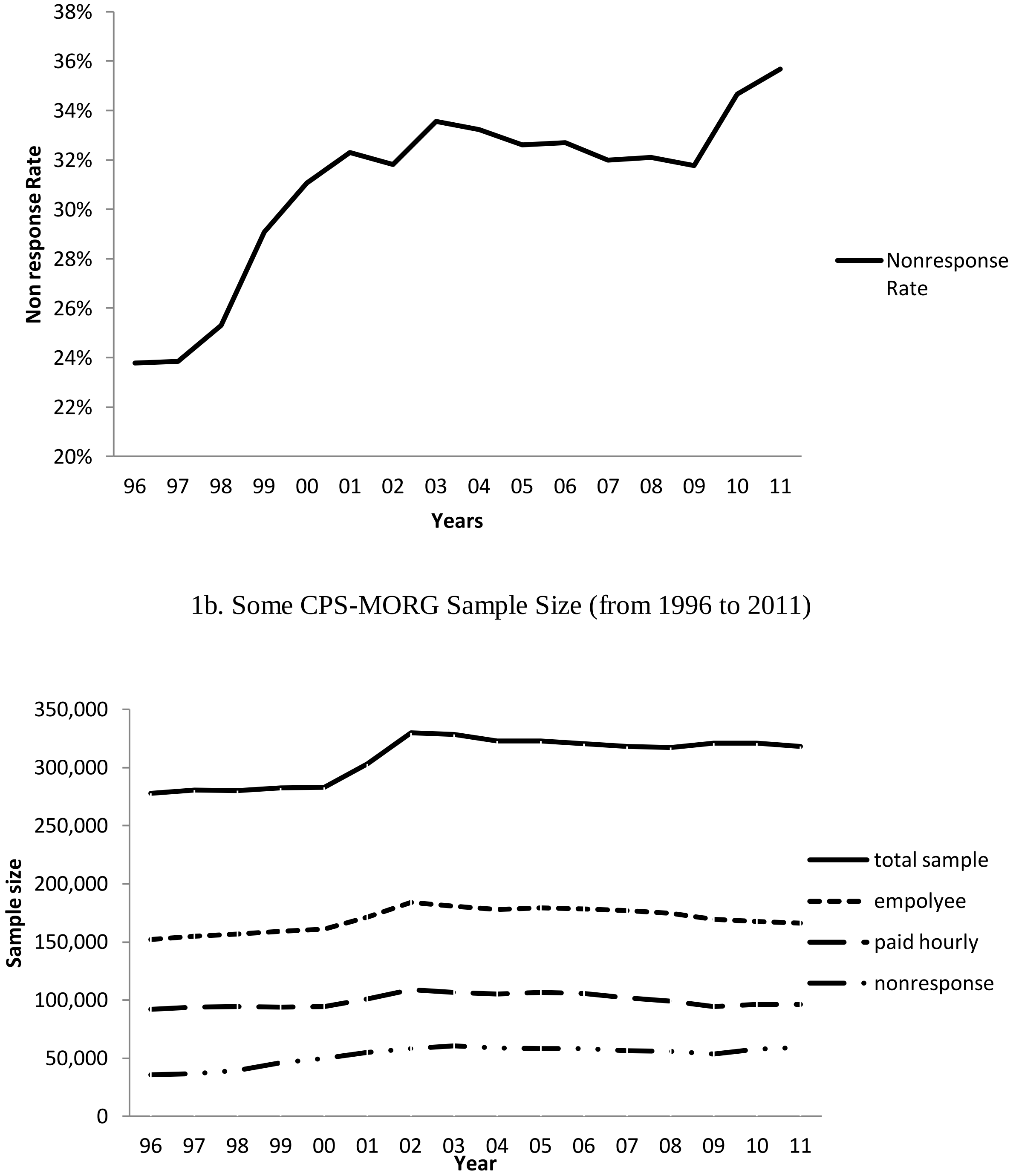
NOTE.- The estimators from the two step selection model are the base values for the percentage differences.

Table 6: Predicted Wages from Different Models

Models	**Original Sample**		**Robustness Check**	
	(1) **Response**	**(2)** **Full Sample**	**(3)** **Response**	**(4)** **Full Sample**
OLS	20.548	20.516	15.365	15.625
ML	22.828	22.790	19.421	19.900
Two Step	21.632	21.604	16.538	16.793
Number of Observations	106,593	165,744	62,348	98,861

NOTE.- When the dependent variable is in the log-normal form, the conditional means for the models is not simply *exp [lnwage]*. Cameron & Trivedi (2009) provide a formula on page 548.

Figure 1: Non-response (Imputation) Rate in the CPS-MORG from 1996 to 2011

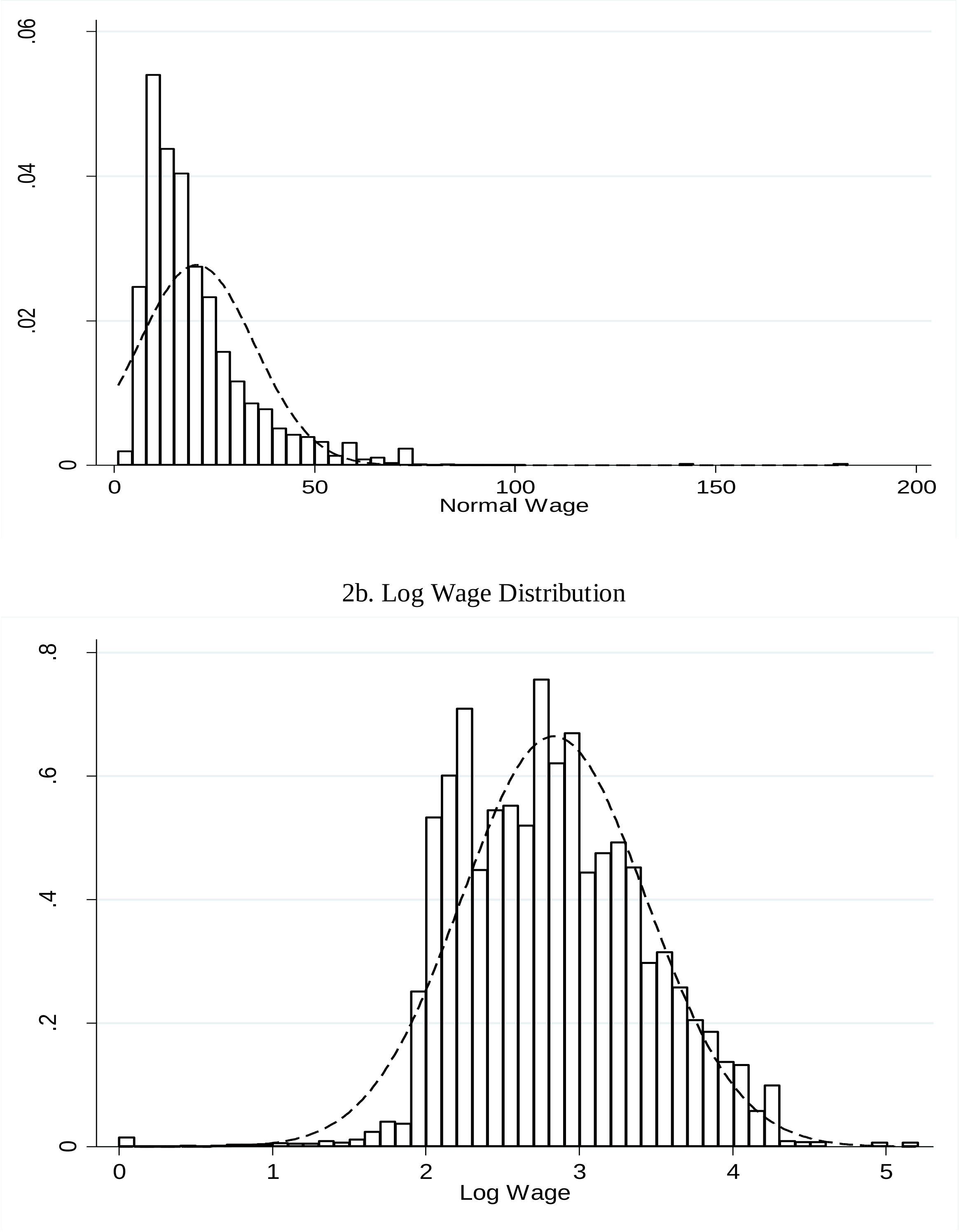


1a. Non-response Rate Changes

1b. Some CPS-MORG Sample Size (from 1996 to 2011)

NOTE.- Before 1996, the CPS does not have consistent allocation flags.

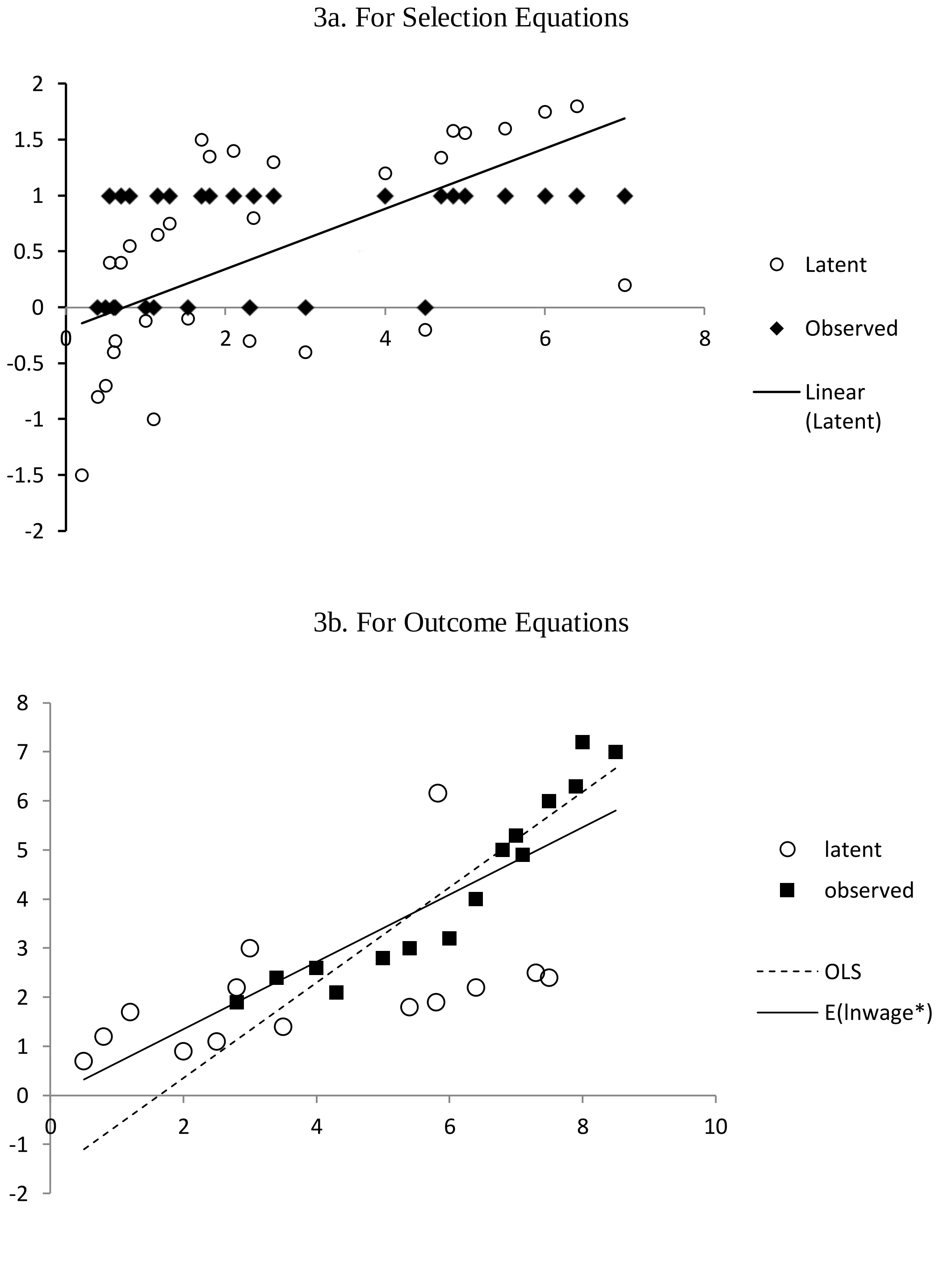
Figure 2: The 2011 CPS Wage Distribution



2a. Normal Wage Distribution

2b. Log Wage Distribution

Figure 3: General Example for a Selection Model

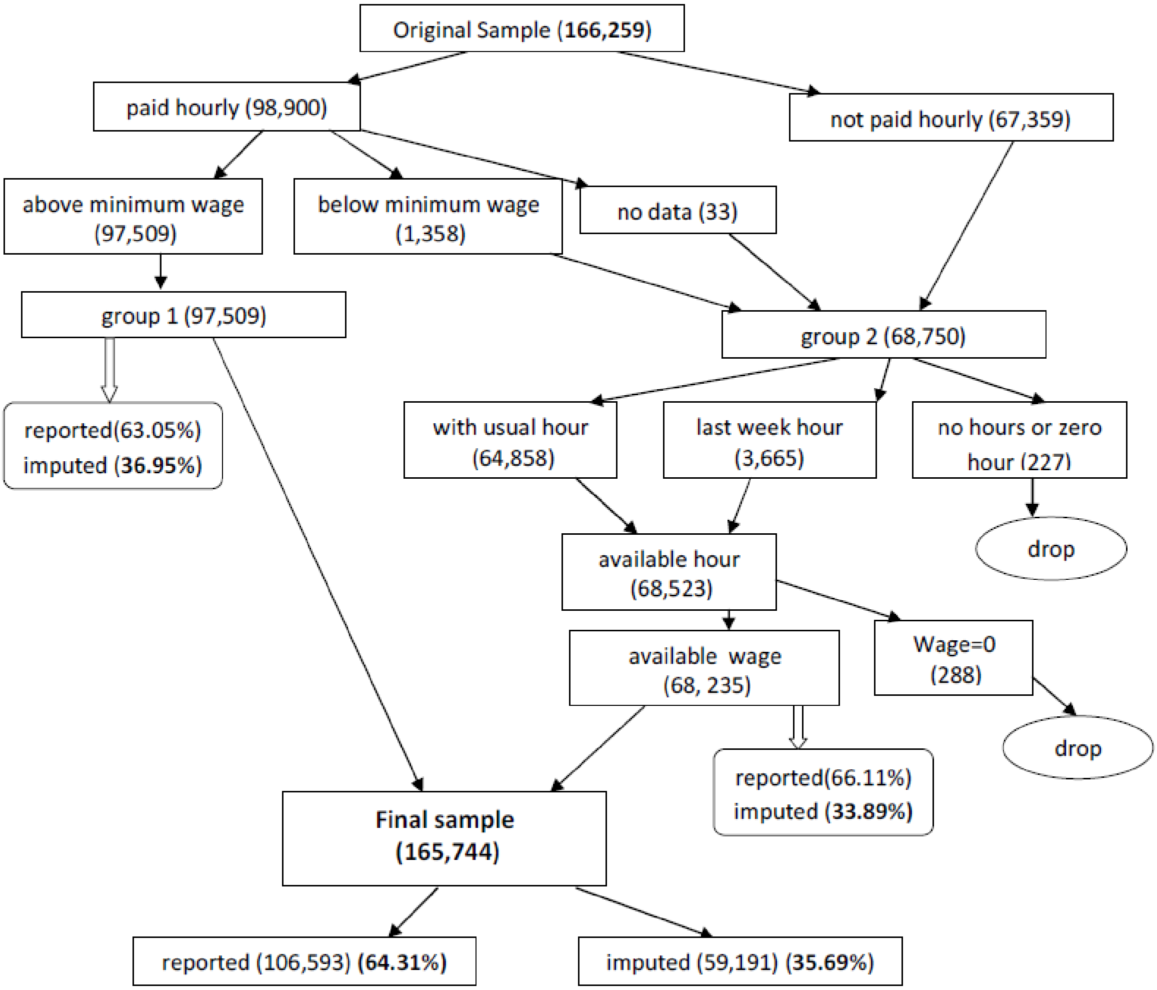


3a. For Selection Equations

3b. For Outcome Equations

Appendix 1: Sample Generating Process for the Imputation Rate

Steps	Change	Total
Original 2011 CPS-MORG	-	318,334
1. Delete unemployment and non-participants	-131,125	187,009
2. Drop self-employed workers	-10,750	166,259


Original Sample (166,259)
paid hourly (98,900)
not paid hourly (67,359)
above minimum wage (97,509)
below minimum wage (1,358)
no data (33)
group 1 (97,509)
group 2 (68,750)
reported(63.05%) imputed (36.95%)
with usual hour (64,858)
last week hour (3,665)
no hours or zero hour (227)
drop
available hour (68,523)
available wage (68, 235)
Wage=0 (288)
drop
reported(66.11%) imputed (33.89%)
Final sample (165,744)
reported (106,593) (64.31%)
imputed (59,191) (35.69%)

Appendix 2: Wage Imputation Flags in the CPS

There are four allocation (imputation) flags in the CPS. The detailed description of the four imputation flags are in the table.

Imputation Flag Name	**Related Variable in the CPS**	**Description**
Item 25a *(I25a)*	uhourse *(usual hour)*	1 if usual working hour is non-response and imputed, 0 otherwise. (for most workers)
Item 25b *(I25b)*	paidhre *(paid by hour dummy)*	1 if paid by hour dummy is non-response and imputed, 0 otherwise. (for every workers)
Item 25c *(I25c)*	earnhre *(earning per hour)*	1 if usual hourly wage is non-response and imputed, 0 otherwise. (only for workers who paid by hour)
Item 25d *(I25d)*	earnwke *(usual weekly earning)*	1 if usual working hour is non-response and imputed, 0 otherwise. (for every workers)

To identify the wage non-respondents in the CPS needs more than one imputation flags. For the people paid hourly, they have hourly wages. I can only use the I25c to find the non-respondents. For the people paid by salary, they only have weekly earnings in the CPS. I have to use I25d and I25a together to find the non-response because working hours also make big difference for the calculated hourly wages.

Appendix 3: Selection Model and Inverse Mills Ratio

The expected log wage ($lnwage_\mathrm{i}$), conditioning on X_i, and Z_i is

$$\begin{aligned} \mathrm{E}\,(lnwage_\mathrm{i} \mid \mathrm{X}_i\,, \mathrm{Z}_i) &= \mathrm{E}\,(lnwage_i^* \mid resp_i{=}1, \mathrm{X}_i, \mathrm{Z}_i\,) \\ &= \mathrm{E}(lnwage_i^* \mid U_i^*{>}0, \mathrm{X}_i, \mathrm{Z}_i) \end{aligned}$$

According to the Moments of Incidentally Truncated Normal Distribution Theorem (see Theorem 19.5 in Greene (2012)). $lnwage_i^*$ and U_i^* have a bivariate normal distribution. Therefore,

$$\begin{aligned} \mathrm{E}(lnwage_i^* \mid U_i^*{>}0, \mathrm{X}_i, \mathrm{Z}_i) &= \mathrm{E}\,(lnwage_i^*) + corr\,(\varepsilon_\mathrm{i}, v_\mathrm{i})\, \sigma_\varepsilon \frac{\phi\,[-\mathbf{E}(U_i^*)/\sigma_v]}{\mathbf{1}-\mathbf{\Phi}\,[-\mathbf{E}(U_i^*)/\sigma_v]} \\ &= \mathbf{X_i'\beta} + \rho\sigma_\varepsilon \frac{\phi\,(-\mathbf{Z_i'\gamma^*})}{\mathbf{1}-\mathbf{\Phi}\,(-\mathbf{Z_i'\gamma^*})} \end{aligned}$$

where ϕ (.) is the probability density function (pdf), and Φ (.) is the cumulative normal distribution function (cdf). σ_ε is the standard deviation of ε_i. ρ is the correlation between the two error terms ε_i and v_i.

In addition, since ϕ (-a) = ϕ (a) for symmetric pdf, and $\Phi\,(-\mathrm{a})$ = 1- $\Phi\,(-\mathrm{a})$ for any cdf

$$\rho\sigma_\varepsilon \frac{\phi\,(-\mathbf{Z_i'\gamma^*})}{\mathbf{1}-\mathbf{\Phi}\,(-\mathbf{Z_i'\gamma^*})} = \rho\sigma_\varepsilon \frac{\phi\,(\mathbf{Z_i'\gamma^*})}{\mathbf{\Phi}\,(\mathbf{Z_i'\gamma^*})} = \rho\sigma_\varepsilon\, \hat{\lambda}_i$$

where $\hat{\lambda}_i = \frac{\phi\,(\mathbf{Z_i'\gamma^*})}{\mathbf{\Phi}\,(\mathbf{Z_i'\gamma^*})}$ (*inverse Mills ratio*). As a result,

$$\begin{aligned} \mathrm{E}\,(lnwage_\mathrm{i} \mid \mathrm{X}_i\,, \mathrm{Z}_i) &= \mathbf{X_i'\beta} + \rho\sigma_\varepsilon\, \hat{\lambda}_i \\ &= \mathbf{X_i'\beta} + \beta_\lambda \hat{\lambda}_i(\mathbf{Z_i'\gamma^*}) \end{aligned}$$

The selection model can now be written in one equation

$$lnwage_i \;=\; \mathbf{X}_i^{'}\boldsymbol{\beta}^* \;+ \beta_\lambda \hat{\lambda}_i(\mathbf{Z_i'\gamma^*}) + \;\varepsilon_{\boldsymbol{i}}^{**} \qquad (6)$$