Problem Set 3: Censoring, Truncation and Selection

Joan Llull
Conghan Zheng

Microeconometrics
Fall 2024

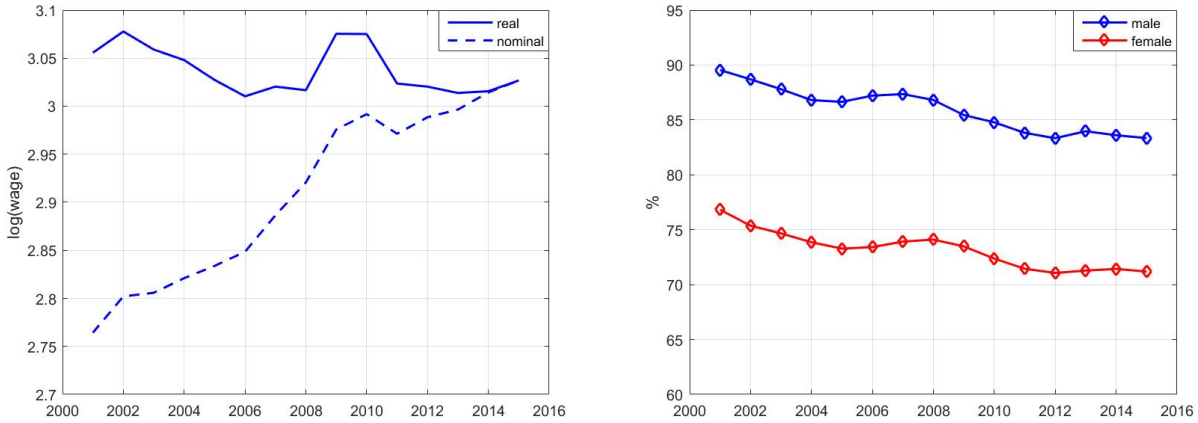
Introduction

This problem set is based on Blundell, Reed and Stoker (2003). Figure 1a below shows that over the years of interest, nominal wages rose steadily, whereas real wages show a (slightly) negative trend. Figure 1b shows a decline in labor force participation for both males and females. We are interested in isolating the potential role of selection in labor force participation, which can cause a upward bias in the trend in Figure 1a.

Using the tools introduced in Chapter 3 and the data that we provide you with, you are supposed to investigate the importance of self-selection in explaining the evolution of real wages over the years 2001-2015 in the U.S. For this problem set, you can omit the sampling weights if including them in the regressions takes too long to run.

Figure 1: Real wages for men and labor force participation, 2001-2015

(a) *Average real hourly wage, in logs* (b) *Labor force participation*



Source: Annual Social and Economic Supplement of the Current Population Survey (US)

Due date October 16th at 12:15 pm, by email to conghan.zheng@uab.cat. You should submit in one zipped folder with:

1. a PDF document containing your analytical solution, if requested in the exercise, any output (figures, tables, etc.) obtained after running the code and your interpretations of the results, instructions on how to run your script(s) if necessary;

2. your code that could run without errors and reproduce the results reported in your explanatory document, specifying the package dependencies if necessary;

3. the name of the zip or rar file, which is `PS3+[your name]`.

Data. The dataset `PS3.dta` is constructed from the *Annual Social and Economic Supplement* (ASEC) of the *Current Population Survey* (CPS), also known as the *March Supplement* of the CPS. It provides annual estimates based on a survey of more than 75,000 households. The ASEC adds to the monthly CPS a set of variables with detailed information on social and economic characteristics, including family and individual income received during the previous year. We restrict our analysis to the period 2001-2015. Our data contains information on weekly and hourly wages, and individual characteristics such as education, age, region, marital status, out-of-work income, and real wages. The sample is limited to men only. Real wages are measured in 2015 dollars.[1] Out-of-work income (variable `benefits`), crucial for the study of labor force participation, is computed using the *OECD Tax-Benefit* model for the US.[2] Earnings information is included only for working individuals, which imply self-selection.

A description of the variables is included in the *Appendix*.

Model. We define a latent variable w^* that represents hourly earnings and is only partially observed. Wages are determined by the linear model

$$w_i^* = x_i'\beta + \varepsilon_i, \tag{1}$$

where the vector x_i contains worker characteristics. We can think of two reasons why wages are only observed partially. First, since wages cannot be negative, the data are left censored with lower limit equal to 0:

$$w_i = \begin{cases} w_i^*, & \text{if } w_i^* > 0 \\ 0, & \text{if } w_i^* \leq 0 \end{cases} \tag{2}$$

Therefore, OLS estimates are not consistent. To estimate (1), we should rather use a Tobit model with density function

$$g(w|\boldsymbol{x}, w > 0) = f(w|\boldsymbol{x})^d F(0|\boldsymbol{x})^{1-d}, \tag{3}$$

where d equals one if $w^* > 0$ and zero otherwise. In particular, we assume that w follows a normal distribution.

[1] Wages are deflated using the CPI index and are expressed in relative prices of 2015.

[2] Benefits comprise both unemployment insurance (UI) and social assistance (SA). They depend on year, marital status, number of children, employment status of the spouse, and earnings. Each year there are 1,206 distinctive groups eligible for UI and SA. Values are in logs. For more information, visit `https://www.oecd.org/social/benefits-and-wages/`.

Second, following the literature on search and matching, we know that unemployed individuals decide to work only if the wage offer they receive, w^*, exceeds their reservation wage $\underline{w}$. The reservation wage is usually attributed to out-of-work benefits such as unemployment insurance. Thus, wages are also affected by self-selection of workers into the labor force:

$$w_i = \begin{cases} w_i^*, & \text{if } w_i^* > \underline{w}_i \\ 0, & \text{otherwise} \end{cases} \tag{4}$$

Additionally, we assume that self-selection is not only governed by the out-of-work benefits b_i but also by worker characteristics z_i:

$$\underline{w}_i = \tilde{\alpha}_1 b_i + \tilde{z}_i' \tilde{\gamma} + \tilde{\nu}_i, \tag{5}$$

where b_i are benefits for individual i, and vector z_i includes a set of worker characteristics, some of which are included in x_i. Therefore, Equation (4) can be rewritten as

$$\begin{aligned} w_i &= \begin{cases} w_i^*, & \text{if } x_i'\beta + \varepsilon_i > \tilde{\alpha}_1 b_i + z_i'\tilde{\gamma} + \tilde{\nu}_i \\ 0, & \text{otherwise} \end{cases} \\ &= \begin{cases} w_i^*, & \text{if } \alpha_1 b_i + z_i'\gamma + x_i'\beta + \nu_i > 0 \\ 0, & \text{otherwise} \end{cases} \end{aligned} \tag{6}$$

where $\alpha_1 = -\tilde{\alpha}_1$; $\gamma = -\tilde{\gamma}$; $\nu_i \equiv \varepsilon_i - \tilde{\nu}_i$; and z_i and x_i may share common regressors (the constant term for sure, and can be merged in this expression).

Exercise 1: Censoring

Ignore the endogenous selection aspect of the data.

1. Estimate Equation (1) using a Tobit model for left-censored data. First, use real hourly wages and then log real hourly wages. Include in x_i the corresponding variables to control for education level, birth cohort, marital status, region, a time trend, and the interaction terms. In particular, add the same combination of regressors as in the wage equation shown in Table 3 of Blundell, Reed and Stoker (2003), but use only a linear time trend. Discuss the results. Interpret the coefficients of *region 2* and its interaction term with the time trend.

2. Compute the marginal effects on the left-censored mean of log wages. Interpret the marginal effect of variable *time*.

3. Estimate a two-part model to explain log real hourly wages. Compare its performance and estimates of the second part with the previous results of the Tobit model.

Exercise 2: Selection

1. Estimate Equation (1) by the Heckman correction method (two-stage approach) on the selection process described in Equation (6). Include in z_i education of the spouse, out-of-work income b_i, and all the regressors in x_i. Discuss why we should be concerned about selection bias in this context.

2. Estimate Equation (1) by OLS on the sample of individuals that work. Compute the average conditional mean of hourly wages

$$\mathbb{E}[\text{hourly wage}|x, \text{hourly wage} > 0]$$

for each year using linear predictions from three models: truncated OLS, the Tobit model, and the selection model by FIML. Plot your results across years and interpret them making reference to the statistics provided in Figure 1b. What is the role of selection empirically?

3. Out-of-work benefits are crucial in this context for identifying participation separately from wages. Given the structure and composition of benefits (in the current U.S. context), comment on whether you think that identification can be compromised in this case.

References

Blundell, Richard, Howard Reed, and Thomas M. Stoker, "Interpreting Aggregate Wage Growth: The Role of Labor Market Participation," *American Economic Review*, 2003, *93* (4), 1114–1131.

Appendix: Description of variables

- `id`: unique person identifier.
- `lfp`: labor force participation, equal to 1 if the individual is in the labor force, and 0 otherwise.
- `region`: U.S. Census region.
- `age`: age of the individual (between 19 and 60).
- `cohort`: birth cohort (b. 1942-1960, b. 1960-1970, b. 1970-1980, b.1980-1996).
- `education`: years of education.
- `education_spouse`: years of education of spouse.
- `educ`: education level (less than high-school graduated, some years in college, college graduated, and further education).
- `educ_sp`: education level of the spouse.
- `married`: marital status.
- `earn_'j'`: earnings $j \in w, h$, where w indicates weekly, and h hourly wage.
- `earn_'j'_r`: real hourly/weekly wage.
- `benefits`: out-of-work income based on the *OECD Tax-Benefit* model, in logs.