Project 2: State Dependence in Union Membership

1 Introduction

A question of economic interest is whether union membership yields a wage premium. That is, do unions raise the wages of their members and, if so, which members? Several authors have attempted to estimate such a *union wage premium* (see, e.g., Vella and Verbeek, 1998). A challenge arising in this setting is that union membership is not randomly assigned. Moreover, whether or not you choose to join a union may hinge on variables privy to the worker.

This exam is about modeling *union membership dynamics.* While this task has little to do with wages as such, it may be thought of a as first step towards estimating a wage premium in that we are aiming at modeling *selection into unions.*

2 Data

You are provided with data `union.csv` on $n = 545$ young men who worked every year of from 1980 through 1987, the initial period $t = 0$ corresponding to 1980. To load the data into Python, import `pandas` and use `read_csv('union.csv')`. The dataset contains the following (basic) variables:

Variable	**Description**
`nr`	Person identifier
`year`	1980 to 1987
`union`	= 1 if in union
`married`	= 1 if married
`educ`	Years of schooling

The dataset has already been sorted ascendingly according to `nr` and then `year`.

3 Assignment

The goal of this assignment is to empirically quantify the level of state dependence in union membership using the data provided. Specifically, you must formulate an econometric model capturing union membership dynamics, estimate the model **using only the first five years of data**, and report the results. Emphasis should be placed on (average) partial effects of interest capturing state dependence in union membership specifically. Your model for union membership should involve a relationship of the form

$$\mathrm{P}\left(union_{it}=1\middle|\, union_{i,t-1},\ldots,union_{i0},\mathbf{z}_i,c_i\right)=G\left(\mathbf{z}_{it}\boldsymbol{\gamma}+\rho\cdot union_{it-1}+c_i\right),$$

$t = 1, 2, \ldots, 7$, where $\mathbf{z}_{it}$ denotes a collection of control variables arising from the table above, $(\boldsymbol{\gamma}', \rho)'$ parameters to be estimated, c_i individual-specific time-invariant unobserved heterogeneity (on which assumptions are to be placed), and $G : \mathbf{R} \to \mathbf{R}$ a link function of your choosing. Your estimation results may come in the form of tables and/or figures.

4 Hints and Suggestions

(1) Remember to properly define all variables and symbols employed and distinguish between them. For example, you should distinguish between the true parameter and an estimate thereof as well as between random variables and values at which they are conditioned on. Strive to employ the notation used in the course/textbook. Make use of boldface and capitalization to avoid confusing scalars, vectors and matrices. Specify dimensions whenever confusion may arise.

(2) When using an estimation procedure, carefully discuss the assumptions required to derive the estimator and establish properties thereof. Assess whether these assumptions are likely to be satisfied in the current empirical setting. (Don't just copy the math.) If not, what are the consequences for the estimator in question (and your results)? Strive to provide a real-world example of behavior that might invalidate a given assumption, carefully linking the behavior or mechanisms to the mathematical symbols in the model.

(3) If you come up with several model specifications and associated estimates, discuss which one seems the most appropriate and justify your decision (e.g., based on formal testing).

(4) Be precise about the statistical tests you use for testing various hypotheses. Explain which null hypothesis you are testing and the alternative you are testing against, how the test statistic is constructed, the decision rule you employ, and the conclusion you reach. If a variance (matrix) has been estimated, discuss the assumptions invoked for consistency. If several choices are possible, justify your choice.

5 Formal Requirements

- You must hand in a report that presents the econometric model, presents your estimation results and results of formal statistical tests (including interpretation and statements on economic and statistical significance), and discusses the potential weaknesses of the model, data and approach. If you present many estimates of the same parameters (e.g. estimators based on different asumptions, or varying the controls or sub-sample used), it may be helpful to present the estimates together in one table to facilitate comparison.
- The report must be written in English using an academic language and uploaded to Peergrade via Absalon as a single PDF file.
- The report must be at most five pages of main text (including mathematics) plus at most two pages of output.
    - For the main text (and mathematics), you must use fontsize = 12p, line spacing = 1.5, and 2.5 cm page margins (as used in this document).
    - The ouput can be any (relevant) tables or graphs as long as they properly formatted and labelled. Place the output at the end of your report, starting on a new page.
- Along with your report, you must upload a compressed zip-folder with all the Python code needed to replicate your results. Make sure that your code is transparent and runs with only minor modifications (e.g. changing relevant paths). There is no character limit on the submitted Python code.
- You are allowed (and strongly encouraged) to work in groups of up to three people. List all group members on the front page of your report in alphabetical order of surnames.
- The assessment criteria are given on the course website in Absalon.

References

Vella, F. and M. Verbeek (1998): "Whose wages do unions raise? A dynamic model of unionism and wage rate determination for young men," *Journal of Applied Econometrics*, 13, 163–183.