Microeconometrics: Part I

-Tilburg University course 35M1C5-

Instructor: Dr. Alexandros Theloudis

Autumn 2022

Course assignment 2

The students should work on this assignment after Lecture 10 'Structural econometrics III'. Please provide brief but complete answers. Hand in your answers by **Monday October 24, 2022** at noon.

1. **(survival analysis; 50 points)** In this question, you will examine how mortality rates of elderly Americans depend on socioeconomic variables, parental longevity, health behavior, and chronic health conditions. The question is based on data from the Health and Retirement Study (HRS), a representative panel that started in 1992 and follows cohorts of (near) elderly Americans through their lives; the data contain an astounding amount of information about them.

   Answer the questions below using `Stata`. Add a do-file to your submission with all your commands from data preparation to final output. There are three pieces of `Stata` code on Canvas; the `Stata` survival analysis manual is an excellent source of more detailed information. Please comment your code adequately.

   You must access two datasets: 1) the RAND HRS longitudinal file 2018; 2) a cross-wave tracker 2014 file. Download the first file yourself from www.rand.org/labor/aging/dataprod/hrs-data.html; register, wait for your password, and download the file. The download includes PDF documents with descriptions of the data. The second file is available on Canvas (`tracker2014.dta`).

   (a) Import each data file into `Stata`. Select or generate the following variables; carry out sample selection:

   - **tracker file**: keep information on `HHID`, `PN`, `BIRTHYR`, `DEGREE`, `GENDER`, `HISPANIC`, `RACE`, NYEAR, `AIWWAVE`. Keep observations if they participate in 1992 interview (`AIWWAVE==1`). Create variable `hhidpn = HHID * 1000 + PN` (generate variable in long format); [2.5/50]

- **RAND HRS data**: keep variables `hhidpn`, `r1iwstat`, `r1momliv`, `r1dadliv`, `r1momage`, `r1dadage`, `r1smokev`, `r1smoken`, `r1drinkr`, `r1bmi`, `r1hibpe`, `r1diabe`, `r1cancre`, `r1lunge`, `r1hearte`, `r1arthre`. You can keep additional variables if they look interesting to you. Keep observations if they participate in the 1992 interview (`r1iwstat==1`). [2.5/50]

(b) Merge the two datasets using `hhidpn` as key. Check whether all observations are merged. Merging is a common source of mistakes in empirical work. If `hhidpn` is not in long format, the merge will include mistakes. Hint: 2 observations will not be matched; drop them. [7.5/50]

(c) Set the data in survival analysis format. The origin is the year of birth, `BIRTHYR`. The time under observation starts in 1992. The end date is the year of decease, `NYEAR`. Not all subjects in the data die. If `NYEAR==0`, the person survives beyond the year 2011; these observations are thus right-censored in 2011. [7.5/50]

(d) Show Kaplan Meier survivor estimators (graphs) separately for men and women, starting at age 55. Please describe your results. [5/50]

(e) Run Weibull proportional hazards models, successively adding more covariates. Many of the covariates first need to be created based on the information in the dataset. You should take care creating these covariates; if you use e.g. commands like `gen black = RACE == 2`, observations with missing race will be mistakenly categorized as non-black.

i. use binary covariates for the following socio-economic characteristics: female, black, hispanic, high-school degree (high-school degree or equivalent, but less than a 2-year college degree), and college degree (2-year college degree or more); [1/50]

ii. in addition to the covariates in i., add variables on longevity of parents: whether the father died before age 65 (excluded dummy), and whether the mother died before age 70 (excluded dummy); [1/50]

iii. in addition to the covariates in ii., add variables on health behavior: whether one is current smoker, past smoker, heavy drinker ($\geq$ 3 drinks per day), moderate drinker (0 <drinks per day< 3), overweight (25 $\leq$ BMI $\leq$ 30), and obese (BMI $>$ 30); [1/50]

iv. in addition to the covariates in iii., add covariates on chronic health conditions: high blood pressure, diabetes, cancer, heart disease, lung disease, and arthritis. [1/50]

Describe and discuss the estimation results in the specifications above. [4/50]

Show the results for all specifications in one table with several columns; the `estout` and `esttab` commands (or alternatively `outreg2`) may be useful for this – but

alternative ways are possible. If these commands are not already installed on your computer, you may have to install them e.g. through `ssc install estout`. [3/50]

(f) Write down and describe the hazard function for the model you estimate in specification (e)i. Also write down and describe the likelihood function for this specification. Use vector notation for the covariates. [9/50]

(g) Describe and explain the assumptions that the proportional hazards model in point (e) is based on, i.e. with respect to the shape of the baseline hazard, the proportionality condition, and the type of heterogeneity allowed for in the data. [5/50]

2. **(GMM; 50 points)** Consider the linear regression model

$$y_i = \underset{1\times q}{\mathbf{x}_i'} \underset{q\times 1}{\boldsymbol{\beta}} + u_i$$

(a) Suppose the regressors are exogenous, that is, $\mathbb{E}[u_i|\mathbf{x}_i] = 0$ holds. What are the implied orthogonality conditions in the model? And what are their sample analog? [5/50]

(b) Write down explicitly the objective function of GMM. [5/50]

(c) Does the GMM estimator have a closed form solution? What is it? [10/50]

(d) Suppose that the regressors are not exogenous but there are relevant instruments $\underset{r\times 1}{\mathbf{z}_i}$ for which $\mathbb{E}[u_i|\mathbf{z}_i] = 0$. Answer all previous questions. [10/50]

(e) Show and justify what conditions q and r must satisfy for identification in (d). [10/50]

(f) Explain which weighting matrix yields the smallest possible variance in (d), assuming the conditions of (e) are met. [10/50]

End of assignment 2