Problem Set 4: Durations Models

Joan Llull
Conghan Zheng

Microeconometrics
Fall 2024

Introduction

This problem set aims at replicating some parts of Farber (2005) using the tools introduced in Chapter 4.

Due date November 1st 23:59, 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 being `PS4+[your name]`.

Data. `PS4_1.dta` and `PS4_2.dta` contain the data used in the original Farber (2005) paper. Data were collected from the taxi leasing companies in New York and was constructed based on trip sheets filled by cabdrivers. The datasets contain information on 12,187 fares during 538 shifts for 15 drivers, and differ slightly from the dataset used in the paper[1]. The main difference between the two files is that `PS4_1.dta` contains records for each trip, whereas `PS4_2.dta` summarizes the information for each shift. Variables are described in the Appendix.

Exercise

All the tables and equations referenced in the following questions can be found in the Farber (2005) paper.

1. Plot the empirical survival function with the duration defined as hours worked in the shift. This is a graph with duration on the horizontal axis, and the fraction of the original sample still working at each period on the vertical axis. Round hours worked to one decimal point.

[1] Farber (2005) analyzes 13,461 trips in 584 shifts for 21 drivers.

2. Construct a dummy variable for high-wage shift (hourly wage above the median). Estimate its proportional impact on working hours and plot the hazard function for both high-wage and low-wage shifts. Discuss the results.

3. Consider two sets of regressors: X_1 consists of `high_wage, rainfall`, and `snowfall`; X_2 contains `high_wage, rainfall, snowfall`, and individual dummies. Estimate Cox regressions of these models and test the proportional hazard assumption. Compare the results.

4. Conduct a test for unobserved heterogeneity of drivers using a Cox regression on X_1. Plot the survival functions of the most "diligent" and the "laziest" drivers. Relate the results to your previous findings.

5. Use the trip data (`PS4_1.dta`) to compute the empirical hazard of stopping by hours and by income as in Table 4 of Farber's paper.[2]

6. If you were to estimate a logit discrete time duration model where the equivalent to the baseline hazard γ_t was specified as a sum of free parameters of time, what cutoffs would you use to parameterize γ_t? Explain whether and why you would censor your sample to estimate the model, and around which duration you would do so.

References

Farber, Henry S., "Is Tomorrow Another Day? The Labor Supply of New York City Cabdrivers," *Journal of Political Economy*, 2005, *113* (1), 46–82.

[2] *Hint*: your results might not be exactly the same as those in the original paper but they should not be too different.

Appendix: Description of Variables

List of variables in `PS4_1.dta`:

- `id`: unique driver identification number.
- `shid`: unique shift id (= `id` $*$ 100000 + `date`).
- `trip`: trip number within each shift.
- `date`: date of the shift.
- `stop`: indicator for last trip of the shift.
- `tot_hrs`: total hours at the end of trip.
- `tot_inc`: total income at the end of trip.
- `hour`: clock hour.
- `dayend`: day shift.
- `raining`: if hourly rain > 0.
- `snowing`: if daily snow > 0.
- `nonmanh`: 1 if outside Manhattan.

List of variables in `PS4_2.dta` that are not included in `PS4_1.dta`:

- `hours`: hours worked.
- `lhours`: log hours worked.
- `wage`: hourly wage.
- `lwage`: log hourly wage.
- `income`: total income per shift.
- `ttripmin`: total time of trips.
- `tbrmin`: total time of breaks.
- `twaitmin`: total waiting time.
- `tripmin`: average trip time.
- `brmin`: average break time.
- `waitmin`: average waiting time.
- `mintemp`: minimum temperature.
- `maxtemp`: maximum temperature.
- `rainfall`: amount of hourly rainfall.
- `snowfall`: amount of daily snowfall.