Replication readme: "Gender Earnings Gap in the Gig Economy"

August 2020

Introduction

This document describes the code and data used in the analyses and included in the replication files. The primary data was provided under a Data Usage Agreement from Uber. This data is proprietary and often contains Personally Identifying Information; our agreement with Uber excludes us from sharing the underlying data.

Computational requirements

Coding environment

The primary analyses we written in Python 2.7.15, with a few supplemental analyses—such as the Gelbach decompositions—run in Stata.

The Python code is provided as a set of Jupyter Notebooks. The file `requirements.txt` includes all packages (with versions) required to run the code. To ease browsing of the code, the folder `html_notebooks` contains HTML versions of each Jupyter Notebook, which can be opened (but not run) in any standard browser without installing any software.

The Stata code was run on Stata MP 16. The only non-standard package is `b1x2`, for running Gelbach decompositions – more information on how to install and use this package is available in Gelbach (2014).

Memory and runtime requirements

The primary analyses were run on a 32-core server with 256GB of memory. Start to finish, the code takes approximately 36 hours to run and often approaches the memory limits of the server.

Description of Code

These files do not need to be run in order, but are presented approximately in the order of which the results appear in the paper.

`1_weekly_analyses_for_entire_US.ipynb`

- This notebook contains all analyses on the weekly Uber data (i.e. most results in section 3 of the paper). This includes summary statistics, weekly gender earnings gap, and the hourly earnings gap over time.
- Outputs: Figure 1, Table 1, Table 2

`2a_chicago_summary_stats.ipynb`

- This notebook uses driver-hour and trip-level data from Chicago to construct statistics including parameter averages and graphs related to earnings, tenure, and learning.
- Outputs: Figure 2, Figure 4, Figure 5, Figure 6, Figure 15, Table 3, Table 15

`2b_chicago_hourly_main_analyses.ipynb`

- This notebook uses driver-hour data from Chicago to and outputs all the main results based on Chicago hourly data, including regressions for where, when, experience, and speed as well as a number of Appendix figures and tables.
- Outputs: Figure 7, Figure 16, Table 4, Table 5, Table 6, Table 7, Table 17, Table 18, Table 19, Table 20

`2c_chicago_geographies.ipynb`

- This notebook takes geohash-level aggregates of Uber driving and geohash features and analyses the features of locations more commonly driven in by men.
- Outputs: Figure 3, Table 13, Table 14

`3_other_cities.ipynb`

- Here we extend our baseline analyses to other cities, using driver-hour level data from each city.
- Outputs: Table 9

`4_gelbach_decompositions.do`

- This Stata script runs all the Gelbach decompositions seen throughout the paper.

- Outputs: Figure 9, Figure 10, part of Table 9

`A1_costs.ipynb`

- This notebook uses data on Uber driver vehicles and their fuel efficiency to test whether men or women pay higher gas costs.
- Outputs: In-line results of A.1

`A2_school_and_football.ipynb`

- This notebook recreates the school and football game results found in Appendix section 2.
- Outputs: Figure 9, Table 10

`A3_taxi_drivers.ipynb`

- This recreates the results relating to traditional taxi drivers, using data from the CPS.
- Outputs: Table 11

`A4_race.ipynb`

- This notebook merges in information on driver race and tests whether controlling for race affects the earnings gap.
- Outputs: Table 15, Table 16

`A5_incentives.ipynb`

- This notebook duplicates the main findings on a measure of earnings that excludes incentives.
- Outputs: Table 12

`A6_learning_curve_analysis.do`

- This do file further explores the learning curve for Uber drivers, as described in Appendix A.9 and A.10.
- Outputs: Figure 11, Figure 12, Figure 13, Figure 14

`A7_speed_NHTS.ipynb`

- This do file uses data from the NHTS on driving speed of men and women outside of Uber.
- Outputs: Table 21

`analyses_helper_functions.py`

- This script contains a few functions that are used frequently throughout many Python analyses, especially for graphing.

Data Sources

Data Availability Statements

The primary data for this project are confidential and were provided by Uber under a Data Use Agreement (DUA) that prohibits making the data publicly available. Researchers interested in working with Uber data should contact Libby Mishkin (mishkin@uber.com), a Senior Economist at Uber, or Jonathan Hall (jvh@uber.com), the Chief Economist.

In addition, for this project we used data from the National Highway Travel Survey (Federal Highway Administration, U.S. Department of Transportation, Washington, DC., 2017) and Current Population Survey (Flood et al., 2018). Each of these are publicly available. Further details are provided below.

List of datasets

`us_weekly_earnings` and `chi_weekly_earnings`

- These datasets contain weekly aggregations of a given driver's earnings and trip characteristics. The Chicago file is a subset of the full US file.

`chi_combined_data` and `chi_reg_subset_w_geo_data`

- These are the primary datasets for the Chicago hourly analyses. The first includes all driver-hours for Chicago drivers. The second limits to just those used in regressions and includes certain geographic location characteristics (e.g., crime levels in pickup location).

`atl_combined_data`, `bos_combined_data`, `det_combined_data`, `hou_combined_data`, and `sf_combined_data`

- Each of these dataset contain the same driver-hour level features as the Chicago version, but for Atlanta, Boston, Detroit, Houston, and San Francisco.

`chi_trip_level_052016_052018`

- This dataset contains individual trips for Chicago drivers, including characteristics such as pay, speed, and wait time.

`geohash_level_aggregates`

- This dataset is at the geohash-level and include features such as number of trips starting there as well as demographics and crime levels.

`chi_gender_proj_ethnicity`

- This dataset includes estimates of each Chicago driver's race and ethnicity.

`driver_vehicles` and `epa_mpg_data`

- These data include the make and model of each driver's most frequently used car as well as an estimate of the MPG.

`cps`

- These data are the 2004-2018 Current Population Survey (CPS) with the Annual Social and Economic (ASEC) supplement for taxi and limousine drivers (Flood et al., 2018).

`nhts_tripfile`

- These data include trips in the National Household Travel Survey, which are used to estimate gender differences in speed outside of Uber driving (Federal Highway Administration, U.S. Department of Transportation, Washington, DC., 2017).

References

Federal Highway Administration, U.S. Department of Transportation, Washington, DC., "National Household Travel Survey," 2017.

Flood, Sarah, Miriam King, Renae Rodgers, Steven Ruggles, and J. Robert Warren., "Integrated Public Use Microdata Series, Current Population Survey," 2018.

Gelbach, Jonah, "B1X2: Stata module to account for changes when X2 is added to a base model with X1," 2014.