Data Availability Guidelines and Code Base for "Penalties for Speeding and their Effect on Moving Violations: Evidence from Quebec Drivers"

Vincent Chandler
Université du Québec en Outaouais

Lealand Morin
University of Central Florida

Jeffrey Penney
University of Alberta

February 3, 2022

Penalties_for_Speeding_in_Quebec

This is the documentation for the code base to accompany the manuscript *Penalties for Speeding and their Effect on Moving Violations: Evidence from Quebec Drivers* by Chandler, Morin, and Penney in the *Canadian Journal of Economics*, 2022

Any updates will be available on the GitHub code repository `Penalties_for_Speeding_in_Quebec` available at the following link:

`https://github.com/LeeMorinUCF/Penalties_for_Speeding_in_Quebec`

Data Availability

All data were obtained from the Société de l'assurance automobile du Québec (SAAQ), the driver's licence and insurance agency for the province of Québec. The administrators responsible for the data can be reached at `https://saaq.gouv.qc.ca/en/reach-us` or at the following mailing address:

Service de la recherche en sécurité routière
Société de l'assurance automobile du Québec
333, boulevard Jean-Lesage, C-4-12
C. P. 19600, succursale Terminus
Québec (Québec) G1K 8J6
Téléphone : 418 528-4095

Once the data are obtained, these primary datasets must be placed in the `Data` folder before running the scripts.

Traffic Tickets

The primary data source is an anonymized record of traffic tickets from the SAAQ for each year in the sample. The data were provided to the authors under an understanding that the data not be made publicly available.

The official name of the database is *Fichiers des infractions au Code de la sécurité routière*. The datasets are named in the format `csYYYY.dta`, with `YYYY` indicating the year in which drivers received tickets. The datasets contain the following variables.

- `pddobt` or *Nombre de points* is the number of demerit points awarded for the offence.
- `dinf` or *Date d'infraction* is the date of infraction in `YYYY-MM-DD` format.
- `dcon` or *Date de condamnation* is the date of conviction in `YYYY-MM-DD` format.
- `seq` or *Numéro séquentiel* is a sequence of unique identification numbers for the drivers.

Statistics for Individual Drivers

The above record of tickets are marked with driver-specific identifier, which serves as a key for a dataset of driver-specific characteristics. This dataset contains the driver identification number, along with the gender and date of birth of each driver. This information is not publicly available to protect the privacy of the drivers.

The official name of the database is *Fichier des numéros séquentiels*. The file `seq.dta` contains licensee data for 3,911,743 individuals who received tickets and includes the following variables.

- `seq` or *Numéro séquentiel* is a sequence of unique identification numbers for the drivers.
- `sxz` or *Sexe* is either 1.0 or 2.0, an indicator for male or female, respectively.
- `an` or *Année de naissance provenant du numéro de permis de conduire* is an integer for the year of birth of each driver.
- `mois` or *Mois de naissance provenant du numéro de permis de conduire* is an integer for the month of birth of each driver.
- `jour` or *Jour de naissance provenant du numéro de permis de conduire* is an integer for the calendar day of birth of each driver.

Aggregate Counts of Drivers

Counts of individual drivers were obtained from the Website of the Banque de données des statistiques officielles sur le Québec, available at `https://bdso.gouv.qc.ca/pls/ken/ken213_afich_tabl.page_tabl?p_iden_tran=REPERRUNYAW46-44034787356%7C@%7Dzb&p_lang=2&p_m_o=SAAQ&p_id_ss_domn=718&p_id_raprt=3370#tri_pivot_1=500400000`.

The statistics were compiled into a single spreadsheet `SAAQ_drivers_annual.csv`, which is available in the `Data` folder and contains the following variables.

- `age_group` is an age range in years.
- `sex` is an indicator for the gender of drivers, either `"M"` or `"F"`.
- `yrYYYY` denotes that the column records the number of drivers in each year `YYYY` on June 1 of each year.

Instructions

All regression results, tables and figures in the manuscript can be obtained by running the shell script `SAAQ_CJE.sh`. The workflow proceeds in three stages: one set of instructions outlines the operations to transform the raw data in the SAAQ database into the dataset that is the input for the statistical analysis in the next stage. In the final stage, the estimation results are used to create the figures and tables for the manuscript.

Data Preparation

Run the scripts in the `Code/Prep` folder, which perform the following operations:

1. Run the R script `SAAQ_tickets.R`, which collects the record of tickets for each year into a single dataset of tickets. This produces the dataset `SAAQ_tickets.csv`, which is the record of events in the regression models.

2. Run the R script `SAAQ_point_balances.R`, which calculates the accumulated demerit point balances for each driver and collects counts of drivers at each demerit point level. This produces the dataset `SAAQ_point_balances.csv`, which is the record of counts of drivers at each demerit point level for each day in the sample period. This is the record of non-events for the subset of drivers *who have ever received tickets*.

3. Run the R script `SAAQ_driver_counts.R`, which collects the public record of the number of drivers in each gender and age group category. It uses linear interpolation to transform the dataset `SAAQ_drivers_annual.csv` into a record of daily counts `SAAQ_drivers_daily.csv`. This dataset is the the record of non-events for the subset of drivers *who have never received tickets*.

4. Run the R script `SAAQ_join.R`, which joins the above datasets into the complete record of events and non-events for all drivers in Quebec. This produces the dataset `SAAQ_full.csv`, which is used in the regression analysis in the next stage.

Statistical Analysis

The script in the `Code/Reg` folder is the main script for the sequence of regression models.

Run this script, `SAAQ_Regs.R`, which estimates all models in the paper in a series of loops. It perform the following operations:

1. Read in the main dataset `SAAQ_full.csv`.

2. Create and modify categorical variables.

3. Define the policy indicator to represent the change in legislation on April 1, 2008 and the sample period over the four-year period centered on this date.

4. Defines the sequence of sets of models to be estimated, including the full sample, high-point drivers, an event study and an analysis by demerit point balances, as well as placebo regressions.

For each model, the script performs the following operations:

1. Define the target variable.

2. Set the relevant sample period, which differs for the placebo regression.

3. Set the sample selection, to select male or female drivers and to select either the full sample or high-point drivers.

4. Estimate the linear and logistic regression model.

5. Calculate HCCME standard errors for the linear probability model.

6. Calculate the marginal effects for the relevant coefficients.

7. Save the estimation results in files stored in the `Estn` folder to produce tables and figures for the manuscript.

Manuscript

Once the estimates are obtained, run a series of scripts to draw from values in the estimation results to produce the figures and tables in the manuscript.

Producing the Output

Run the scripts in the `Code/Out` folder perform the following operations:

1. Run the script `SAAQ_Tables.R`, which produces tables of estimates from the results in the `Estn` folder. These tables are all output to the `Tables` folder.

2. Run the script `SAAQ_Estn_Figs.R`, which produces the figures from the estimation of the event studies and the estimation with granular demerit-point categories. These figures are output to the `Figures` folder and are ultimately named `Figure3.eps` and `Figure4.eps`.

3. Run the script `SAAQ_Count_Figs.R`, which produces the figures of the frequency of tickets from aggregate data by month. This produces `num_pts_5_10.eps` and `num_pts_7_14.eps`, which are both output to the `Figures` folder and are ultimately named `Figure1.eps` and `Figure2.eps`. It also outputs a dataset `Point_Freq_Gender_Ratio.csv`, which is used to calculate the summary statistics in Table 2.

Producing the Tables Separately

All tables in the manuscript were output to the folder `Tables`.

1. Table 1 was produced manually and appears in the file `Penalties.tex`.

2. Table 2 was produced by an Excel spreadsheet `Point_Freq_Gender_Ratio.xlsx` from the outputs in `Point_Freq_Gender_Ratio.csv` and appears in the file `Point_Freq_Gender_Ratio.tex`.

3. Tables 3, 4, 5, 6 and 7 were produced together from the commands on lines 248 to 258 of the script `SAAQ_Tables` using the regression results obtained above and the function library `SAAQ_Tab_Lib.R` in the folder `Code/Lib`.

Producing the Figures Separately

All figures in the manuscript were output to the folder `Figures`.

1. Figure 1 was produced from the commands on lines 263 to 284 of the script `SAAQ_Count_Figs.R`.
2. Figure 2 was produced from the commands on lines 310 to 331 of the script `SAAQ_Count_Figs.R`.
3. Figure 3 was produced from the commands on lines 156 to 189 of the script `SAAQ_Estn_Figs.R` using the regression results obtained above.
4. Figure 4 was produced from the commands on lines 252 to 292 of the script `SAAQ_Estn_Figs.R` using the regression results obtained above.

Libraries

The above programs use functions defined in the following libraries, which are stored in the `Code/Lib` folder.

- The script `SAAQ_Agg_Reg_Lib.R` defines functions for running regressions with data aggregated by the number of driver days for each combination of the dependent variables. Since weighted regression is used in different contexts, this library makes adjustments, such as for degrees of freedom, to make the results equivalent to those which would be obtained from the full dataset with one observation per driver per day. Since most drivers do not get tickets on most days, this library effectively compresses the dataset by a factor of one thousand, from billions of driver days to millions of unique observations.
- The script `SAAQ_Agg_Het_Lib.R` defines functions for the calculation of heteroskedasticity-corrected standard errors with aggregated data.
- The script `SAAQ_Reg_Lib.R` defines helper functions for data formatting and preparation for regressions.
- The script `SAAQ_MFX_Lib.R` defines functions to calculate marginal effects.
- The script `SAAQ_Tab_Lib.R` defines functions to generate LaTeXtables from regression results.

Computing Requirements

All the tables and figures in the paper can be performed on a single microcomputer, such as a laptop computer. The particular model of computer on which the statistical analysis was run is a Dell Precision 3520, running a 64-bit Windows 10 operating system, with a 4-core x64-based processor, model Intel(R) Core(TM) i7-7820HQ CPU, running at 2.90GHz, with 16 GB of RAM.

Software

The statistical analysis was conducted in R, version 4.0.2, which was released on June 22, 2020, on a 64-bit Windows platform x86_64-w64-mingw32/x64.

The attached packages include the following:

- `foreign` version 0.8-81, to open datasets in `.dta` format.
- `data.table`, version 1.13.0 (using 4 threads), to handle the main data table for data preparation and analysis in the scripts in the `Code/Prep` and `Code/Reg` folders.
- `xtable`, version 1.8-4, to generate LaTeXtables for Tables 3, 4, 5, 6, and 7.
- `scales` version 1.1.1, to format numbers in LaTeXtables.

Upon attachment of the above packages, the following packages were loaded via a namespace, but not attached, with the following versions:

- `Rcpp` version 1.0.5
- `RcppParallel` version 5.0.2
- `parallel` version 4.0.2
- `compiler` version 4.0.2
- `pkgconfig` version 2.0.3
- `haven` version 2.3.1
- `stringr` version 1.4.0
- `withr` version 2.4.2
- `tidyr` version 1.1.3
- `car` version 3.0-10
- `scales` version 1.1.1
- `stringi` version 1.5.3

Acknowledgements

The authors would like to thank François Tardif for his help with the data in the early stages of this project, as well as Catherine Maclean for helpful suggestions and valuable comments. Jeffrey Penney acknowledges support from SSHRC. The authors are especially grateful to the editor and two anonymous referees for comments and suggestions that led to substantial improvements from the original manuscript. The authors have no conflict of interest to disclose. The usual caveat applies.

References

Société de l'assurance automobile du Québec, *Fichiers des infractions au Code de la sécurité routière*, 1998–2010, accessed February 2012.

Société de l'assurance automobile du Québec, *Fichier des numéros séquentiels*, 1998–2010, accessed February 2012.

Banque de données des statistiques officielles sur le Québec, *Nombre de titulaires d'un permis de conduire ou d'un permis probatoire selon le sexe et l'âge, Québec et régions administratives* https://bdso.gouv.qc.ca/pls/ken/ken213_afich_tabl.page_tabl?p_iden_tran=REPERRUNYAW46-44034787356%7C@%7Dzb&p_lang=2&p_m_o=SAAQ&p_id_ss_domn=718&p_id_raprt=3370#tri_pivot_1=500400000.