Stata Code Sample [*]

Xiling (Celia) Zhu xiling@uchicago.edu

Dec 23, 2020

0. Overview

This task was inspired by IO research on the automobile industry. I was provided with car model sales data in five European markets. `car data` contains the manufacturing characteristics and the quantity sold of each car model within each market, from 1970 to 1990. `market data` contains the GDP, population, and tax rate of each market in each year.

1. Data Cleaning

The central task of the section is to merge `car data` and `market data` into one panel with 3 dimensions: car model (i), market (j), and year (t). In each row, it contains car model manufacturing characteristics, price, and quantity sold, market GDP, population, and tax rate.

Set up directories to store the data, figures, tables, and log, respectively.

```
. clear all

. set more off

. set varabbrev off

. 
. * Data directory
. global data "/Users/celiazhu/Box/projects/ra_code_sample/io_code_sample/data"

. 
. * Figure directory
. global figure "/Users/celiazhu/Box/projects/ra_code_sample/io_code_sample/output/figure"

. 
. * Table directory
. global table "/Users/celiazhu/Box/projects/ra_code_sample/io_code_sample/output/table"
```

Append all car data.

```
. local carcsv: dir "$data/car_data" files "*.csv"

. foreach file of local carcsv {
  2.   preserve
  3.   qui insheet using "$data/car_data/`file´", clear
```

[*] Generated by `markstat`. For source code, please see the link here.

```
  4.   qui cap destring sp, force replace // change sp (maximum speed) from string to numeric
  5.   qui replace ye = 1900 + ye // use full year
  6.   foreach var of varlist ma loc {
  7. qui replace `var´ = "United Kingdom" if `var´ == "UK" // use full country name
  8.   }
  9.   qui save temp, replace
 10.   restore
 11.   qui append using temp
 12. }

. save "$data/clean_data/car_data_all.dta", replace
file /Users/celiazhu/Box/projects/ra_code_sample/io_code_sample/data/clean_data/car_data_all.dta saved

. rm temp.dta
```

Append all market data.

```
. clear

. local mktcsv: dir "$data/market_data" files "*.csv"

. foreach file of local mktcsv {
  2.   preserve
  3.   qui insheet using "$data/market_data/`file´", clear
  4.   qui save temp, replace
  5.   restore
  6.   qui append using temp
  7. }
```

Spell out full country names in appended market data.

```
. local country_list `" "Belgium" "France" "Germany" "Italy" "United Kingdom" "´

. /* Abbreviated market name corresponds to the first letter of the full country
> name, and the correspondence is unique in this dataset. */
. foreach country in `country_list´ {
  2. qui replace ma = "`country´" if ma == substr("`country´", 1, 1)
  3. }

. save "$data/clean_data/market_data_all.dta", replace
file /Users/celiazhu/Box/projects/ra_code_sample/io_code_sample/data/clean_data/market_data_all.dta saved

. rm temp.dta
```

Merge car and market data.

```
. merge 1:m ye ma using "$data/clean_data/car_data_all.dta", nogen

    Result                           # of obs.
    -----------------------------------------

    not matched                             0
    matched                             7,679
    -----------------------------------------
```

Fix missing values of fuel consumption variables.

```
. qui destring li*, force replace

. replace li = (li1 + li2 + li3)/3 if li == .
(1,919 real changes made)

. foreach var of varlist li1 li2 li3 {
  2.   qui replace `var´ = 0 if `var´ == .
  3.   qui replace `var´ = li*3 - (li1 + li2 + li3) if `var´ == 0
  4. }
```

Fix data type.

```
. qui destring ac, force replace // ac (time to acceleration) is float

. foreach var of varlist ma loc brand model cla frm {
  2.   qui encode `var´, gen(`var´_code) // transform string to categorical variables
  3. }
```

Label needed variables.

```
. label var hp "Horsepower (kW)"

. label var li "Fuel consumption (liter per km)"

. label var eurpr "Price in common currency"
```

Store cleaned data.

```
. qui save "$data/clean_data/all_data.dta", replace
```

2. Data Exploration

This section uses the model-market-year panel dataset to visualize the relationship between fuel consumption (`li`) and horsepower (`hp`) in the years 1970 and 1990.

```
. preserve

. // a temporary dataset for 1970 data
. tempfile data_70

. * keep 1970 data
. keep if ye == 1970
(7,407 observations deleted)

. * keep neede variables
. keep ye qu hp li ma_code model_code

. qui save `data_70´, replace

. restore

. 
. preserve

. // a temporary dataset for 1970 data
. tempfile data_90

. * keep 1990 data
. keep if ye == 1990
(7,281 observations deleted)

. * keep needeed variables
. keep ye qu hp li ma_code model_code

. qui save `data_90´, replace

. restore
```

For the years 1970 and 1990, group cars by decile of observed horsepower in that year, and then compute the sales-weighted average of fuel consumption for cars in each horsepower decile.

For each year, produce a scatter plot of the sales-weighted average of fuel consumption versus the midpoint of each horsepower decile.

For each year, regress fuel consumption on a constant, horsepower, and log(horsepower), using sales as sample weights. Display the fitted curves on the scatterplot.

```
. /* 1) group cars by decile of horsepower; 2) fit curves; 3) summary statistics
> of horsepower and sales in each decile group;
>
> compute midpoint of each decile group;
>  compute sales-weighted average of fuel consumption; 4) fit curves*/
. foreach year of numlist 70 90 {
  2.   use `data_`year´´, clear
  3.   // 1) group cars by decile of horsepower
.   xtile hp_decile = hp, nq(10)
  4.   bysort hp_decile: asgen wtd_avg_li = li, weight(qu)
  5.   // 2) fit curves; use sales as sample weights
.   qui gen loghp = ln(hp)
  6.   sort ye ma_code model_code
  7.   qui reg li hp loghp [pw = qu]
  8.   qui predict li_hat
  9.   // 3) summary statistics of horsepower and sales in each decile group
.   qui gen hp_mid = .
 10.   qui gen hp_min = .
 11.   qui gen hp_max = .
 12.   qui gen qu_total = .
 13.   * loop over each horsepower decile group
.   forvalues  i = 1/10 {
 14. * horsepower summary statistics
. qui su hp if hp_decile == `i´
 15. local hp_min r(min)
 16. local hp_max r(max)
 17. qui replace hp_min = `hp_min´ if hp_decile == `i´
 18. qui replace hp_max = `hp_max´ if hp_decile == `i´
 19. * midpoint of each decile group for x-axis of scatterplot
. qui replace hp_mid = (hp_min + hp_max)/2 if hp_decile == `i´
 20. * sales summary statistics
. qui su qu if hp_decile == `i´
 21. local qu_total r(sum)
 22. qui replace qu_total = `qu_total´ if hp_decile == `i´
 23.   }
 24.   qui save `data_`year´´, replace
 25. }
.
. * Aggregated data of fuel consumption and horsepower in 1970 and 1990
. use `data_70´, clear

. append using `data_90´
(label model_code already defined)
(label ma_code already defined)

. collapse (first) wtd_avg_li hp_min hp_max hp_mid li_hat qu_total, by(ye hp_decile)

. label var wtd_avg_li "Sales-weighted average of fuel consumption"

. label var hp_mid "Midpoint of horsepower decile"
```

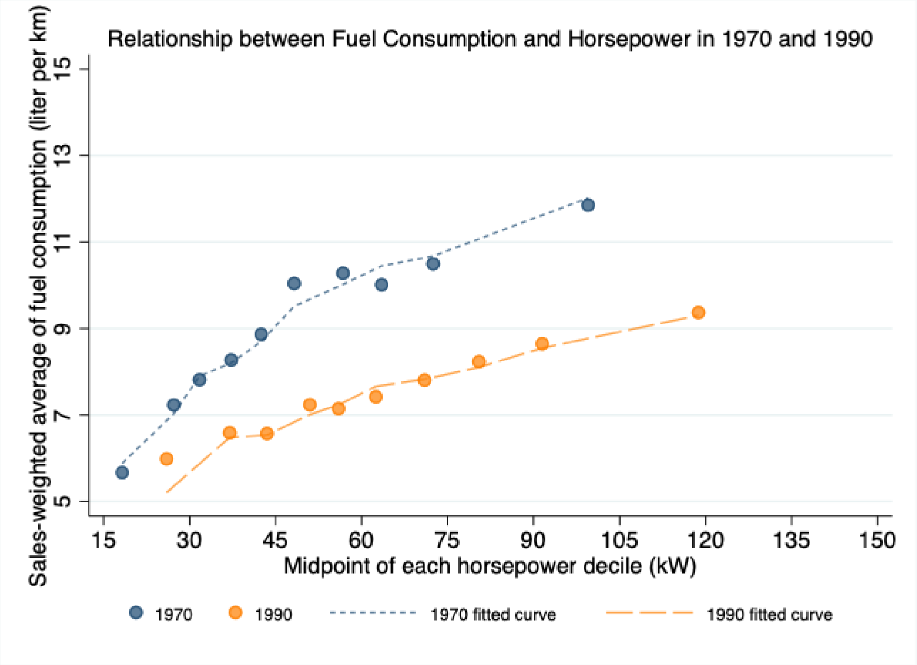
Visualize the relationship between fuel consumption and horsepower.

```
. local file_name relation_li_hp

. twoway (scatter wtd_avg_li hp_mid if ye == 1970, mcolor(navy%70)) ///
> (scatter wtd_avg_li hp_mid if ye == 1990, mcolor(orange%70)) ///
> (line li_hat hp_mid if ye == 1970, sort lcolor(navy%70) lpattern(shortdash)) ///
> (line li_hat hp_mid if ye == 1990, sort lcolor(orange%70) lpattern(longdash)), ///
> xlabel(15(15)150) ylabel(5(2)15) ///
> xtitle("Midpoint of each horsepower decile (kW)", size(medsmall)) ///
> ytitle("Sales-weighted average of fuel consumption (liter per km)", size(medsmall) margin(top)) ///
> title("Relationship between Fuel Consumption and Horsepower in 1970 and 1990", ///
> size(medsmall) color(black)) ///
> legend(label(1 "1970") label(2 "1990") ///
> label(3 "1970 fitted curve") label(4 "1990 fitted curve") ///
> nobox region(lcolor(white)) size(small) rows(1)) ///
> graphregion(color(white))

. graph export "$figure/`file_name´.png", replace
(file /Users/celiazhu/Box/projects/ra_code_sample/io_code_sample/output/figure/relation_li_hp.png written i
```

Figure 1: Relationship between fuel consumption and horsepower by year



Based on figure 1, in general, the sales-weighted average fuel consumption increased as the horsepower increased. Comparing with 1970, the sales-weighted average fuel consumption decreased in 1990 in all horsepower decile groups, and the fitted curve was flatter in 1990. We can extrapolate that, controlling for horsepower, the sales-weighted average fuel consumption decreased over time – cars in the five European markets became more fuel-efficient.

Suppose a social cost of carbon was imposed across Europe in 1991, causing the price of gas to increases across all five markets. This imposed social cost of carbon would disincentivize consumers to buy cars that have a high fuel consumption, or the cars that have a high horsepower, which positively correlates with fuel consumption. As a result, the sales-weighted average fuel consumption of the cars in the high horsepower group would fall in 1991. The fitted curve in 1991 would be flatter than the one in 1990, especially in the tail of the curve.

Please see table 1 for summary statistics of the sales-weighted average of fuel consumption by decile of horsepower in 1970 and 1990.

```
. * prepare data to generate texsave table
. qui gen wtd_avg_li_3 = string(wtd_avg_li,"%4.3f") // round up to 3 decimal places

. qui drop wtd_avg_li

. rename wtd_avg_li_3 wtd_avg_li

. qui reshape wide wtd_avg_li hp_min hp_max hp_mid li_hat qu_total, i(hp_decile) j(ye)

. foreach year of numlist 1970 1990 {
  2.   * horsepower range of each decile group
.   egen hp_range`year´ = concat(hp_min`year´ hp_max`year´), punct("--")
  3.   * insert comma in total sales quantity
.   gen qu_total`year´_comma = string(qu_total`year´, "%15.0fc")
  4.   drop qu_total`year´
  5.   rename qu_total`year´_comma qu_total`year´
  6. }
.
. * reorder variables
. order hp_decile wtd_avg_li1970 hp_range1970 qu_total1970 ///
> wtd_avg_li1990 hp_range1990 qu_total1990
.
. * drop un-needed variables
. drop hp_min* hp_max* hp_mid* li_hat*
.
. * generate summary statistics table
. local file_name hp_sum_table

. local title title("Sales-weighted average of fuel consumption by decile of horsepower")

. local midliners "\cmidrule(lr){2-4} \cmidrule(lr){5-7} \addlinespace[-2.5ex]"

. local colnames "{Decile} &{Fuel consumption} &{Horsepower range} &{Sales} &{Fuel consumption} &{Horsepowe

. local headerlines headerlines("& \multicolumn{3}{c}{1970} & \multicolumn{3}{c}{1990}" "`midliners´" "`col

. local fn footnote("Fuel consumption represents the sales-weighted average of fuel consumption.")

. local marker marker("tab:1")

. local size size("small")

. texsave using "$table/`file_name´.tex", replace ///
> loc(H) frag nonames `headerlines´ `marker´ `fn´ `size´ `title´
```

Table 1: Sales-weighted average of fuel consumption by decile of horsepower

	1970			1990		
Decile	Fuel consumption	Horsepower range	Sales	Fuel consumption	Horsepower range	Sales
1	5.666	13–23.5	814,581	5.985	19–33	2,197,737
2	7.232	25–29.5	1,095,414	6.586	34–40	1,556,830
3	7.815	30.5–33	769,165	6.569	41–46	1,045,728
4	8.270	34–40.5	1,157,252	7.237	48–54	1,024,291
5	8.867	41–44	234,318	7.147	55–57	1,042,327
6	10.041	45–51.5	495,287	7.419	59–66	944,924
7	10.279	53–60.5	285,753	7.806	67–75	470,646
8	10.012	61–66	287,505	8.233	76–85	494,383
9	10.495	67–78	326,636	8.648	87–96	345,608
10	11.855	81–118	133,559	9.369	96.5–141	232,187

Fuel consumption represents the sales-weighted average of fuel consumption.

3. Estimation and Causal Inference

For each car model i, market j, and year t, construct the outcome variable $Y_{ijt} = log(S_{ijt}) - log(S_{0ijt})$, where N_{jt} is the number of consumers in market j year t, assuming the average size of family is four and each family potentially buys one car or no cars in a given year, S_{ijt} is the market share for car model i in market j year t, S_{0jt} is the share of consumers who buy no cars in market j year t.

A standard logit demand model:

$$U_{cijt} = \beta_c \text{FuelConsumption}_{ijt} + \alpha_c Price_{ijt} + \zeta_{ijt} + \epsilon_{cijt}$$

where c is an index for car consumers, i car model, j market, and t year. U is the indirect utility, as a function of the car models' average fuel consumption and price. ζ_{ijt} is the un-observables on the year-market-model level. ϵ_{cijt} is un-observables on the year-market-model-consumer level.

Assume the structural parameters, β and α, do not depend on individual consumer (every consumer has the same taste for fuel consumption and price), ϵ_{cijt} has type-one extreme value function, and all consumers who did not buy a car chose the outside option, which gives zero utility. Then, we can estimate the structural parameter through the following log-linear model:

$$log(S_{ijt}) - log(S_{0ijt}) = \beta \text{FuelConsumption}_{ijt} + \alpha Price_{ijt} + \zeta_{ijt} \tag{1}$$

ζ_{ijt} is a shock on the demand side.

The results of this regression is in model (1) of table 2. We estimate the coefficient of fuel consumption, β, to be -0.214. It means that, holding everything else equal, the consumers' indirect utility decreases by 0.214 unit as the fuel consumption of the car increases by 1 unit.

```
. use "$data/clean_data/all_data.dta", clear

.
. * Generate the number of consumers in market j in year t
. gen n_consumers = pop/4 // 4 is average family size; assume each family potentially buys one car

. label var n_consumers "Number of consumers"

.
. * Generate market share of car model i in market j in year t
. gen share = qu/n_consumers

. label var share "Market share"

.
. * Generate share of consumers who by no cars in market j in year t
. bysort ma_code ye: egen qu_total = total(qu)

. gen share_no_cars = 1 - qu_total/n_consumers

. label var share_no_cars "Share of consumers who buy no cars"

.
. * Generate outcome variable
. gen outcome = ln(share) - ln(share_no_cars)
. label var outcome "Outcome"

.
. * Regress outcome variable on a constant, fuel consumption, and price
. eststo clear

. eststo: qui reg outcome li eurpr, r
(est1 stored)

. qui estadd local fe "None"
```

Price is an endogenous variable in model 1. Price is correlated with ζ_{ijt}, the demand shock. Assuming the demand shock is observable to manufacturers, then manufacturers would adjust the prices accordingly.

There are two different sources of endogeneity that could bias our estimate of the structural parameter of price.

One source of endogeneity is the un-observalbe features of the car models, markets, or years. For example, consumers may prefer one particular type of car models for reasons other than its fuel consumption or price, but for the models' un-observable features. Denote them as ϕ_i. Similarly, consumers may experience some market-specific, model-and-year-invariant shock, ϕ_j, or some year-specifc, model-and-market-invariant shock, ϕ_t.

We can mitigate this type of endogeneity by including car model, market, and year fixed effects in model 1.

$$log(S_{ijt}) - log(S_{0ijt}) = \beta \text{FuelConsumption}_{ijt} + \alpha Price_{ijt} + \zeta_{ijt} + \phi_i + \phi_j + \phi_t \quad (2)$$

We report the coefficients of fuel consumption and price in model (2) of table 2.

```
. * Including car model, market, and year fixed effects
. sort ye ma_code model_code

. eststo: qui reg outcome li eurpr i.ma_code i.model_code i.ye, r
(est2 stored)

. qui estadd local fe "car model, market, and year"

.
. * Output table for two models
. local file_name linear

. local title "Structural parameters for fuel consumption and price"

. local label \label{tab:2}

. sort ye ma_code model_code

. local reg_tbl_setting "se booktabs width(\textwidth) label" // settings used for all regression tables

. qui esttab using "$table/`file_name´.tex", replace ///
> drop(*.ma_code *.model_code *.ye) scalars("fe Fixed effects") ///
> title("`title´ `label´") nofloat ///
> `reg_tbl_setting´
```

Table 2: Structural parameters for fuel consumption and price

	(1) Outcome	(2) Outcome
Fuel consumption (liter per km)	-0.214*** (0.00926)	-0.138*** (0.0178)
Price in common currency	-0.0000578*** (0.00000421)	-0.000103*** (0.0000102)
Constant	-5.067*** (0.0784)	-4.610*** (0.217)
Observations	7679	7653
Fixed effects	None	car model, market, and year

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Another source of endogenity comes from the market structure. For example, if the car markets in the five European countries are not perfectly competitive, there still exists potential correlation between price and demand shock even if we controlled for all fixed effects.

We need to use an instrument variable for price. One candidate of the instrument variable is the transportation cost: the cost needed to transport cars from their manufacturing locations to their markets. This is because the transportation cost is the supply-shifter that does not shift the demand: it's uncorrelated with the demand shock ζ_{ijt}.