Stata Code Sample

Xiling (Celia) Zhu xiling@uchicago.edu

February 14, 2021

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.

2 Data Exploration

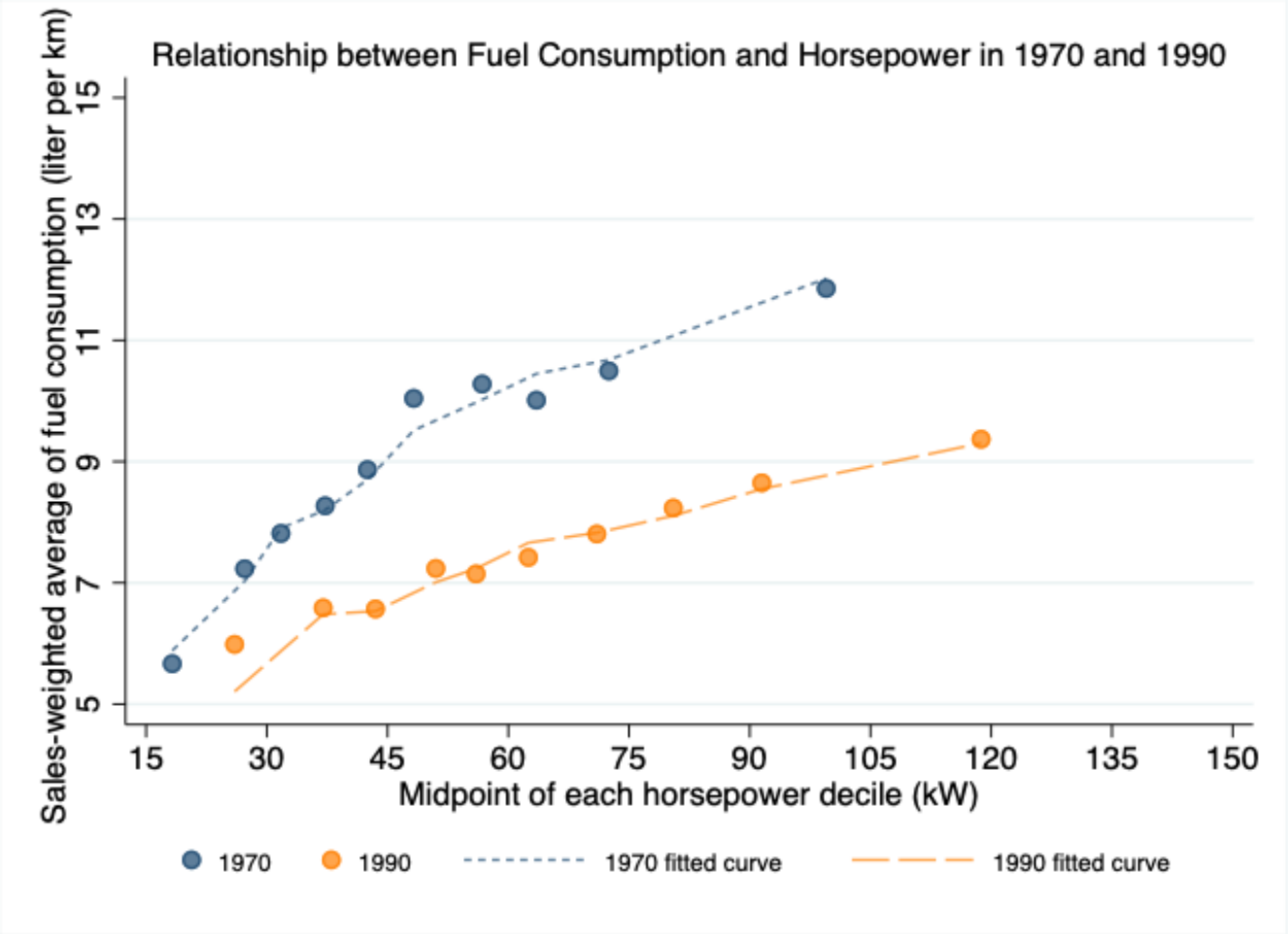
This section uses the model-market-year panel dataset to visualize the relationship between fuel consumption (`li`) and horsepower (`hp`) in the years 1970 and 1990.

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.

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.

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

Decile	1970			1990		
	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:

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

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 unobservable variables on the year-market-model level. ϵ_{cijt} is unobservable variables 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:

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

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

Table 2 reports the results of this regression. 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.

Table 2: Structural parameters for fuel consumption and price

	(1) Outcome	(2) Outcome
Fuel consumption (liter per km)	-0.214***	-0.214***
	(0.009)	(0.031)
Price in common currency	-0.000***	-0.000***
	(0.000)	(0.000)
Constant	-5.067***	-5.053***
	(0.078)	(0.295)
Observations	7679	7653
Fixed effects	None	None
Standard errors	Robust	Clustered on car models

Standard errors in parentheses

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

Table 2 also reports the standard errors under different specifications. Column (1) reports the heteroskedastic robust standard errors and column (2) the standard errors clustered on car models. To cluster the standard errors on car models, we assume the observations are independent across car models but not necessarily within each model. This assumption is not entirely plausible because different car models may be similar in design. Other specifications of standard errors are less convincing. For instance, to cluster on years, we need to assume that observations are independent across years. But it is reasonable to assume that consumers' preferences are auto-correlated.

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.

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

Controlling for car model, market, and year fixed effects, we report the coefficients of fuel consumption and price in table 3

Table 3: Structural parameters for fuel consumption and price (added fixed effects)

	(1) Outcome	(2) Outcome
Fuel consumption (liter per km)	-0.138***	-0.138***
	(0.018)	(0.035)
Price in common currency	-0.000***	-0.000***
	(0.000)	(0.000)
Constant	-4.610***	-4.610***
	(0.217)	(0.360)
Observations	7653	7653
Fixed effects	Car model, market, and year	Car model, market, and year
Standard errors	Robust	Clustered on car models

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.

To mitigate the second source of endogenity, 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}.