

Pseudo-panel approach for repeated cross-sectional data - an introduction

Yang Meng

Sheffield Alcohol Research Group
School of Health and Related Research
University of Sheffield



© The University of Sheffield

Researchers prefer longitudinal panel data

- Longitudinal panel data, where the same measures are observed repeatedly over time for the same individuals, are generally regarded as superior for identifying relationships among variables compared with other types of data (e.g., cross-sectional and aggregate time series data).
- Advantages of panel data include:
 - Individuals can be used as their own controls in panel models (e.g., fixed-effect models, random-effect models) to control for both observed and, more importantly, unobserved time-constant heterogeneities among different individuals.
 - Some researchers argue only panel data is able to identify causal relationships among variables, while cross-sectional data can only identify association among variables.

© The University of Sheffield



But, scarcity of panel data, abundant of repeated sections

- Unfortunately, such individual-level panel data is scarce and not available in most countries. In stead, cross-sectional data are much more common, and many of these are established population surveys and are carried out repeatedly over a long period of time.
- Cross-sectional surveys
 - Data collected by observing many subjects (e.g., individuals, households) at the same point of time, or without regard to differences in time. Analysis of cross-sectional data usually consists of comparing the differences among the subjects.
- Some advantages of cross-sectional data over genuine panel data:
 - Suffer less from attrition and nonresponse
 - Generally larger in sample size
 - Long time period that the survey spans

© The University of Sheffield



Examples of UK cross sections

- Examples of large scale repeated cross-sectional surveys in the UK
 - General Lifestyle Survey (formally known as the General Household Survey): started from 1971 and around 8,000/15,000 households/individuals are surveyed annually.
 - Health survey for England: started from 1994 and around 10,000 individuals are surveyed annually.
 - Living Cost and Food Survey (formally known as the Expenditure and Food Survey): started from 2001 and around 5000 households are surveyed annually.
 - Labour Force Survey
 - British Crime Survey
 - ...

© The University of Sheffield

One solution – pseudo panel approach

- When repeated cross-sectional data is available, the pseudo-panel approach offers a method to “*longitudinalise/panelise*” the cross-sectional data and enables the use of panel models on the constructed “*pseudo-panels*”.
- A member of the pseudo-panel is defined as a subgroup of the population with fixed membership, individuals of which can be identified as they appear in repeated cross-sectional surveys. Examples of subgroups in a pseudo panel:
 - Males born between 1976 and 1980 who are of white ethnicity
- Key features of the pseudo panel approach
 - Replace individual observations in genuine panel with subgroup means.
 - Track subgroups through time in repeated cross sections.
 - Time series for the subgroup means can be used as if panel data were available.
 - A possible link between individual level data and (national) aggregate data.

Some references

- The pseudo-panel approach has been around for few decades, and has been developed ever since:
 - Deaton 1985
 - Moffitt 1993
 - Verbeek and Vella 2005
 - Verbeek 2008
- And the method has been empirically applied to different disciplines, for example:
 - Car ownership (Dargay et al. 1999)
 - Price elasticities of alcohol demand (Meng et al. 2014)

Defining panel members in a pseudo-panel

- The factors for defining subgroups (i.e., members of the pseudo-panel) need to be time-invariant (e.g., birth year, gender, ethnicity) or can be reasonably assumed to be time-invariant.
- Given a fixed number of individuals in the repeated cross-sectional datasets, N , there is a trade off between the number of defined subgroups, C , and the number of individuals within each defined subgroups for each time period, n_c
 - More subgroups \gg increased heterogeneity/variations of members in the pseudo-panel.
 - More individuals in a subgroup per time period \gg more robust estimation of subgroup- and period-specific means .
 - Both are preferable, but you can't have both! $N = C * n_c * T$
 - Empirical guidance: $n_c > 100$.

References

- Dargay, J.M., Vythoulkas, P.C., 1999. Estimation of a dynamic car ownership model – a pseudo-panel approach. *Journal of Transport Economics and Policy* 33, 287–301.
- Deaton, A., 1985. Panel data from time-series of cross-sections. *Journal of Econometrics* 30, 109–126.
- Meng Y, Brennan A, Purshouse R, Hill-McManus D, Angus C, et al. Estimation of own and cross price elasticities of alcohol demand in the UK – a pseudo-panel approach using the Living Costs and Food Survey 2001 to 2009. *Journal of Health Economics* 2014 34: 96-103.
- Moffitt, R., 1993. Identification and estimation of dynamic-models with a time-series of repeated cross-sections. *Journal of Econometrics* 59, 99–123.
- Verbeek, M., 2008. Pseudo-panels and repeated cross-sections. In: Matyas, L., Sevestre, P. (Eds.), *The Econometrics of Panel Data*. Springer-Verlag, Berlin, pp.369–383.
- Verbeek, M., Vella, F., 2005. Estimating dynamic models from repeated cross-sections. *Journal of Econometrics* 127, 83–102.

Estimation of own and cross price elasticities of alcohol demand in the UK – a pseudo-panel approach

Yang Meng, Alan Brennan, Robin Purshouse,
Daniel Hill-McManus, Colin Angus, John
Holmes, Petra Meier

Sheffield Alcohol Research Group

School of Health and Related Research

University of Sheffield



© The University of Sheffield

Background

- Estimation of own- and cross-price elasticities of alcohol demand for different beverages is important for the evaluation of price-based policies such as tax/duty increases, setting floor prices for retail sale (e.g., minimum unit price, ban on blow-cost sales).
- There have been numerous studies on own-price elasticities for main beverage types (e.g., beer, spirits, wine) and for alcohol as a whole (see for example Fogarty et al 2010; Wagenaar et al 2009; Gallet 2007).
- However, there have been limited studies examining cross-price elasticities, or own-price elasticities for alcohol sold in different premises (off- vs on-licensed), which are potentially important for evaluating real-world policy interventions.

© The University of Sheffield

Background

- Longitudinal panel data, where both consumption and price-paid for different beverages are observed repeatedly over time for the same individuals, are generally superior to cross-sectional and aggregate time series data for estimating elasticities.
- One key advantage of panel data is that individuals can be used as their own controls in panel models (e.g., fixed-effect models, random-effect models) to control for both observed and unobserved time-constant heterogeneities among different individuals.
- Panel data offers advantages to identify causal interrelationships among variables, compared with other types of data.

Pseudo-panel approach

- Unfortunately, such individual-level panel data is not available in most countries. In stead, cross-sectional data or aggregate time series data are mostly commonly used data for estimating elasticities.

Data and defining the pseudo-panel

- Data: 9 Living Cost and Food Surveys (LCF, formerly Expenditure and Food Survey) from 2001/2 to 2009.
 - Prices and income were adjusted with December 2009 as the base period.
 - Alcohol purchase quantity were uplifted to be in line with per capita sales data (beverage specific annual adjustment factors were used).
- In the base case, a pseudo-panel with 72 subgroups were defined by:
 - Gender (male, female); 12 birth cohorts (born between 1930-1934, 1935-1940, ..., 1985-1989); and 3 social-economic groups (higher, middle and lower).
 - $n_c = 140$, where $N=90,652$, $C=72$, $T=9$.
 - Observations where $n_c < 30$ were excluded from the analysis to ensure robust estimation of mean statistics
- Sensitivity analyses were performed:
 - 96 subgroups (4 vs 3 socio-economic groups)
 - 48 subgroups (2 regions vs 3 socio-economic groups)
 - 96 subgroups (4 regions vs 3 socio-economic groups)

Dependent and independent variables

- Dependent: mean units purchased for 10 modelled beverages by each subgroup each time period, C_{ijt} .
- Independent:
 - All models: mean price per unit (PPU) for the 10 beverages paid by each subgroup each time period, P_{ijt} ; mean income by each subgroup each time period, $Income_{jt}$; Year dummy.
 - Time-variant (tested for all models): proportion of individuals having children, being married, being unemployed, and smoking by each subgroup each time period, KID_{jt} , MRD_{jt} , UNE_{jt} , SMK_{jt} ; square of the mean age of the subgroup, Age^2_{jt} .
 - Time-invariant (tested for random-effect models only): gender, socio-economic status, birth year.

Model specification

- Types of models tested:
 - Fixed effect models, Random effect models, Ordinary least squares (OLS)
- Assumptions for the base case:
 - Models were fitted separately for the 10 beverages.
 - Lagged dependent variables were excluded.
 - Log-log functional form for the dependent and independent variables of P_{ijt} and $Income_{it}$.
 - Other dependent variables were tested as levels (original measurement).
 - η_e used as weights for fitting fixed effect and OLS models.
- Model tests and coding:
 - Hausman tests for fixed vs random effect models
 - t-test and F-test for inclusion/exclusion of non-PPU/Income independent variables
 - Development in STATE/SE 12.1

Model specification

- Example of the unrestricted fixed-effect model for off-trade beer:

$$\ln C(offbeer)_{it} = \beta_1 \ln P(offbeer)_{it} + \beta_2 \ln P(offcider)_{it} + \beta_3 \ln P(offwine)_{it} + \beta_4 \ln P(offspirit)_{it} + \beta_5 \ln P(offRTD)_{it} + \beta_6 \ln P(onbeer)_{it} + \beta_7 \ln P(oncider)_{it} + \beta_8 \ln P(onwine)_{it} + \beta_9 \ln P(onspirit)_{it} + \beta_{10} \ln P(onRTD)_{it} + \beta_{11} \ln Income_{it} + \beta_{12} KID_{it} + \beta_{13} MRD_{it} + \beta_{14} UNE_{it} + \beta_{15} SMK_{it} + \beta_{16} Age^2_{it} + \gamma YearDummies + a_i + u_{it}$$

- **Model selection:**
 - Fixed effect models appear to be more appropriate than random effect models based on Hausman tests
 - Non-PPU/Income independent variables are jointly significant for majority of models tested
- **Results (base case fixed effect models):**

		Purchase									
		Off-beer	Off-cider	Off-wine	Off-spirits	Off-RTDs	On-beer	On-cider	On-wine	On-spirits	On-RTDs
Price	Off-beer	-0.980*	-0.189	0.096	-0.368	-1.092	-0.016	-0.050	0.253	0.030	0.503
	Off-cider	0.065	-1.268*	0.118	-0.122	-0.239	-0.053	0.093	0.067	-0.108	-0.194
	Off-wine	-0.040	0.736*	-0.384*	0.363	0.039	-0.245	-0.155	0.043	-0.186	0.110
	Off-spirits	0.113	-0.024	0.163	-0.082	-0.042	0.167	0.406	0.005	0.084	0.233
	Off-RTDs	-0.047	-0.159	-0.006	0.079	-0.585*	-0.061	0.067	0.068	-0.179*	0.093
	On-beer	0.148	-0.285	0.115	-0.028	0.803	-0.786*	0.867	1.042*	1.169*	-0.117
	On-cider	-0.100	0.071	0.043	0.021	0.365	0.035	-0.591*	0.072	0.237*	0.241
	On-wine	-0.197	0.094	-0.154	-0.031	-0.093	-0.276	-0.031	-0.871*	-0.021	-0.363
	On-spirits	0.019	-0.117	-0.027	-0.280	-0.145	-0.002	-0.284	0.109	-0.890*	0.809*
	On-RTDs	0.079	0.005	-0.085	-0.047	0.369	0.121	-0.394	-0.027	-0.071	-0.187

- **Own-price elasticities:**
 - All negative and 8 out of 10 are statistically significant (except for off-trade spirits and on-trade RTDs).
 - Range from -0.08 (off-trade spirits) to -1.27 (off-trade cider).
 - In the off-trade, apart from cider, beer being most elastic (-0.98), followed by RTDs (-0.59), wine (-0.38) and spirits (-0.08).
 - In the on-trade, spirits being most elastic (-0.89), followed by wine (-0.87), beer (-0.79), cider (-0.59) and RTDs (-0.19).
 - For wine and spirits, on-trade is more elastic than off-trade. The opposite for beer, cider and RTDs.

- Cross-price elasticities
 - Mix of positive and negative signs (46 vs 44).
 - Only 6 out of 90 were statistically significant, among which 5 out of 6 have positive signs.
 - Jointly significant for the demand of on-trade wine, spirit and beer.
 - Some level of substitution effect of on-trade demand with respect to off-trade prices (15 out of 25 have positive signs in the top right corner of the matrix).

- Estimated own-price elasticities are broadly in line with historic estimates, though most previous estimates have not separated off- vs on-trade, and/or cider and RTDs.
- Challenging to compare estimated cross-price elasticities due to lack of similar studies. Although mostly insignificant individually, they are potentially useful for 1) improving estimation of own-price elasticities, and 2) are jointly significant for the demand of several beverages.

- LCF collect purchasing not consumption data (issues regarding inventory behaviour, purchasing for others, etc.).
- Although the use of fixed effect model on pseudo-panel data could reduce the endogeneity problem (e.g., controlled for unobserved time-constant omitted variable), endogeneity is still a problem:
 - Simultaneity (drinkers choose demand and price simultaneously).
 - Measurement error (e.g., price faced vs price paid).
 - Omitting time-variant variables such as preference for brand and packaging.
- The estimated elasticities are less robust to be applied for policy evaluation when:
 - Only a single or small number of beverages are affected.
 - Price changes are substantial.
- **References**
- **Meng Y**, Brennan A, Purshouse R, Hill-McManus D, Angus C, et al. Estimation of own and cross price elasticities of alcohol demand in the UK – a pseudo-panel approach using the Living Costs and Food Survey 2001 to 2009. *Journal of Health Economics* 2014 34: 96-103.