



Empirical Project - Assignment 2 (week 2)

Determinants of health care decisions

Purpose

The purpose of this assignment is to familiarize students with the empirical application of instrumental variables methods. You will compare IV/TSLS estimators with OLS estimators and consider some of the subtleties involved when considering instrumental variables. To test for the necessity of using instruments and test the validity of the instruments. You will investigate an interesting economic question involving (adverse) selection and moral hazard. OLS seems inappropriate here, but will instrumental variables estimators solve all problems?

Report:

Write a report of maximally two pages (A4, times new roman 12 pnts or equivalent) a short analysis of the determinants of health care decisions. Use the common structure: Title, Authors, Abstract. (all of which can be on the front page) Then on the two subsequent pages: 1 Introduction. 2. Theoretical background and method. 3 Data. 4 Results and preferred empirical model. 5. Conclusion. 6. Appendix (this is in addition to the two pages and should contain all the background material and regressions, specification tests etc.

The questions posed are there to help you to think about problems involved. You should not put questions and answers in the appendix. The appendix should only contain information and details to back up claims you make in the report.

Submit a pdf version of this report before Friday 11 June 23:59 on Canvas. Also upload your Stata do file, which should run properly and contain some brief comments on what you have done. Do not forget the names and student numbers of all group members.

Assessment:

This assignment will be marked out of 10. It is worth 20% of the final mark for the course 'Empirical Project'.

Literature:

- Heij, C., Boer, P. de, Franses, P.H., Kloek, T. and Dijk, H.K. van, 2004, *Econometric Methods with Applications in Business and Economics*, Oxford University Press;
- Harmon, C., and Walker, I., 1995, Estimates of the economic return to schooling for the United Kingdom, *American Economic Review*, vol. 85, pp.1278-1286.
- Shen, C., 2013, Determinants of health care decisions: insurance, utilization, and expenditures, *Review of Economics and Statistics*, vol. 95, pp. 142-153.

Background:

Shen (2013) analyses the relation between health expenditure and health insurance. Being insured or not will depend on the expected health costs (adverse selection) which can depend on private information the insurance taker has, but not the insurance company (asymmetric information). Finally, the health care consumed and cost incurred will depend on the presence of the individuals health insurance (moral hazard). Given the endogeneity, OLS is inconsistent and the standard solution is to use instrumental variables and TSLS. Shen (2013) also solves a number of other problems including for instance the fact that there are so many observations with zero expenditure. The paper also uses semi-parametric methods (which do not make specific assumptions on the underlying disturbance term, as is done in ML) but ignores the zero observations. You will also revisit the issue of zero observations in this assignment.

Data:

- *Description:* Medical Expenditure Panel Survey (MEPS), cohort 2011 (HC-147 2011 Full Year Consolidated Data File), Pseudo panel of US household survey data since 1996. Every member of a household is being interviewed (parents answer for young children). The data have been restricted to individuals older than 18 because of missing information for younger individuals. The resulting number of observations is 25465. There are three measurements per year which sometimes gives additional information (see below)
- *Additional information:* http://meps.ahrq.gov/mepsweb/data_stats/data_overview.jsp and on Canvas.
- *Specific data for the assignment:* 10 separate STATA files are available. It is the last digit of the the first student named on the assignment (and names should be in alphabetical order). So if this is a 0 you use the 6528 observations in data2_0.dta.

Version	Selection
data2_0	$18 \leq \text{age} \leq 30$: 6528 observations
data2_1	$31 \leq \text{age} \leq 40$: 4635 observations
data2_2	$41 \leq \text{age} \leq 50$: 4509 observations
data2_3	$51 \leq \text{age} \leq 60$: 4270 observations
data2_4	$61 \leq \text{age}$: 5523 observations
data2_5	female == 1: 13608 observations
data2_6	female == 0: 11857 observations
data2_7	$0 \leq \text{adult_bmi} \leq 25$: 8710 observations
data2_8	adult_bmi > 25: 15592 observations
data2_9	geen selectie: 25465 observations

Information on variables:

Special codes: are used for certain variables:

- -1: question not applicable for this respondent;
- -2: question not posed because situation did not change since last survey.
- -7: respondent refuses to answer.
- -8: respondent does not know the answer
- -9: interviewer did not report the answer.

Available variables:

Variables are based on Shen (2013) but with some differences and additions.

- `adult_bmi` = `bmi`;
- `age` = age in years;
- `child_bmi` = `bmi` (reported while child). Despite the age restriction `child_bmi` can be useful when `adult_bmi` is unavailable;
- `civ_stat` = civil status: = 1 (married), 2 (widow/widower), 3 (formally divorced), 4 (separated), 5 (other), 6 (younger than 16);
- `duid` = number of the household (identifier);
- `ethnic_gr` = ethnisc group = 1 (white), 2 (black), 3 (native American), 4 (Asian), 5 original Hawaiian), 6 (other);
- `fam_size` = number of people in the household;
- `female` = dummy = 1 for women (0 for men);
- `flu_vac` = dummy = 1 if vaccinated against flu in 2011;
- `h_expend` = annual health expenditures in \$;
- `hisp` = dummy =1 for Hispanics (first language Spanish) (0 otherwise);
- `inc_fam` = total family income in \$;
- `inc_pers` = total personal income in \$;
- `insur` = health insurance = 1 (privately insured), 2 (public-insurance), 3 (not insured);
- `insur_fr` = fraction of the year (`#months/12`) that a person was health insured;
- `kessler`: the Kessler (K-6)-index. Sum of 6 variables indicating the mental health of the individual: individual feels (a) nervous (b) desperate (c) restless (d) depressed (e) anything is too much effort (f) useless. Each of these variables can take the values 0 (never), 1 (rarely) 2 (sometimes) 3 (often) 4 (all the time)

- mh_perc = Mental health average of the self reported mental health during the 3 interviews in 2011. Possible answers per interview: 1 (excellent), 2 (very good), 3 (good), 4 (reasonable) 5 (bad), -9 (If no answer is given at any stage during the three interviews);
- n_comorb = number of co-morbidities, the sum of 12 dummy variables indicating if a person suffers from a particular illness (some further information below. This variable differs slightly from ?);
- ph_perc = Physical health: average of the self reported physical health during the 3 interviews in 2011. Possible answers per interview: 1 (excellent), 2 (very good), 3 (good), 4 (reasonable) 5 (bad), -9 (If no answer is given at any stage during the three interviews);
- pid = person identifier within the household (101, 102, 103: person 1, person 2, person 3 ...);
- pregnant = dummy variable = 1 if woman is pregnant during one of the three interviews, =0 otherwise;
- p_weight = weight given to observation to render a representative sample vs the US population;
- region = 1 (Northeast), 2 (Midwest), 3 (South) or 4 (West);
- smoker = dummy variable = 1 for smokers, =0 for non-smokers;
- use_seatb = dummy variable = 1 if individual wears seatbelt in a car;
- y_educ = years of education/training;
- A number of variables with respect to profession: EMPST31, EMPST42, EMPST53, MORJOB31, MORJOB42, MORJOB53, SELFCM31, SELFCM42, SELFCM53, OCCAT31, OCCAT42, OCCAT53, INDCAT31, INDCAT42, INDCAT53. These variables are straight from the data source

Notes with respect to changed definitions of variables and differences from Shen (2013):

- General. Shen (2013) uses 2005 data. You have data available for 2011. Over time a number of changes appear to have been made in the survey. Moreover, the definitions in Shen (2013) are not always very clear, but we have tried to find comparable variables. We also provide some additional variables.
- Sample. Shen (2013) uses data of employed individuals with an age in the range 22-64 years without public health insurance and a relatively high BMI (>30). Your data only has the age > 17 restriction.

- `civ_stat` = civil Status: the definition in the data description (p. C-23) is not clear (except option 6 "Under 16 - Inapplicable"). The following information can be found in the survey questions: RE13OV ===== {(Are/Is)/On December 31, {YEAR}, (were/was)} (PERSON) {now} married, widowed, divorced, or separated? Possible answers: 1: MARRIED, 2: WIDOWED, 3: DIVORCED, 4: SEPARATED, -7: REF, -8: DK. The variable "Married" in Shen (2013) provides less information;
- `duid` and `pid`: Allow the identification of individuals from the same household.;
- `ethnic_gr` and `hisp`. Provide more detailed information than "Race" in Shen (2013);
- `income`: the definition of income in Shen (2013) is unclear. The available data here includes both personal income (`inc_pers`) as well as household income (`inc_fam`).
- `mental illness`: the information in Shen (2013) and the 2011data are insufficient to deduce this variable. We provide the alternative `kessler` and `mh_perc` variables;
- `n_comorb` = number of co-morbidities. Defined co-morbidities in the 2011 and 2005 surveys appear to be different. In 2011 individuals are asked about the diagnoses: (a) high blood pressure (b) Coronary artery disease (c) angina (d) have you suffered from a heart attack (e) other heart diseases (f) high cholesterol (g) brain haemorrhage (h) lung disease (COPD) (i) cancer (j) diabetes (k) arthritis (l) asthma. This has been put into 12 dummies : =1 if diagnosed, =0 if not diagnosed) and added together;
- Profession and sectorial variables: `EMPSTij`, `MORJOBij`, `SELFCMij`, `OCCCATij`, `INDCATij` with `ij`=31, 42 of 53 indicating the moment at which the question was asked.
 - `EMPSTij` = employment status respondent at moment of interview `ij`. Possible values 1 (currently employed), 2 (temporarily laid off), 3 (worked during reference period) and 4 (unemployed).
 - `MORJOBij` = having more than one job. Possible values 1 (yes) en 2 (no).
 - `SELFCMij` = being self employed. Possible values: 1 (yes) and 2 (no).
 - `OCCCATij` = occupational category. Possible values are explained in the file "occ3.pdf" available on Canvas.
 - `INDCATij` = industrial category (bedrijfstak). Possible values are explained in the file "ind3.pdf" available on Canvas. Note that variables "white collar" (see definition in Shen (2013, p. 147) and "Industry Insurance Rate" cannot be deduced from this information. You are expected to change the available variables on professions and sector in relevant and useful variables in your econometric analysis.
- Addition variables (relative to ?): `adult_bmi`, `child_bmi`, `insur_fr`, `flu_vac`, `ph_perc`, `pregnant`, `p_weight`, `use_seatb`.

Instruments:

Shen (2013) uses profession and sectorial variables as instruments. In addition `flu_vac` and `use_seatb` might be useful instruments. Start out with assuming all these variables are valid and relevant. In preparing the report this is something that needs to be determined.

Choice of variables:

Note the large number of available variables. Still, this is only a small proportion of the total number available (more than 2000 variables!). A number of variables give similar kinds of information but with some differences. You will have to take sensible decisions and make sensible choices to obtain the best model possible. There will not be a perfect model. You will have to justify your choices of variables based on economic theory (and common sense). Econometric tests, measures, and insights can also help. In any case you need to justify your choices.

Background questions and thoughts

Do NOT submit answers to the following questions. They are just there to help you do the analysis.

Just use them to think about the problem and the issues involved in analysing personal health expenditure.

You should analyse the personal health costs and the insurance decision.

1. Start with explaining healthcare expenditure. Look for the best possible model assuming all explanatory variables are exogenous. So estimate with OLS (why?)
 - (a) Determine on the basis of economic theory and the available data which variables you will include in the model. Make a distinction between (i) the dependent variable and (ii) the exogenous potential explanatory variables. Should insurance be in the model?
 - (b) The variable *insur* is actually not suitable for the analysis. Why? Create two dummy variables for private and public insurance against medical expenses. Estimate a model with both dummies.
 - (c) Is there a difference in effect between the two health insurance dummies? Are the estimated effects significant?
2. The insurance variable should be in the model on theoretical grounds, but is probably endogenous (why?). Assume that this is the only (possibly) endogenous explanatory variable.
 - (a) The above text provides a number of possible instruments for the endogenous explanatory variable health insurance. Which requirements must an instrument meet? Check theoretically whether the instruments listed above (could) comply with these conditions (you should also investigate this empirically, but we get back to this later).
 - (b) Give the descriptive statistics of the variables and instruments mentioned in (a) and the instruments (report: mean, standard deviation, minimum and maximum).
3. Hausman test for exogeneity.

- (a) Check the optimal (OLS based) specification for exogeneity (Hausman test). Do this for both health insurance variables.
 - (b) Check whether the instruments you intended have a direct impact on health costs
4. IV / 2SLS estimate.
- (a) Perform an IV / 2SLS estimation and evaluate the results
 - (b) Compare the estimation results with the OLS estimates. Are the the estimates very different? Are the standard errors very different?
5. Discuss the strength of the instruments used. The usual way is to see if the instruments are relevant is based on an F-test in the first stage regresion in the TSLS procedure (i.e. the regression on the endogenous explanatory variables). In practice, it is not really tested. The resulting F-statistic must be greater than 10 according to the commonly used rule of thumb. If that is the case, the instruments are strong. If the F-statistic is smaller than 10 then we have weak instruments. Harmon and Walker (1995) use this test to check the strength of the instruments.
6. The IV/TSLS restults are based on the exogeneity of the instruments.
- (a) Test the validity (exogeneity) of the instruments.
 - (b) If the instruments are really exogenous, then functions of them will also be exogenous and can be used as instruments. In principle you can create many instruments, but there must be a trade of. What happens to the strength of additional instruments? What happens if you use n instruments?
7. Shen (2013) argues that publicly insured individuals have no choice. What does this imply for the exogeneity of the publicly insured dummy? How different are the results when treating this group differently?
8. There are many observations with zero health expenditure (why?). This poses a problem for the basic linear regression model with continuously (normally) distributed disturbances. Explain why. The Probit/Logit and Tobit models are especially designed for these kinds of situation. Look the models up in e.g. Heij et al. (2004). Assume that all variables that determine the insurance decision are exogenous (which is restrictive, but keeps the empirical analysis doable).
- (a) Estimate the Logit and Probit models.
 - (b) Estimate the Tobit model .Why are the estimates obtained by iteration?
 - (c) Compare the marginal effect of the Tobit model with your earlier OLS/IV estimates.