Empirical Case Studies

Christoph Schottmüller

1 Selection in long term care insurance

The dataset "FMcGdata.csv" contains data about an elderly sample (average age 78) from the US. The dataset contains whether these elderly entered a nursing home in the period 1995-2000 and whether they had long term care insurance in 1995. A survey in 1995 asked

- "Of course nobody wants to go to a nursing home, but sometimes it becomes necessary. What do you think are the chances that you will move to a nursing home in the next five years?" (answer in the variable "belief")

The following table gives an overview over the variables in the dataset:

variable name	description
dEnterNH	binary variable that is 1 if individual went to nursing home between 1995 and 2000
belief	self reported belief of entering nursing home between 95 and 00 (elicited in 95)
insPrediction	prediction of insurance company that person enters nursing home between 95 to 00
ltcinsd	binary variable which is 1 if person has long term care insurance in 95
asset1	binary, 1 if person is in top wealth quartile
asset2	wealth quartile 2
asset3	wealth quartile 3
prevent	percentage of gender specific preventive activities undertaken (between 0 and 1)
seatbelt	use of seatbelt (values 1-4 for "all or most", "sometimes", "rarely", "never")
dSeatbelt	binary variable that is 1 if seatbelt is used "all or most" times, 0 else
inc1	binary, 1 if person is in top income quartile

1. What proportion of the elderly entered a nursing home between 1995 and 2000?
2. What is the average belief that a person enters nursing home? What is the average prediction of the insurance company?
3. What proportion of the elderly had long term care insurance in 1995?
4. Are those with a higher belief of entering a nursing home more likely to have long term care insurance?
5. Are people with a higher belief in fact more likely to enter a nursing home?
6. Do people have private information about their probability of entering a nursing home? (Hint: that is, does the belief contain information about the likelihood to enter a nursing home beyond the prediction of the insurance company?)
7. Is there adverse selection, i.e. are those that buy long term care insurance more likely to enter a nursing home?
8. The variables "prevent" and "dSeatbelt" can be viewed as measures of risk aversion (admittedly in other domains). How is risk aversion (measured by these variables) correlated with the purchasing decision of long term care insurance and the probability of entering a nursing home?
9. How is wealth correlated with the purchasing decision of long term care insurance and the probability of entering a nursing home?

2 Minimum deductible in the Netherlands

In the lecture we discussed some evidence for moral hazard. The main question here is whether we can find some empirical evidence for moral hazard ourselves.[1]

In the Netherlands health insurance is provided by a handful semi-private health insurers. The base coverage is fixed by law and the law also mandates, since 2008 , a minimum deductible for all insured *above age 18*, i.e. there is no deductible or other copayment for kids. Copayments that do not take the form of a deductible are not used and most insured have a deductible equal to the legal minimum. The minimum deductible has been increased over time from the initial 150€ per year. (2009: € 155; 2010: € 165; 2011: € 170; 2012: € 220; 2013: € 350; 2014: € 360; 2015: € 375; 2016, 2017 en 2018: € 385). Cost data for the Netherlands is available on `http://www.vektis.nl/index.php/vektis-open-data` where for each (age,gender,postcode) triple you can find the total health care costs split up into different categories. On the course website I provide a simplified version of the data sets from 2011 and 2014 in which I changed the variable names to English and aggregated all the costs that fall under the deductible into one variable.

You can do the exercises below in a spreadsheet app (like Microsoft Excel or *OpenOffice Calc*) but even better suited would be a statistics software (like *R* or Stata or SPSS) or a data analysis package in a general pupose programming language (like *Pandas for Python* or *DataFrames for Julia*). (options in italics are free, open source and available for all common operating systems; as you might have guessed, I like Julia best)

1. Download the data for 2011 and open it in your software of choice. Do you understand what the number in the different cells mean?
2. Can you find out how many people had health insurance in the Netherlands in 2011?
3. Can you make a plot with age on the x-axis and average costs under the deductible on the y-axis? (If you are not familiar with the software this might be tricky and you might want to proceed without it.)
4. Can you find the average costs under the deductible of 17 year olds?
5. Can you find the average costs under the deductible of 19 year olds?
6. How do you interpret the difference in average costs between 17 and 19 year old?
7. To get a better idea of the difference plot the distribution of costs (this is called a "histogram") for 17 and 19 year olds. (again this can be a bit tricky)
8. Why would it make sense to repeat some of the analysis with the 2014 data?
9. Can you give a demand elasticity for the deductible, i.e. if we increase the deductible by 100% by how much do expenditures decrease?
10. Can the estimate of the previous exercise be compared to the famous -0.2 demand elasticity from the RAND health insurance experiment?

3 Hospitals in Germany

Check the dataset *hospitalBirths.csv*. The data is from 2018 and comes from mandatory quality reports hospitals have to submit on an annual basis. Population numbers and city locations are taken from the official city level reports of the "Bundesamt für Kartographie und Geodäsie". Hospital locations were determined using *openstreetmaps.org*. The following table explains the important variables.

[1] This case study is based on material prepared by Jan Boone, see section "Regulation in health care markets" here.

variable name	description
institutionenkennzeichen	hospital id
Standortnr	location id (only relevant for hospitals with several locations)
betten	number of hospital beds
privat	dummy: 1 if hospital is private and 0 else
freigemein	dummy: 1 if hospital is a non-profit (usually owned by a religious organization)
births	number of births
caesarean	number of Caesarean sections
nCompBirths	number of birth stations within a 50km radius (excluding the hospital itself and hospitals with the same "insitutionenkennzeichen")
nBirthsInRadius	number of births within a 50km radius (as the crow flies)
demandPotMax	approximate number of people living within a 50km radius

In the following we want to look at the following question: Could there be demand inducement by hospitals in the sense of pushing for more Caesarean sections among all births.

1. Construct a new variable "shareC" that contains the share of Caesarean sections among all births.
2. Construct a variable "density" that contains a measure of birth station density. (hint: use the variables *nCompBirths* and *demandPotMax*)
3. Do you expect *density* and *shareC* to be correlated? Check in the data whether your expectation is met! (Can you use a plot to get a better idea of the correlation? Are there certain hospitals that you would like to exclude from the analysis?)
4. Would you expect stronger or weaker correlations between *density* and *shareC* for hospitals that are (i) public, (ii) private, (iii) non-profits? Check in the data whether your expectation is met! (Extra exercise: check how *shareC* differs across ownership types! Any explanations for these differences?)
5. Is this data set suitable to check for supplier induced demand? Why (not)?

Finally, have a brief look at the file "reimbursement$_{\text{OPS5377fluctuation.csv}}$". It contains the reimbursement of a hospital for a 77 year old man getting a pacemaker/defilibrator (OPS code 5.377.1, diagnosis ICD I44.1, DRG: F12I), staying 5 days in hospital in several years. (The federal base rate is taken from here. The grouping was done using this online tool.) The case numbers are all men aged between 75 and 80 with OPS 5.377 according to `https://www-genesis.destatis.de`. Plot the reimbursement (or caseweights) over time and do the same with the case numbers. How can such data be used to check for supplier induced demand?