

MICRO DATA ANALYSIS

CLASS 2

Wiktor Budziński

UPDATED PLAN

Class 1: Brief introduction to R, OLS, building econometric models

Class 2: Finishing OLS (heteroskedasticity), binary and ordered choice

Class 3: Multinomial choice

Class 4: Count data

Class 5: Endogeneity (2SLS and other methods for non-linear models)

BINARY DATA

Many variables take a form of a binary response

- Yes or No answers
- Voting for one of the two candidates
- Buying a product or not

The linear regression model is not well-suited for such data as it assumes continuous distribution rather than a discrete one

- Also, the predictions from the model can be outside of the $[0,1]$ interval
- Binary data are also heteroskedastic by nature

BINARY DATA

Models for binary data are usually formulated in one of the three ways:

1. Using an indicator variable: $y^* = \mathbf{X}\boldsymbol{\beta} + \varepsilon$
where $y = \begin{cases} 1 & \text{if } y^* > 0 \\ 0 & \text{if } y^* \leq 0 \end{cases}$
2. Using a generalized linear model formulation: $E(y | \mathbf{X}) = f(\mathbf{X}\boldsymbol{\beta})$
where $f()$ is a 'link' function such that $f : \mathbb{R} \rightarrow [0,1]$
3. Using a random utility specification
 - We will cover that in the next class with multinomial variables

BINARY DATA

Different models arise depending on the assumption about error term distribution in the indicator function specification or the link function in the GLM setting

Usual model forms:

- Logit: $P(y = 1 | \mathbf{X}\boldsymbol{\beta}) = \frac{\exp(\mathbf{X}\boldsymbol{\beta})}{1 + \exp(\mathbf{X}\boldsymbol{\beta})}$
- Probit: $P(y = 1 | \mathbf{X}\boldsymbol{\beta}) = \Phi(\mathbf{X}\boldsymbol{\beta})$
- Complementary log-log: $P(y = 1 | \mathbf{X}\boldsymbol{\beta}) = 1 - \exp(-\exp(\mathbf{X}\boldsymbol{\beta}))$

CONTINGENT VALUATION

One of elicitation formats in non-market valuation

- The aim is usually to learn the value that consumers put on some policy program
- Sometimes also used in marketing

It basically asks respondents directly how much they would be willing to pay for the program

Question can be framed in a few different ways:

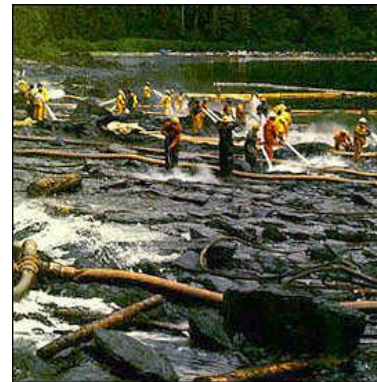
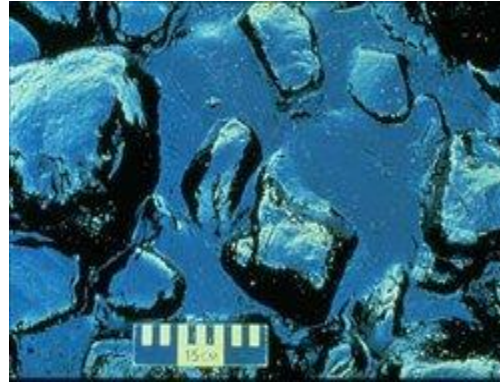
- How much are you willing to pay?
- Would you be willing to pay X?
- Indicate maximum value you would be willing to pay (so called payment card)

Question format will determine modelling strategy

EXXON VALDEZ

Giant oil spill at Prince William Sound

- Although environmental damage was huge, no one was harmed
- Exxon was sued by environmental organizations
 - Environmental damage was estimated to be around 3-15 billions USD
- Contingent valuation methods were used for estimation



CONTINGENT VALUATION

EXXON's accident forced a debate regarding the use of CVM

- Conclusion was that it can be used reliably if some conditions are fulfilled

Since then non-market valuation methods were used widely in cost benefit analyses and policy making

- US Clear Water Act
- Clear Air for Europe

Used in variety of settings: environmental, health, transportation...

EXERCISE 1: BINARY DATA

1. Analyze simulation example in *Sim_examples2.r*
2. Read *oil1.rds* data into R
 - CVM data regarding Shell oil spill in San Francisco Bay (1988)
 - The study was conducted few years later, and it was about governmental program of prevention and mitigation of damages from oil spills in the future
 - Respondents were informed how the program would look and how much would it cost them in increased taxes
 - Respondents could vote for the program or against
3. Estimate basic logit model
4. Interpret the results with marginal effects
5. Calculate willingness to pay for this program

HETEROSCEDASTICITY

As usually in a microdata binary variable are often characterized by heteroskedastic error terms

- This can lead to biased estimates

There is no test for it except for estimating the model with heteroscedasticity

Index function equation becomes: $y_i^* = \mathbf{X}_i\boldsymbol{\beta} + \sigma(\mathbf{Z}_i\boldsymbol{\gamma})\varepsilon_i$

- Where $\sigma(\mathbf{Z}_i\boldsymbol{\gamma}) = \exp(\mathbf{Z}_i\boldsymbol{\gamma})$

6. Estimate logit/probit models with heteroscedasticity

WORKBOOK 2

Now try to conduct a similar analysis for the exercises in Workbook2.R

- Exercise 1

ORDERED DATA

Often a discrete variable is ordered, even if its values does not have any absolute interpretation

- It often represents a consumer choice on a given scale

Examples include:

- Consumers rating a product by giving stars
- Respondents rating some statement on a Likert scale
 - I definitely agree, I rather agree, I neither agree nor disagree, I rather disagree, I definitely disagree

ORDERED DATA

As the levels of the variable do not have an absolute interpretation, and the support is usually finite, count data models should not be used for such variables

Usually the model is described in terms of the index function: $y_i^* = \mathbf{x}_i \boldsymbol{\beta} + \varepsilon_i$

- y_i^* is unobserved to us, used as an index to construct a model
- Epsilon is usually either normally (probit) or logistically (logit) distributed, with variance normalized to 1

For the variable for which we observe J different values we will additionally estimate $J-1$ threshold parameters

- We will assume that order variable takes a given value if an index variable is between two thresholds:

$$y_i = j \quad \text{for} \quad \alpha_{j-1} < y_i^* < \alpha_j$$

ORDERED DATA

$$y_i = 1 \quad \text{dla} \quad y_i^* \leq \alpha_1$$

We will observe: $y_i = 2 \quad \text{dla} \quad \alpha_1 < y_i^* \leq \alpha_2$

...

$$y_i = J \quad \text{dla} \quad y_i^* > \alpha_{J-1}$$



Usually the model is estimated without a constant

ORDERED DATA

Likelihood can then be easily calculated as:

$$P(y_i = 1 | \mathbf{X}_i) = F(\alpha_1 - \mathbf{X}_i \boldsymbol{\beta})$$

$$P(y_i = 2 | \mathbf{X}_i) = F(\alpha_2 - \mathbf{X}_i \boldsymbol{\beta}) - F(\alpha_1 - \mathbf{X}_i \boldsymbol{\beta})$$

...

$$P(y_i = J | \mathbf{X}_i) = 1 - F(\alpha_{J-1} - \mathbf{X}_i \boldsymbol{\beta})$$

Thresholds need to be positive and increasing: $\alpha_1 < \alpha_2 < \dots < \alpha_{J-1}$

EXERCISE 2: ORDERED DATA

1. Analyze what covariates can explain how much people are worried about the environmental status of Baltic Sea (*envw*)
2. Interpret the results using marginal effects

ORDERED DATA

As usually in a microdata ordered variable are often characterized by heteroskedastic error terms

- This can lead to biased estimates

There is no test for it except for estimating the model with heteroscedasticity

Index function equation becomes: $y_i^* = \mathbf{x}_i \boldsymbol{\beta} + \sigma(\mathbf{z}_i \boldsymbol{\gamma}) \varepsilon_i$

- Where $\sigma(\mathbf{z}_i \boldsymbol{\gamma}) = \exp(\mathbf{z}_i \boldsymbol{\gamma})$

EXERCISE 2: ORDERED DATA

3. Estimate ordered model with heteroscedasticity

WORKBOOK 2

Now try to conduct a similar analysis for the exercises in Workbook2.R

- Exercise 2