MICROECONOMETRICS
CLASS 3

Wiktor Budziński
Marek Giergiczny

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 not 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

We will observe:

$$y_i = 1 \quad \text{dla} \quad y_i^* \le \alpha_1$$
$$y_i = 2 \quad \text{dla} \quad \alpha_1 < y_i^* \le \alpha_2$$
$$\ldots$$
$$y_i = J \quad \text{dla} \quad y_i^* > \alpha_{J-1}$$

Usually the model is estimates without a constant

ORDERED DATA

Likelihood can then be easily calculated as:

$$P\left(y_i = 1 | \mathbf{X}_i\right) = F\left(\alpha_1 - \mathbf{X\beta}\right)$$
$$P\left(y_i = 2 | \mathbf{X}_i\right) = F\left(\alpha_2 - \mathbf{X\beta}\right) - F\left(\alpha_1 - \mathbf{X\beta}\right)$$
$$\ldots$$
$$P\left(y_i = J | \mathbf{X}_i\right) = 1 - F\left(\alpha_{J-1} - \mathbf{X\beta}\right)$$

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

EXERCISE 1: 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: $y_i^* = \mathbf{X}_i\boldsymbol{\beta} + \sigma(\mathbf{Z}_i\gamma)\varepsilon_i$

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

EXERCISE 2: ORDERED DATA

1. Estimate ordered model with heteroscedasticity

WORKBOOK 3

Now try to conduct a similar analysis for the exercises in Workbook3.R

- Exercise 1

BINARY DATA

Binary data arise as a special case of an ordered variable with only two levels

- Because of that only one threshold coefficient will be estimated
- For binary data threshold is usually fixed at 0, and instead a constant is estimated

Usual model forms:

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

Binary models can be interpreted as a random utility models

- More about it in the next class

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:

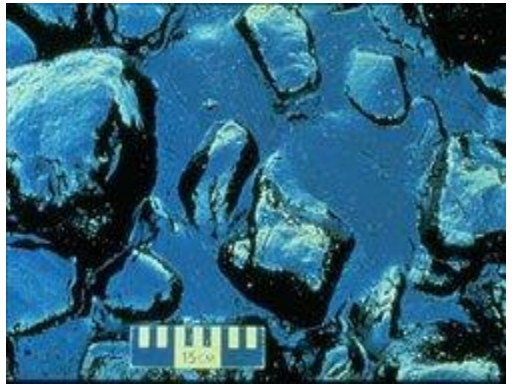
- How much are you willing to pay?
- Would you be willing to pay X?
- Indicate maximum value you would be willing to pay (so called payment card)

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 3: BINARY DATA

1. Analyze simulation example in *Sim_examples3.r*
2. Read *oil1.rds* data into R
   1. CVM data regarding Shell oil spil in San Francisco Bay (1988)
   2. 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
   3. Respondents were informed how the program would look and how much would it cost them in increased taxes
   4. Respondents could vote for the program or against
3. Estimate basic logit model and compare it with ordered logit
4. Interpret the results with marginal effects
5. Calculate willingness to pay for this program

COUNT DATA

In count data covariate is some integer value (0, 1, 2, …), which usually represent a number of something, for example

- No. of visits to a doctor
- No. of visits in a National Park
- No. of crimes
- No. of kids

Usually have to be defined over some period of time

- For example, in the last 12 months

COUNT DATA

To account for the characteristics of this data some specific probabilistic distribution needs to be assumed, for example Poisson

$$P(y_i|\lambda_i) = \frac{\lambda_i^{y_i}\exp(-\lambda_i)}{\Gamma(y_i - 1)}$$

Where $\lambda_i = \exp(\mathbf{X}_i\boldsymbol{\beta})$

- It of course relies on the assumptions of Poisson distribution, namely that mean of the distribution is equal to its' variance

TRAVEL COST METHOD

One of the applications of this kind of model is so called Travel Cost Method

- Used in environmental economics to estimate recreational welfare from, for example, National Parks

It basically relies on estimation of the demand function for number of visits to the given site (e.g. National Park)

Travel costs incurred by the individual is considered to be a price of the good

- For example, how much someone spend on petrol or train tickets to get there
- Usually the cost of time is also included

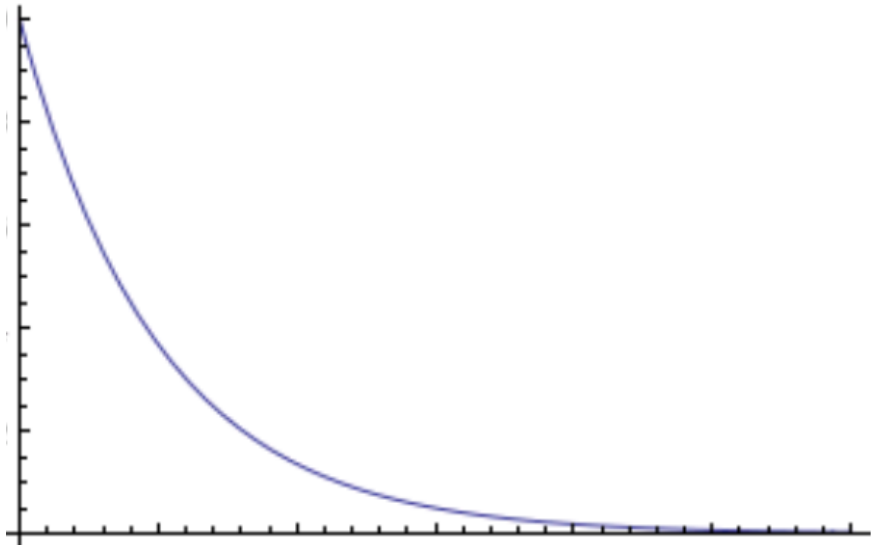
Having demand function estimated one can calculate consumer surplus as a welfare measure

TRAVEL COST METHOD

For example, using Poisson model we estimate demand as $\mathrm{E}\left(y_i \mid \mathbf{X}_i\right) = \lambda_i = \exp\left(\mathbf{X}_i\boldsymbol{\beta} - \alpha TC_i\right)$

Consumer surplus is given then by: $\mathrm{CS} = \int_{TC_i}^{+\infty} \exp\left(\mathbf{X}_i\boldsymbol{\beta} - \alpha y\right) dy = \frac{\lambda_i}{\alpha}$

If we divide it by lambda, we obtain CS per trip



EXERCISE 5: POISSON MODEL

1. Read *me.baltic.rds* into R
    1. International study regarding recreation at the Baltic Sea
2. Estimate demand function with Poisson regression
3. Calculate CS per trip as well as average CS for each country

POISSON DIAGNOSTICS

In order to analyze the functional form of the model, one can conduct test RESET

- In models estimated with MLE, this is usually called a LINKTEST
- Furthermore one can use quantile residuals and plot them against independent variables

Equidispersion is unique feature of Poisson model in which mean of distribution is equal to its' variance

- It is often not true in data

POISSON DIAGNOSTICS

Analogous test to Breush-Pagan test can be conducted to check for equidispersion

If the distribution is equidispersed then Z statistics should be normally distributed and be independent from lambda

$$z_i = \frac{\left(y_i - \hat{\lambda}_i\right)^2 - y_i}{\hat{\lambda}_i}$$

Auxiliary regression is calculated in which Z is explained by lambda

- If lambda is found to be significant then this is an evidence for overdispersion

EXERCISE 6: DIAGNOSTICS

1. Conduct LINKTEST to test functional form of the analyzed model
2. Predict quantile residuals and plot them against travel cost variable
3. Test for equidispersion
   - Analyze the second simulation example in *Sim_examples3.r*

NEGATIVE BINOMIAL MODEL

If we identify overdispersion in the data then some other model than Poisson should be applied

- Negative binomial regression is usually a first choice

NB can be thought of as a Poisson regression with additional random term

$$\mathrm{E}\left(y_i \mid \mathbf{X}_i, u_i\right) = \lambda_i u_i = \exp\left(\mathbf{X}_i\boldsymbol{\beta} + \log(u_i)\right)$$

If we assume that *u* has Gamma distribution then conditional probability can be calculated as:

$$P\left(y_i \mid \lambda_i, \theta\right) = \frac{\theta^{\theta}\lambda_i^{y_i}\Gamma\left(\theta + y_i\right)}{\Gamma\left(y_i + 1\right)\Gamma\left(\theta\right)\left(\lambda_i + \theta\right)^{\theta + y_i}}$$

NEGATIVE BINOMIAL MODEL

Theta is an additional parameter to estimate which controls for overdispersion, variance of NB distribution is given by:

$$\lambda_i \left(1 + \left(1/\theta\right)\lambda_i\right)$$

Testing $1/\theta = 0$ could be done to test for overdispersion

- It is usually easier to conduct LR test

HYPOTHESIS TESTING IN MLE

Wald test:

- Allow to test nonlinear hypothesis, H0: $\mathbf{c}(\boldsymbol{\theta}) = \mathbf{q}$
- Test statistics is given by: $W = \left(\mathbf{c}\left(\hat{\boldsymbol{\theta}}\right) - \mathbf{q}\right)' \left(AVC\left(\mathbf{c}\left(\hat{\boldsymbol{\theta}}\right) - \mathbf{q}\right)\right)^{-1} \left(\mathbf{c}\left(\hat{\boldsymbol{\theta}}\right) - \mathbf{q}\right) \sim \chi^2_J$
- Intuitively: if H0 does not hold, then $\mathbf{c}\left(\hat{\boldsymbol{\theta}}\right) - \mathbf{q}$ will be large, and the statistic will be larger than critical value

Likelihood Ratio test:

- Used to compare two models. One of the models should be nested in the other.
- Nested model is equivalent to the other model, but with some restrictions
- H0: restrictions are valid
- Test statistics (requires estimation of both models): $LR = -2\left(\ln \hat{L}_R - \ln \hat{L}_U\right)$

EXERCISE 7: NB MODEL

1. Estimate Negative Binomial regression
2. Compare model's fit with Poisson regression

ZERO-INFLATED MODELS

Another frequent issue in count data is an excess number of observations with 0 values

- Often not accounted for by regular distributions

Zero-inflated models account for that by *inflating* probability of zero response

- Usually interpreted as having two types of consumers: on the market who just had 0 demand, and out of the market
- This segmentation is latent, and not observed by researcher

ZERO-INFLATED MODELS

In such model 0 can arise from two reasons, and therefore probability function looks in the following way:

$$P\left(Y=y_i \mid \mathbf{X}_i, \mathbf{Z}_i\right)=\begin{cases} p_i\left(\mathbf{Z}_i\right)+\left(1-p_i\left(\mathbf{Z}_i\right)\right)F\left(y_i \mid \mathbf{X}_i\right) & if \quad y_i=0 \\ \left(1-p_i\left(\mathbf{Z}_i\right)\right)F\left(y_i \mid \mathbf{X}_i\right) & if \quad y_i \neq 0 \end{cases}$$

$p_i\left(\mathbf{Z}_i\right)$ is probability of being out-of-market, usually modelled as a logistic function of some covariates, whereas $F\left(y_i \mid \mathbf{X}_i\right)$ is probability of the number of events for someone in-the-market

- Both equations are estimated jointly

VOUNG TEST

Usually we compare nested models which can be easily done with LR test

- One model is nested in the other if upon fixing some coefficients, the model becomes equivalent to the other model

To test between non-nested models one can use Voung test $V = \sqrt{N}\,\frac{\mathbf{E}(m)}{std(m)}$

- Where $m_i = \log(L_i^1) - \log(L_i^2)$ is a difference in log-likelihoods between two non-nested models

It should be normally distributed if both models are equally far away from the true DGP

One could use some additional penalty within a model, similarly as in AIC and BIC

Usually used to compare Zero-inflated models with regular Poisson model

EXERCISE 8: ZERO-INFLATED MODELS

1. Estimate zero-inflated models
2. Compare them with a regular Poisson and NB models
3. Use simulation to check whether Voung test can be used for testing for zero inflation

HURDLE MODELS

Another solution to the issue of the excess of zeroes are so called hurdle models

They treat the dependent variable as a two step process

- First individuals decides whether they have positive demand or not
- If they decided that they have positive demand, they choose actual number of events

First step is modelled as a binary process (usually logit), second as a count data process truncated at 0

- Two models can be estimated separately

Likelihood is given by:

$$P(Y=y_i|\mathbf{X}_i,\mathbf{Z}_i)=\begin{cases}1-p_i(\mathbf{Z}_i) & if \quad y_i=0\\ p_i(\mathbf{Z}_i)\dfrac{F(y_i|\mathbf{X}_i)}{1-F(0|\mathbf{X}_i)} & if \quad y_i\neq 0\end{cases}$$

EXERCISE 9: HURDLE MODELS

1. Estimate hurdle models
2. Compare the interpretations of additional equations between Zero-inflated models and hurdle models
3. Compare model's fit to data with Zero-inflated models

WORKBOOK 3

Now try to conduct a similar analysis for the exercises in Workbook3.R

- Exercise 2