MICROECONOMETRICS
CLASS 2

Wiktor Budziński
Marek Giergiczny

GENERALIZED LINEAR MODELS

GLMs are based on Exponential dispersion model for which general formula for probability is given by:

$$f\left(x \mid \mu, \sigma^2\right) = h\left(x, \sigma^2\right) \exp\left(-\frac{d(x,\mu)}{2\sigma^2}\right)$$

μ is a mean of the distribution

$d(x,\mu)$ is called a unit deviance

- For example, for normal distribution: $d(x,\mu) = \left(x-\mu\right)^2$

$h\left(x,\sigma^2\right)$ is a normalizing term

- For example, for normal distribution: $h\left(x,\sigma^2\right) = \left(2\pi\sigma^2\right)^{-0.5}$

GENERALIZED LINEAR MODELS

Examples of the distributions from the exponential dispersion model include: inverse-gaussian, gamma, binomial and poisson

Similarly, as in linear regression we assume that mean of the distribution is some linear function of the covariates $\mu_i = \mu(\mathbf{X}_i\boldsymbol{\beta})$

- This is usually called a "link function"
- If the dependent variable is positive, we will usually use logarithm as a link function

EXERCISE 1: GLM

1. Continue to work on the *wine.xls* data, with a functional form from the previous class
2. Estimate GLM's with gamma and inverse-Gaussian distributions
3. Compare their model fit to the regression model we tried before
4. Analyze first simulation example in *Sim_examples2.R*

EXERCISE 1: GLM

1. Continue to work on the *wine.xls* data, with a functional form from the previous class
2. Estimate GLM's with gamma and inverse-Gaussian distributions
3. Compare their model fit to the regression model we tried before
4. Analyze first simulation example in *Sim_examples2.R*

Let's go to 

DIAGNOSTICS IN GLMS

In linear regression most assumptions were tested using residuals from the model

In GLMs it not exactly clear how to define residuals

- Response residuals are defined as $e_i^R = y_i - \mu_i$
  - In GLMs this is less useful as variance of the distribution is usually a function of the fitted values
- Person's residuals are given by $e_i^P = \frac{y_i - \mu_i}{Var(\mu_i)}$
  - Account for non-constant variance
- Deviance residuals are defined as $e_i^D = sign(y_i - \mu_i)\sqrt{d(y_i, \mu_i)}$
- Finally, quantile residuals can be calculated as $e_i^Q = \Phi^{-1}\left(F\left(y_i \mid \mu_i, \sigma^2\right)\right)$
  - It could be shown that they are normally distributed, the other are only approximately normal, and even that is not always true
  - If the dependent variable is discrete then quantile residuals employ additional randomization to make it more smooth

EXERCISE 2: GLM

1. Predict different types of residuals for the GLMs
2. Check whether quantile residuals are normally distributed
3. Plot residuals against fitted values and covariates
4. Calculate influence measures

EXERCISE 2: GLM

1. Predict different types of residuals for the GLMs
2. Check whether quantile residuals are normally distributed
3. Plot residuals against fitted values and covariates
4. Calculate influence measures

Let's go to 

QUANTILE REGRESSION

In GLMs we modelled conditional mean of the distribution $E\left(y_i \mid \mathbf{X}_i\right) = \mu\left(\mathbf{X}_i\right)$

We also showed that for Gaussian distribution you can also specify the functional form of the conditional variance: $\mathrm{var}\left(y_i \mid \mathbf{Z}_i\right) = \sigma^2\left(\mathbf{Z}_i\right)$

- Sometimes distribution could be more complex and these two moments may not describe the full distribution

Quantile regression is a semiparametric method which allow us to model separately each quantile of the distribution

QUANTILE REGRESSION

As a remainder: quantile τ is such a value λ_τ , that $F\left(\lambda_\tau\right)=\tau$

In Quantile regression we define conditional quantiles: $Q_\tau\left(\mathrm{y}_i \mid \mathbf{X}_i\right)=F_{y_i}^{-1}\left(\tau \mid \mathbf{X}_i\right)$

Usually similar specification as for linear regression: $Q_\tau\left(\mathrm{y}_i \mid \mathbf{X}_i\right)=\mathbf{X}_i \boldsymbol{\beta}_\tau$

- Different coefficients for different quantiles

QUANTILE REGRESSION

The simplest case is a median regression, for which we will minimize the following objective function:

$$Q\left(\boldsymbol{\beta}_{0,5}\right) = 0,5\sum_{i=1}^{N}\left|y_i - \mathbf{X}_i\boldsymbol{\beta}_{0,5}\right|$$

In more general case:

$$Q\left(\boldsymbol{\beta}_{\tau}\right) = \sum_{i:y_i \geq \mathbf{X}_i\boldsymbol{\beta}_{\tau}} \tau\left|y_i - \mathbf{X}_i\boldsymbol{\beta}_{\tau}\right| + \sum_{i:y_i \leq \mathbf{X}_i\boldsymbol{\beta}_{\tau}} \left(1-\tau\right)\left|y_i - \mathbf{X}_i\boldsymbol{\beta}_{\tau}\right|$$

This is sometimes called Least Absolute Deviations estimator

EXERCISE 3: QUANTILE REGRESSION

1. Estimate quantile regression for 0.25, 0.5 and 0.75
2. Compare coefficients across different quantiles and with OLS

EXERCISE 3: QUANTILE REGRESSION

1. Estimate quantile regression for 0.25, 0.5 and 0.75
2. Compare coefficients across different quantiles and with OLS

Let's go to 

QUANTILE REGRESSION

Advantages:

- No assumptions for the distribution of the dependent variable
- More robust to outliers than KMRL
- It can imply heteroscedasticity in the data
- Quantile of the transformed variable is equal to the transformed quantile of the original variable
  - If the transformation is increasing
  - Not true for mean, for example

$$Q_\tau\left(g\left(y_i\right)|\mathbf{X}_i\right)=g\left(Q_\tau\left(\mathrm{y}_i|\mathbf{X}_i\right)\right)$$

- Disadvantages: standard errors are either approximated or have to be simulated with bootstrap

EXERCISE 4: QUANTILE REGRESSION

1. See how heteroscedasticity affects results from QR in *Sim_examples2.r*
2. Estimate QR for quantiles: 0.1, 0.25, 0.5, 0.75 and 0.9
    1. Test whether coefficients differ between quantiles
3. Plot coefficients for different quantiles to analyze differences visually

WORKBOOK 2

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

- Exercises 1 and 2

Let's go to 

EXERCISE 4: QUANTILE REGRESSION

1. See how heteroscedasticity affects results from QR in *Sim_examples2.r*
2. Estimate QR for quantiles: 0.1, 0.25, 0.5, 0.75 and 0.9
    1. Test whether coefficients differ between quantiles
3. Plot coefficients for different quantiles to analyze differences visually

Let's go to 

ZERO RESPONSES

Very often the dependent variable is continuous but can also contain 0 values

- Can correspond to some solutions of consumer optimization problem, for example, expenditures

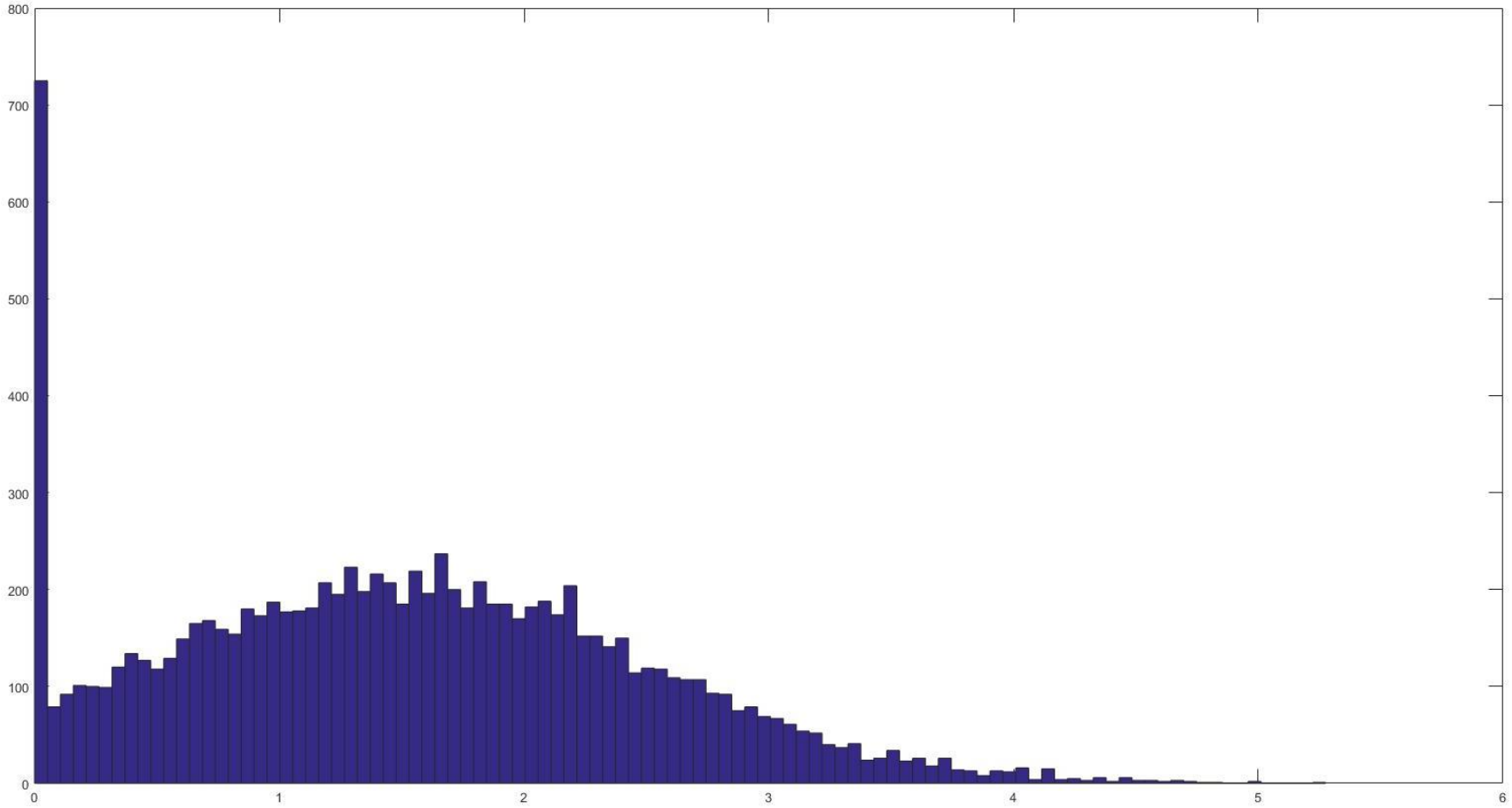
As most of responses has positive values we would like to take logarithm of it

- Because of 0 responses we cannot do that

Possible solutions:

- Assume linear relationship $y_i = \mathbf{X}_i\boldsymbol{\beta} + \varepsilon_i$
- Add small constant to the variable and use logs: $\log(y_i + \Delta) = \mathbf{X}_i\boldsymbol{\beta} + \varepsilon_i$
- Use Poisson regression
- Use Tobit model
- Use two step models (for example Heckman selection)

ZERO RESPONSES



POISSON

Poisson probability is given by the formula: $P(y_i \mid \lambda_i) = \frac{\lambda_i^{y_i} \exp(-\lambda_i)}{\Gamma(y_i - 1)}$

- It is generally for dependent variables that are integers
- Can be fitted to the positive continuous data nonetheless
  - It is not a valid probabilistic model though, log-likelihood does not make sense
  - It is often called a pseudo-maximum likelihood estimator
- It is well defined for zero values of the dependent variable
- Usually based on logarithmic link function: $\lambda_i = \exp(\mathbf{X}_i\boldsymbol{\beta})$
  - Elasticity can be easily calculated

TOBIT

Can also be called a censored regression model

Data are assumed to be normally distributed with censoring at 0

$$\begin{cases} y = y^* & y^* > 0 \\ y = 0 & y^* \leq 0' \end{cases} \quad y^* = \mathbf{X\beta} + \varepsilon, \quad \varepsilon \sim N(0, \sigma)$$

y^* corresponds to 'uncensored' variable

Another way to look at it: $y = \max\{\mathbf{X\beta} + \varepsilon, 0\}$

Estimated with Maximum Likelihood Method:

$$L_i = \begin{cases} \frac{1}{\sigma}\phi\left(\frac{y_i - \mathbf{X}_i\boldsymbol{\beta}}{\sigma}\right) & y_i > 0 \\ 1 - \Phi\left(\frac{\mathbf{X}_i\boldsymbol{\beta}}{\sigma}\right) & y_i = 0 \end{cases}$$

MARGINAL EFFECTS

In nonlinear models coefficients usually do not have direct interpretation as in linear regression

- Usually we look only at signs of coefficients to get qualitative interpretation

Marginal effects are calculated to obtain absolute interpretation

- This is just a simple derivative

$$ME_{ij} = \frac{\partial \mathbf{E}\left(y_i \mid \mathbf{X}\right)}{\partial X_j}$$

It takes different values for different respondents

- Either average it over respondents or calculate at mean value of **X**

MARGINAL EFFECTS

For Poisson model we will have:

$$ME_{ij} = \frac{\partial \mathbf{E}\left(y_i \mid \mathbf{X}_i\right)}{\partial X_j} = \frac{\partial \lambda_i}{\partial X_j} = \frac{\partial \exp\left(\mathbf{X}_i\boldsymbol{\beta}\right)}{\partial X_j} = \exp\left(\mathbf{X}_i\boldsymbol{\beta}\right)\beta_j$$

For Tobit:

$$ME_{ij} = \frac{\partial \mathbf{E}\left(y_i \mid \mathbf{X}_i\right)}{\partial X_j} = \frac{\partial\left(\Phi\left(\frac{\mathbf{X}_i\boldsymbol{\beta}}{\sigma}\right)\left(\mathbf{X}_i\boldsymbol{\beta} + \sigma\lambda_i\right)\right)}{\partial X_j} = ?$$

Where: $\lambda_i = \phi\left(\frac{\mathbf{X}_i\boldsymbol{\beta}}{\sigma}\right) \Big/ \Phi\left(\frac{\mathbf{X}_i\boldsymbol{\beta}}{\sigma}\right)$

EXERCISE 5: ZERO RESPONSES

1. Read *me.usahealth.rds* dataset into R
2. Compare different approaches of finding an effect of coinsurance rate on medical expenditures
   1. Try OLS, OLS with logarithmic transformation of the dependent variable, Tobit and Poisson model
   2. Calculate marginal effects to interpret results from the model
   3. Compare models fit to data

EXERCISE 5: ZERO RESPONSES

1. Read *me.usahealth.rds* dataset into R
2. Compare different approaches of finding an effect of coinsurance rate on medical expenditures
   1. Try OLS, OLS with logarithmic transformation of the dependent variable, Tobit and Poisson model
   2. Calculate marginal effects to interpret results from the model
   3. Compare models fit to data

Let's go to 

WORKBOOK 2

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

- Exercise 3

Let's go to 

INTERVAL REGRESSION

Often covariates obtained from the surveys are censored on some intervals

- For example instead of asking directly for someone's income, we could ask respondent to indicate to which interval (e.g. 0-1500 PLN, 1500-3000 PLN, 3000-5000 PLN and so on) their income belongs to
- By doing it we of course limit the amount of the information obtained from the survey
- On the other hand such information could be more reliable (more truthful, less prone to error)

In such data research does not have full information about value of the covariate, he only knows which interval it lies within

OLS will lead to biased estimates as it assumes that the values of dependent variable are exact

INTERVAL REGRESSION

More formally, let assume that $F(y_i \mid \beta)$ is a CDF function of the distribution that data are assumed to come from (e.g. normal)

If we only know that $a_i < y_i \leq b_i$, then probability of it will be given by:

$$L_i = F\left(b_i \mid \beta\right) - F\left(a_i \mid \beta\right)$$

This could be used to construct maximum likelihood estimator

- Any continuous distribution could be chosen
- Choice of the distribution will usually be data driven

EXERCISE 6: INTERVAL REGRESSION

1. Read *WTPint.xls* dataset into R
2. Construct an interval variable used by *survival* package
3. Estimate interval regressions with different distributional assumptions

EXERCISE 6: INTERVAL REGRESSION

1. Read *WTPint.xls* dataset into R
2. Construct an interval variable used by *survival* package
3. Estimate interval regressions with different distributional assumptions

Let's go to