The lectures will survey the topic of count regression with emphasis on the role of unobserved heterogeneity. The first part will cover the basic Poisson regression model and its most commonly used extensions. The second part will cover newer topics involving multivariate count models and models with endogenous regressors. The approach will emphasize practical modeling considerations. Illustrations will be drawn from my recent collaborative research in health economics.

**Topics to be covered**

1. Basic Poisson regression and assumptions
2. Unobserved heterogeneity and motivation
3. Models with overdispersion
4. Generalizations of the Poisson
   - Negative binomial regression
   - Hurdle model
   - Zero inflated model
   - Latent class model
5. Maximum likelihood estimation
6. Robust variance estimation
7. Goodness of fit measures
8. An Empirical Example
9. Interpretation of Results
10. Partially parametric models.
11. Bivariate and multivariate count models
12. Count models with endogenous regressors
   - MLE and simulated MLE estimation
   - Copula based models
   - Moment based estimation
13. Estimating the impact (treatment effects) of endogenous variables
14. Empirical illustrations
General references


Useful for empirical work using STATA


Specific references

Topics 1-4:   RACD – Chapters 1, 3, 4.2, 4.7, 4.8, 6;
MMA – Chapter 20.1 – 20.4;
Topics 5-6:   RACD – Chapter 2.5
MMA -- Chapter 20.5
Topic 7: RACD -- Chapter 5
Topic 9: MMA, Chapter 20.5-20.6.
Topics 10-14: MMA, Chapter 20.5-20.6.
C. Li and P.K. Trivedi, “Disparity in Medical Care Use in the USA: The Role of Health Insurance Coverage and Selection Bias”. Discussion paper, 2005.
This assignment is concerned with practical issues regarding modeling of count data. The questions concern modeling a relationship between health care utilization and health insurance, controlling for several economic and socio-demographic factors. Please use the computer program STATA 8.0 or 9.0 to implement the required data analysis tasks.

1. You will receive by email a STATA data file which contains a subsample of data from the famous RAND Health Insurance Experiment.

In tackling this assignment you should use STATA 8.0 or 9.0. In addition to the excellent online help available in STATA you will find it very useful to consult the book, *Regression Models for Categorical and Dependent Variables Using STATA* by J. Scott Long and Jeremy Freese. Chapter 7 of this book will be very useful in implementing the analysis required here.

**Data:** The data are derived from one of the most well known social experiments conducted in the USA during the 1970s and 1980s. In this experiment the participants were “exogenously assigned” insurance plans with different coinsurance rate. The explanatory variable that corresponds to this is \( \log c \). This variable serves the same role as “out-of-pocket price of doctor visit”; it is the central focus of studies of health utilization as a function of its determinants. Another “price variable” is \( \text{IDP} \). The income variable is \( \text{LINC} \), log of income. The demographic variables are family size, \( \text{age} \), \( \text{female} \), \( \text{child} \), \( \text{black} \), \( \text{educed} \). Health status variables are \( \text{physlim} \), \( \text{ndisease} \), \( \text{hlthg} \), \( \text{hlthf} \), \( \text{hlthp} \) - - all are defined below. In your regression \( \text{mdvis} \) is the dependent variable. Your explanatory variables should exclude \( \text{fmde} \) and any other variables in the data file but not mentioned below in the table.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>MDVIS</td>
<td>Number of outpatient visits to an MD</td>
</tr>
<tr>
<td>LOGC</td>
<td>( \ln(\text{coinsurance+1}) ), ( 0 \leq \text{coinsurance} \leq 100 )</td>
</tr>
<tr>
<td>IDP</td>
<td>1 if individual deductible plan, 0 otherwise</td>
</tr>
<tr>
<td>LPI</td>
<td>( \ln(\text{max}(1,\text{annual participation incentive payment})) )</td>
</tr>
<tr>
<td>FMDE</td>
<td>0 if IDP=1, ( \ln(\text{max}(1,\text{MDE}/(0.01 \text{ coinsurance}))) ) otherwise</td>
</tr>
<tr>
<td>LINC</td>
<td>( \ln(\text{family income}) )</td>
</tr>
<tr>
<td>LFAM</td>
<td>( \ln(\text{family size}) )</td>
</tr>
<tr>
<td>AGE</td>
<td>age in years</td>
</tr>
<tr>
<td>FEMALE</td>
<td>1 if person is female</td>
</tr>
<tr>
<td>CHILD</td>
<td>1 if age is less than 18</td>
</tr>
<tr>
<td>FEMCHILD</td>
<td>FEMALE * CHILD</td>
</tr>
<tr>
<td>BLACK</td>
<td>1 if race of household head is black</td>
</tr>
<tr>
<td>EDUCDEC</td>
<td>education of the household head in years</td>
</tr>
<tr>
<td>PHYSLIM</td>
<td>1 if the person has a physical limitation</td>
</tr>
<tr>
<td>NDISEASE</td>
<td>number of chronic diseases</td>
</tr>
<tr>
<td>HTHG</td>
<td>1 if self-rated health is good</td>
</tr>
<tr>
<td>HTHF</td>
<td>1 if self-rated health is fair</td>
</tr>
<tr>
<td>HTHP</td>
<td>1 if self-rated health is poor</td>
</tr>
<tr>
<td></td>
<td>omitted category is excellent self-rated health</td>
</tr>
</tbody>
</table>
Data Analysis Tasks

(a) Using appropriate STATA commands estimate Poisson and negative binomial regression with \( \text{mdvis} \) as the dependent variable and the following explanatory variable:

\[
\log c \quad \text{idp} \quad \text{linc} \quad \text{female} \quad \text{edudec} \quad \text{xage} \quad \text{black} \quad \text{hlthg} \quad \text{hlthf} \quad \text{hlthp}
\]

Carry out a likelihood ratio tests of the null hypothesis that the following variables have zero effect on \( \text{mdvis} \): \( \log c \) \( \text{idp} \).

Estimate the Poisson and NB models with regular and robust variance matrix options. Compare the results and comment on the appropriateness of the robust variance estimation procedure in each case.

(b) Test for overdispersion in the Poisson regression using the formulations (3.39) and (3.40) in the Cameron-Trivedi, RACD, chapter 3. Which version of the variance formulation gets more support from the data? What do you conclude from this exercise?

(c) Estimate the negative binomial model using the \texttt{nbreg} command in STATA. Compare the estimate of the overdispersion parameter with that in part (b) above? Explain the similarities and differences.

Using the results from the Poisson and the NB regressions, compare the actual and fitted frequency distribution of counts. Does the NB model provide a better fit? How good is the fit in the right tail of the distribution?

(d) Using the results from the \texttt{nbreg} estimation above, compare the estimated marginal effect of a change in \( \log c \) (examine the descriptive statistics for \( \log c \)) for an average individual in excellent health (baseline) and an average individual in poor health (\( \text{hlthp} = 1 \)).

(e) For the above \texttt{nbreg} specification estimate the “hurdle version” consisting of a zero-part (logit or probit) and a positive part, (truncated-at-zero negative binomial). Compare these results with those from a regular negative binomial model. Analyze the similarities and differences between the implications of the two models. Based on your analysis, which model do you regard as a better explanation of the data?
SOME USEFUL COMMANDS FOR COUNT MODELS ESTIMATED IN STATA

describe displays a summary of the contents of the data in memory or the data stored in a Stata-format dataset.

summarize calculates and displays a variety of univariate summary statistics. If no varlist is specified, summary statistics are calculated for all the variables in the data. Option: detail produces additional statistics including skewness, kurtosis, and the four smallest and four largest values, along with various percentiles.

tabulate produces one- and two-way tables of frequency counts along with various measures of association.

poisson estimates a Poisson maximum-likelihood regression of depvar on varlist, where depvar is a nonnegative count variable.

nbreg estimates a negative binomial (Poisson with overdispersion) maximum-likelihood regression of depvar on indepvars, where depvar is a nonnegative count variable.

See help gnbreg for a version of the negative binomial model in which the overdispersion parameter can be modeled as a function of other variables.

nbreg will estimate two different parameterizations of the negative binomial model. The default, also given by the option dispersion(mean), has dispersion for the i-th observation equal to 1 + alpha*exp(x_i*b + offset); i.e., the dispersion is a function of the expected mean of the counts for the i-th observation: exp(x_i*b + offset). The alternative parameterization, given by the option dispersion(constant), has dispersion equal to 1 + delta; i.e., it is a constant for all observations.

fitstat computes the goodness of fit statistics for regression just run.

zip estimates a maximum-likelihood zero-inflated Poisson regression of depvar on varlist, where depvar is a nonnegative count variable. Options: inflate(varlist[, offset(varname)]) specifies the equation that determines whether the observed count is zero. Note that inflate() is not optional. Conceptually, omitting inflate() would be equivalent to estimating the model with poisson (zip) or nbreg (zinb).

zinb estimates a maximum-likelihood zero-inflated negative binomial regression of depvar on varlist, where depvar is a nonnegative count variable.
mfx compute calculates marginal effects after a count regression

prvalue computes predicted probabilities for values of independent variable specified with \texttt{x( )} and \texttt{rest( )}.