INDUCTIS
AN EXL COMPANY
IDEAS
ANALYSIS
EXECUTION

Modeling Techniques

Workers Compensation (WC) Team
Maitreyi Mandal
Namit Chopra

TRAVELERS

Presentation Roadmap

- Project Overview

- Censored/Truncated Regression
  - –Tobit Model

- Sample Selection Bias
  - –Inverse Mills Ratio
  - –Heckman's Two Stage Estimation Procedure

- Count Data Regression
  - –Poisson and Negative Binomial
  - –Zero Inflated Models

Post RTW Medical Payment Analysis

- Belongs to the family of ‘RTW’ Projects
- Early warning system based on Return to work (RTW_Date) date and RTW condition RTW_Qualifier).
- 50% claims do not have any Post RTW bills.
- Dependent variables: Post RTW/Total Medical Ratio, No. of RTW Bills.

Presentation Roadmap

- Project Overview

- Censored/Truncated Regression
  - Tobit Model

- Sample Selection Bias
  - Inverse Mills Ratio
  - Heckman's Two Stage Estimation Procedure

- Count Data Regression
  - Poisson and Negative Binomial
  - Zero Inflated Models

Censored/Truncated Regression

- Censored Regression/ Sample is used where the variable of interest is only observable under certain conditions.

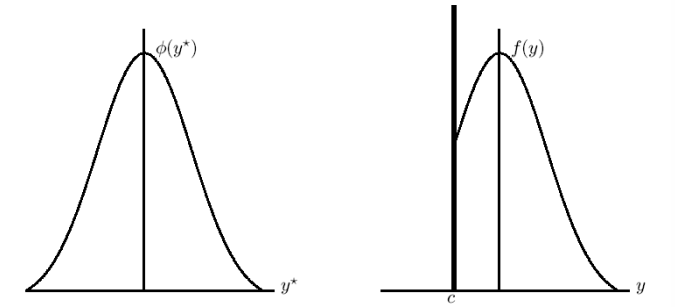


Figure 1: Normal Variable y^* and Censored Variable y

- Truncated regression is used for data, where whole observations are missing so that the value for the dependent and the independent variable is unknown
- Censoring Vs Truncation

Tobit Model

- The **Tobit Model** belongs to the category of limited dependent variable models wherein the dependent variable is roughly continuous over strictly positive values but is zero for a nontrivial fraction of the population i.e. it is censored. It was first proposed by James Tobin (1958).

- Example is the amount an individual spends on alcohol in a given month. In the population of people over age 21 in the United States, this variable takes on a wide range of values. For some significant fraction, the amount spent on alcohol is zero. (Wooldridge, 2nd Ed)

- Mathematically, Tobit Model can be expressed as:

$$Y_i = \beta_i + \beta_2 X_{2i} + u_{2i} \quad \text{if RHS} > 0$$
$$= 0, \text{ otherwise.}$$

Tobit-Example 2

- Problem Statement (Gujarati, 3rd Ed.)

  Find out the amount of money a person or family spends on a house in relation to socioeconomic variables (say, income, mortgage interest rate, number of people in the family, etc.)

- Challenge :

  If a consumer does not purchase a house, obviously we have no data on housing expenditure for such consumers; we have such data only on consumers who actually purchase a house.

- Possible Solution :

  Divide consumers into two groups, one consisting of, say, $n1$ consumers about whom we have information on the regressors as well as the regressand (amount of expenditure on housing) and another consisting of $n2$ consumers about whom we have information only on the regressors but not on the regressand.

- Can we estimate regression using only $n1$ observations and not worry about the remaining $n2$ observations?

- The answer is no, for the OLS estimates of the parameters obtained from the subset of $n1$ observations will be *biased as well as inconsistent;* that is, they are biased even asymptotically.

Tobit-Project Application

- To analyze the ratio of Post RTW medical amount to Total Medical Amount Paid across various dimensions (e.g. injury groups, industry, tenure, age, market etc).
- However, 50% of claims had zero Post RTW medical paid.
- Hence it was decided to use Tobit Model in this project

SAS Computation (Tobit)

- Proc QLIM

```
proc qlim data = out.reg_model_tobit;
model postrtw_total_amount = &independent;
endogenous postrtw_total_amount ~ censored (lb = 0 );
run;
```

SAS Computation (Tobit)

- **Proc LifeReg**

```
proc LifeReg
data out.modeling_data_lifereg;
set out.modeling_data;
if postrtw_total_amount = 0 then
lower = .;
else lower = postrtw_total_amount;

proc lifereg data = out.modeling_data_lifereg;
model (lower,  postrtw_total_amount) = &Independent
/d=normal;
output out = out.lifereg_result;
run;
```

Presentation Roadmap

- Project Overview

- Censored/Truncated Regression
  - –Tobit Model

- Sample Selection Bias
  - –Inverse Mills Ratio
  - –Heckman's Two Stage Estimation Procedure

- Count Data Regression
  - –Poisson and Negative Binomial
  - –Zero Inflated Models

Problems

- Problem of changed signs encountered across proc reg/proc QLIM and proc reg/proc Lifereg.

- Possibility of Sample Selection Bias

- Measures Taken: Inverse Mills Ratio and Heckman's Two Stage Estimation Procedure

Inverse Mills Ratio (IMR)

- It is the ratio of the probability density function over the cumulative distribution function of a distribution.
- It is used to take account of a possible selection bias
- Heckman (1976) proposed a *two stage estimation procedure* using the inverse Mills ratio to take account of the selection bias.

Heckman Two Stage Computation

- **Stage I – Computation of IMR using a Probit Model**

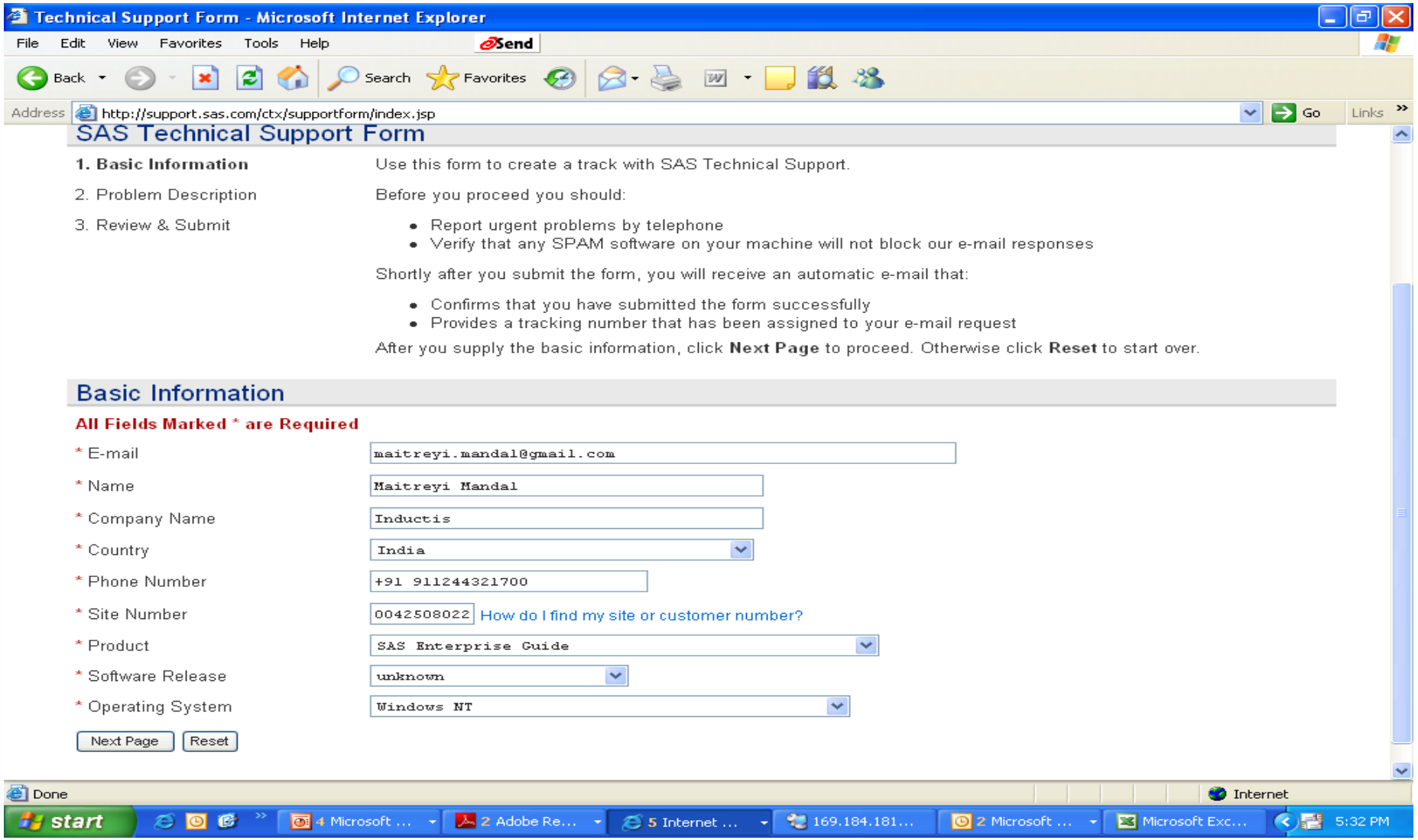
```
proc qlim data = out.reg_model;
output out = out.tobit_mills xbeta;
model postrtw_total_amount_pos = &independent;
run;
data out.reg_model_imr;
set out.tobit_mills;
imr = pdf ('normal', xbeta)/cdf('normal',xbeta);
if cdf ne 0 or .;
run;
```

- **Stage II-Estimating the Final Model using OLS**

```
proc reg data = out.reg_model_imr;
model postrtw_total_amount_pos = &independent;
run;
```

SAS Support for Inductis / Travelers

- Any problem should be reported to: www.support.sas.com



Presentation Roadmap

- Project Overview

- Censored/Truncated Regression
  - Tobit Model

- Sample Selection Bias
  - Inverse Mills Ratio
  - Heckman's Two Stage Estimation Procedure

- Count Data Regression
  - Poisson and Negative Binomial
  - Zero Inflated Models

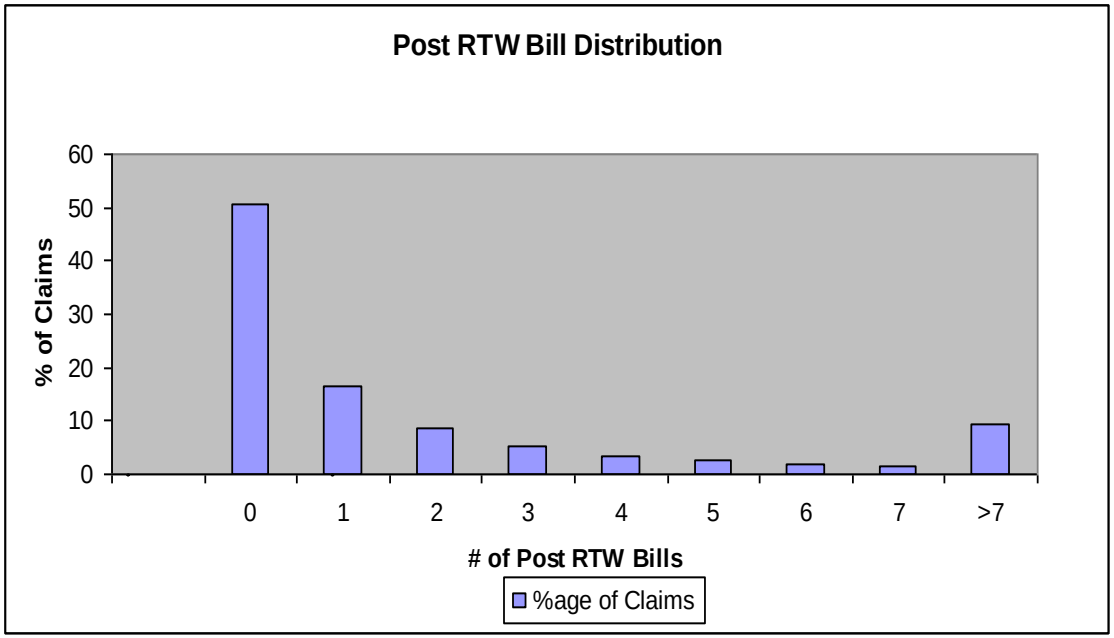
What is Count Data?

- Non-negative integers
- Represent the number of occurrences within a fixed period but can parameterize duration or "exposure"
- King (1989) notes that "one of the most fundamental features of event count data is that the variance of the count increases with the expected number of events."
- Presidential vetoes, US uses of force, war casualties, number of coups, etc.

of Post RTW Bill Distribution

Number of Post RTW Bills	# Claims	%age of Claims
0	99,691	50.76
1	32,147	16.37
2	16,866	8.59
3	10,486	5.34
4	6,806	3.47
5	5,037	2.56
6	3,898	1.98
>7	2,987	1.52

*50.76 % of claims do not have any Post RTW bills



Increasing Model Flexibility

- Poisson
- Negative Binomial
- Generalized Negative Binomial
- Generalized Event
- Count
- Hurdle
- Zero-Inflated Poisson
- Zero-Inflated Negative Binomial

Poisson

- Most basic count model

- Has several restrictive assumptions:
  1. Constant arrival rate and
  2. All events are independent

- One implication of the model specification is that the mean and variance are equivalent ($\mu = V = \lambda_i$)

- $Pr(Y = y) = e^{-\lambda} \lambda_i / y!$ where $\lambda_i = x_i\beta$

Negative Binomial

- Allows for correction of over dispersion
- Result of contagion (non-independent observations)
- Random variation over time (heterogeneity)
- Loosens Poisson restrictions by allowing arrival rate (λ) to vary systematically

SAS Computation

- Proc GenMod

```
proc Genmod data= out.modeling_data_genmod
/*ORDER=INTERNAL*/;
model nfreq =&independent /dist = negbin link = log;
output out    =out.OLS_Resid
p = Pred
resraw= Resid;
run;
```

Zero Inflated Models

- Alternate response to modeling
- Over dispersion
- Believe that the excessive number of zeros may be the result of different DGPs.
- Classic example: number of fish caught in a given park
- Some zeros result from fishing and not catching any fish
- Some zeros result from not fishing at all
- Zero-inflated models allow one to model each process separately
- Usually maps logit onto a count model

(A,b,0) Distribution

- In probability, a discrete probability density function of a random variable X is said to be a member of the (a, b, 0) class of distributions if

$P_k / P_{k+1} = a + b/k$, $k = 1,2,3\ldots$

- where $p_k = P(X = k)$ (provided a and b exist and are real).
- Easy way to determine if a sample was taken from a distribution from the (a,b.0) class is by graphing the ratio of two consecutive observed data (multiplied by a constant) against the x-axis. If a linear trend is seen then it can be assumed that the data is taken from an (a,b,0) distribution.

Proc CountReg (ZIP & ZINB)

- **/*****ZERO INFLATED POISSON (ZIP) ANALYSIS*****/**

proc countreg data = out.modeling_data_genmod type = zip ;
model nfreq =&independent /zi(link = logistic, var =&independent);
run;

- **/*****ZERO INFLATED NEGATIVE BINOMIAL (ZINB) ANALYSIS****/**

proc countreg data = out.modeling_data_genmod type = zinb ;
model nfreq =&independent /zi(link = logistic, var =&independent);
run;

References

- Gujarati,D.N,4th Edition, Basic Econometrics, McGraw-Hill Inc.
- Wooldridge,J.M, 2nd Edition, Introductory econometrics-A Basic Approach, Prentice Hall.
- Liu, W and Cella J (2008),Count Data Models in SAS, Paper 371-2008, SAS Global Forum 2008.
- Wikipedia
- Business Communication with Takeshi Yamaguchi, US WC Team.

Thank You!