Zero-Truncated Count Data Model

- GNU R in SAS/IML Studio-

Chel Hee Lee[1]

[1]Collaborative Biostatistics Program
School of Public Health
University of Saskatchewan

October 12, 2011

Before starting my presentation,

I would like to express my thanks to

- **Matthew Malczewski**, Communication Manager in SAS
- **Jacqueline Quail**, Health Quality Council
- **Gopinath Narasimhan**, Research Officer in Epidemiology Saskatchewan Cancer Agency,
- **Hitesh Bhatt**, Biostatistician in Clinical Research Support Unit College of Medicine

Let's Think Together and Share Knowledge!
(At this time, I try to reflect all your viewpoints and hope this presentation is informative to your future work).

Outline

General Information

SAS/IML Studio 3.2

- is known as **SAS Stat Studio**.
- can **interface with the R system**.

See the chapter "**Calling Functions in the R language**" in SAS/IML Studio 3.2: User's Guide.

GNU R

- is the **open source** software based on the **GNU** philosophy.
- is a matrix-based programming language such as **Matlab**.
- and "**many new statistical methods are first programmed in R**" (Retrieved October 11, 2011, from `http://support.sas.com/rnd/app/studio/Rinterface2.html`).

General Information

One may ask:

- **Why?** Commercial software is always slow to deliver recent academic implementation of brand-new methodology,
- **How?** Exchange data sets or matrices between SAS and GNU R through SAS server-and-client functionality.

GNU (Sharing knowledge) provides:

- Fast knowledge transfer,
- Tight collaboration between business and academia,
- Open minded discussion environment for better research.

Statistical Justification

Statistical Models for Count Data

- **P**oisson
- **N**egative **B**inomial
- **Z**ero **I**nfated **P**oisson
- **Z**ero **T**runcated **P**oisson

Various extensions

- Regression approach
- Hurdle model
- Latent class model approach
- Bayesian approach
- (more)

Statistical Software

Table: Built-in functions for count models in current statistical software

Count Models	STATA STATISTICA	SAS	IBM-SPSS Minitab	Mplus	R NumPy
Basic Form					
P	Yes	Yes	Yes	Yes	Yes
NB	Yes	Yes	Yes	Yes	Yes
ZIP	Yes	Yes	Yes	Yes	Yes
ZINB	Yes	Yes	—	Yes	Yes
ZTP	Yes	—	—	—	—
ZTNB	—	—	—	—	—
Regression					
P	Yes	Yes	Yes	Yes	Yes
NB	Yes	Yes	Yes	Yes	Yes
ZIP	Yes	Yes	—	Yes	Yes
ZINB	Yes	Yes	—	Yes	Yes
ZTP	—	—	—	—	—
ZTNB	—	—	—	—	—

Analytic Solutions

Poisson Model

$$f(y|\mu) = \frac{e^{-\mu}\mu^y}{y!}, \quad y = 0, 1, 2, \ldots$$

$$\mu = E(Y) = V(Y)$$

Truncated Poisson Model

Zero-truncated means the response variable cannot have a value of 0. (Example) the duration patients are in hospital

$$\begin{aligned} f_{ZTP}(y|\mu) &= \frac{e^{-\mu}\mu^y}{y!(1-e^{-\mu})}, \quad y = \cancel{0}, 1, 2, \ldots \\ E_{ZTP}(y|\mu) &= \frac{\mu}{(1-e^{-\mu})} \\ V_{ZTP}(y|\mu) &= \frac{\mu^2}{(1-e^{-\mu})^2}(1-e^{-\mu}-\mu e^{-\mu}) \end{aligned}$$

In your mind,

- Should I buy the STATA?
- Should I wait until the SAS provide the function for this model?
- Can I solve it by hand?
- **Let's do some numerical analysis!**

Which algorithm is appropriate for our problem?

- Grid searching technique (GS),
- Bisectional method (BS),
- Fisher's scoring method (FS),
- Iterative weight least square approach (IWLS).

Let's consider the scenario

- You are the SAS programmer who knows IML but has a little knowledge of R.
- Because of problems of estimating parameters in your statistical model, you cannot make your research move forward.
- You google some information and find that R provides several functions of which may help your estimation problem.
- What will you do?

Difference Between SAS/IML and GNU R?

SAS/IML

```
x = 1:3;
m = {1 2 3, 4 5 6, 7 8 9};


    x               m
1   2   3      1    2    3
               4    5    6
               7    8    9
```

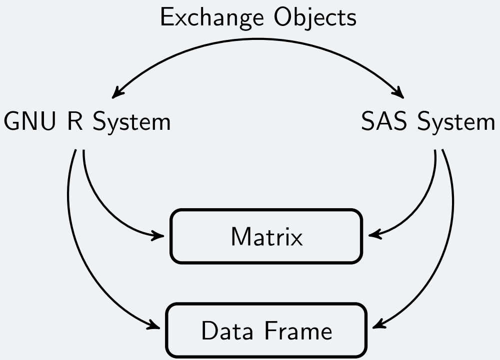
GNU R

```
submit / R;
  rx <- 1:3;
  rm <- matrix( 1:9, nrow=3);
endsubmit;

     [,1] [,2] [,3]
[1,]    1    4    7
[2,]    2    5    8
[3,]    3    6    9
```

How Does It Work?

SAS server-and-client system allows to



Syntax

SAS → R

```
run ExportDataSetToR("SAS.obj", "R.obj" );
run ExportMatrixToR("SAS.obj", "R.obj" );

submit / R;
   < Applying R functions to the data frame named 'R.obj'>
   < Some R programming >
endsubmit;
```

R → SAS

```
run ImportDataSetFromR("R.obj", "SAS.obj");
run ImportMatrixFromR("R.obj", "SAS.obj");

< SAS/IML programming >
```

Testing Numerical Accuracy

ZTP Engine Written By Chel Hee Lee

```
# Set the initial values
mu.par <- 1;
n.samples <- 100;

# Simulate random variates
y <- y.ztp.fn(n=n.samples, mu=mu.par);

# Estimate
est <- ztp.engine(y=y);

# Compare the estimate with the value obtained by theory
theory.mean <- 1/(1-dpois(0, est$mu))*est$mu;
```

```
The estimated mean (from samples) =  1.594
The expected mean (by theory)     =  1.594
```

This ZTP engine is available (by request).

Is It Necessary To Call R In SAS?

There are no numerical difference between SAS and GNU R

- Depending on your programming preference,
- Someone who loves SAS may say “SAS/IML is sufficient to develop a brand-new method”.
- Someone who loves R may say “Do we need the SAS software?”

Next Plan

- ZTNB, ZTPREG, ZTNBREG, Bayesian models will be completed by October 31, 2011

Questions?

Please feel free to contact me

- if you have some statistical concerns that you wish to discuss
- **Decision Making in Chronic Disease Surveillance using Administrative Health Data**,
- **Statistical Reasoning on Incomplete, Imperfect, and Fuzzy Information**,
- **Computational Statistics**, and **GNU R**
- `chl948@mail.usask.ca`

References

[1] SAS Institute Inc. 2009. SAS/IML®Studio 3.2: User's Guide. Cary, NC: SAS Institute Inc.
[2] R Development Core Team (2010). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/