Paper 265

Modeling Event Count data with Proc GENMOD and the SAS System

Matthew Flynn
The Hartford

Introduction

Event count data are distinguished by being positive and integer valued with often small numbers of unique values. Examples include just about anything measured by counts or summary frequency data. Standard ordinary least squares (OLS) regression modeling requires the assumption that the model errors, ε_i, are independently and identically distributed, (i.i.d.)~ $N(0,\sigma)$, normal random variables. In practice, however, departures from this ideal situation are common. This broader class of regression models can be handled under a powerful and flexible set of models called General Linear Models, or GLMs, see McCullagh and Nelder (1989) .

SAS provides a number of tools built to accommodate a variety of different modeling situations, but unfortunately, many users are on unfamiliar ground, and consequently are less confident in these modeling situations. This article attempts to illustrate, through example and discussion, some of the unique features of modeling event count data with SAS and Proc GENMOD.

Preliminary Discussion - Poisson Distribution

Using a set of parameters ranging from a mean of 1.5 to 10.5, the following graphs (figures 1-4) illustrate the shape the Poisson distribution for increasing levels of the single parameter, the mean. (The graphs reproduce the output in Long (1997), figure 8.1, page 219.) At low levels of the mean, this discrete distribution is sharply skewed, but quickly tends to a normal distribution as the mean goes above ten. If one has a frequency distribution of counts with a mean at fifteen or above, the results for modeling a Poisson distribution will be very close to an OLS normal model fit.

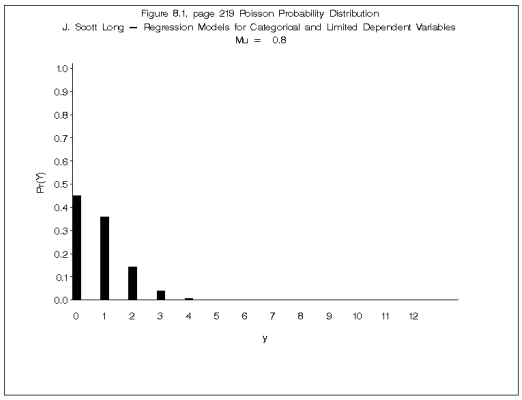

Figure 1

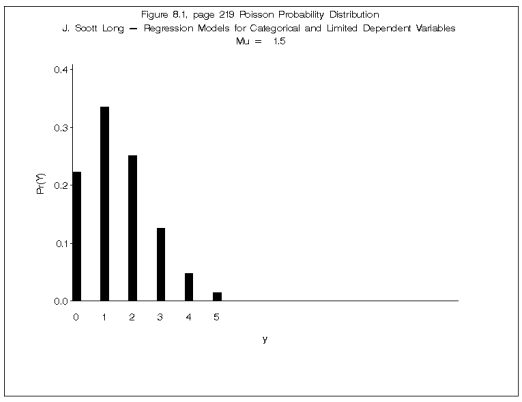

Figure 2

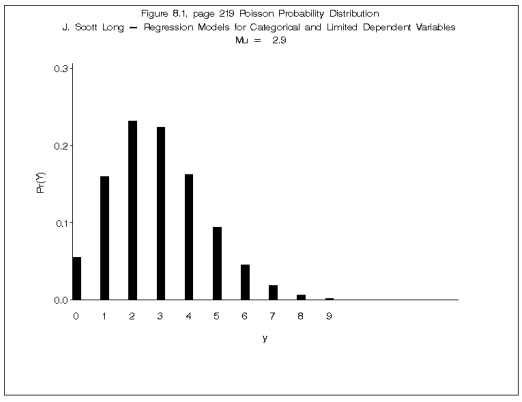

Figure 3

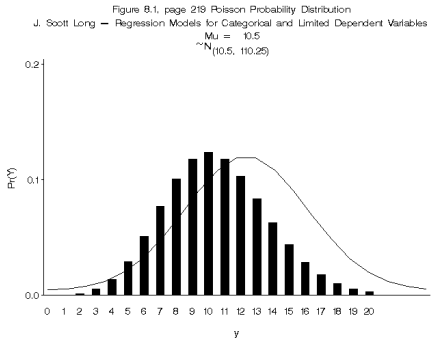

Figure 4

Example

The example data set describes the frequency of reported damage incidents to cargo ships taken from a modern classic, Generalized Linear Models, by McCullagh and Nelder (1989, Table 6.2, page 205). The data set is available over the web, along with many other example data sets at Carnegie Mellon University, Dept. of Statistics -Statlib: http://temper.stat.cmu.edu/datasets/ships.

Exploratory Data Analysis (EDA)

The analyst's first motto is: Know Thy Data!, SAS provides a number of convenient tools to describe data. Using some value and picture formats, data transformations and put statements, one can reproduce Table 6.2 in McCullagh and Nelder (1989). (The entire code describing this analysis is available via E-mail from the author).

McCullagh & Nelder Example 6.3.2, Page 205
Table 6.2 Number of reported damage incidents and aggregate mont
service by ship type, year of construction and period of operati

Obs	Ship type	Year of construction	Period of operation	Aggregate months service	Number damag incide
1	A	1960-64	1960-74	127	0
2	A	1960-64	1975-79	63	0
3	A	1965-69	1960-74	1095	3
4	A	1965-69	1975-79	1095	4
5	A	1970-74	1960-74	1512	6
6	A	1970-74	1975-79	3353	18
7	A	1975-79	1960-74	0	0
8	A	1975-79	1975-79	2244	11
9	B	1960-64	1960-74	44882	39
10	B	1960-64	1975-79	17176	29
⋮					

The data give the number of damage incidents, the total number of months in service (the period for which a ship was at risk for the event, an important consideration), and three classifying factors: ship type, year of construction, and service period. Note that ships built after 1975 cannot have sustained damage in the period prior to 1974. This produces what is called a structural zero, or a necessarily empty cell (which describes an impossibility such as the class of pregnant males, or drivers 18 years old driving more that 20 years, etc.). Structural zeros must be treated differently than accidental zeros (cells which simply did not have an event, perhaps because they did not have a lot of exposure in that cell or combination of factors.

Preliminary statistics can be quickly generated with workhorse SAS tools:

```
Proc univariate data=ships
    normal plot;
  var damage;
 output out=outstat n=n mean=mu
std=std var=var;
run;
```

The analyst will want to utilize several of the printed statistics of the damage variable for later use. The estimated parameters of the empirical distribution of the dependent variable are captured and output to a temporary SAS dataset. These will then be utilized below. A great aspect of the SAS system is the degree of flexibility in accomplishing a given set of tasks. This flexibility can be confusing to the beginner, but allows one to efficiently solve analysis problems. In this example, the calculated sample mean (mu) can be combined back with the original data set:

```
data ships;
   if _N_ = 1 then set outstat;
  set ships;
Proc print data=ships(obs=5); run;
```

N	MU	VAR	SHIP	YEAR	PERIOD	MONTHS
8	3.09646	11.6837	A	1975-79	1975-79	2244
7	3.09646	11.6837	A	1975-79	1960-74	0
6	3.09646	11.6837	A	1970-74	1975-79	3353
5	3.09646	11.6837	A	1970-74	1960-74	1512
4	3.09646	11.6837	A	1965-69	1975-79	1095

Another method places the statistics in SAS MACRO variables, which can be utilized both in later calculations as well as in titles or footnotes:

```
call symput('mu',compbl(mu)); run;
 title 'Ships dataset Average
 damage=&mu'; run;
```

Graphics or data visualization often does wonders in statistical analysis. As a reminder, here we are interested in identifying variables that affect the relative incidence or frequency of the dependent variable, which is described by a positive-valued integer count variable. One can quickly generate a compact description of that count with:

```
Proc FREQ data=ships;
  tables damage / out=outfreq;
run;
```

Again, the output is saved in a temporary SAS data set and immediately used below:

```
data outfreq;
  set outfreq;
/* using SAS function */
predY = pdf('POISSON', damage, &mu);

/* hardcoded */
pry2  = (exp(-mu)*mu**y)/gamma(y+1);
run;
```

The PDF function generates a predicted value from a theoretical Poisson distribution with a mean value equal to the sample mean (gathered above in a macro var.) of the empirical distribution. We will next graph both the empirical distribution and the generated theoretical distribution to view how well the theoretical Poisson distribution fits the sample data.

```
Proc gplot data=outfreq;
 plot pct*damage predY* damage /
overlay;
```

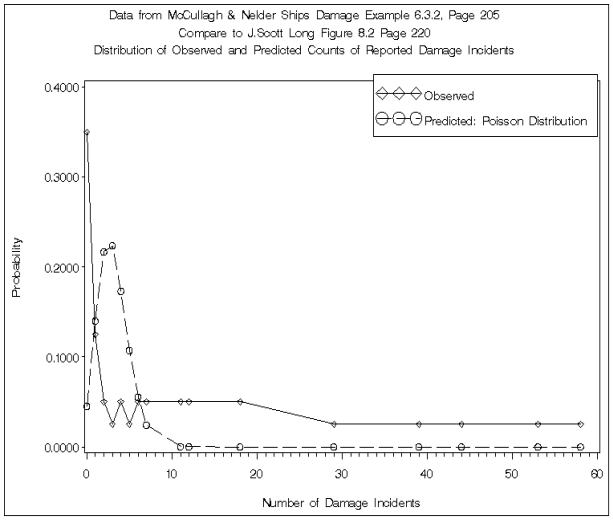

figure 5

The plot illustrates several typical aspects of modeling event count data. The observed frequency is positively skewed, and is overdispersed. Overdispersion or extra-dispersion means that the variance is larger than the mean. One view of the source of overdispersion in this example is that it is could be due to inter-ship variability in the likelihood of a damage incident. The Poisson distribution has only one parameter, which succinctly describes the entire distribution, but that also requires that the variance equal the mean of the distribution, which rarely occurs in practice. A consequence of overdispersion evident in figure 5 is that a theoretical Poisson distribution will under-predict the number of zeros, over-predict values in the center of the distribution, or near the mean, and under-predict large values.

Continuing with our data exploration (know thy data!) prior to modeling, we will

generate a table of how the rate of damage incidence varies by independent variables. Frequently these types of table are called one-way or two-way tables. They can be usefull in summarizing model fit as well as potentially identifying possible interaction effects.

```
proc summary data=ships nway;
 class ships ear;
 var months damage;
output out=shipout(drop=_type_
_freq_) sum=;

 data ships;
  set ships;
  rate = (damage / months)*100;

proc tabulate data=shipout missing
noseps format=12.1 formchar=...;
class ship year;
var rate;
label year='Year of construction';
  keylabel sum=' ';
 table ship='', year*mean=' '*(rate='
')
/ box='                          ship
type';
```

McCullagh & Nelder Example 6.3.2, Page 208
Table 6.4 Observed rate of damage incidents (x1,000 per ship month at risk) by ship type and year of construction

Ship Type	Year of construction			
	1960-64	1965-69	1970-74	1975-79
A	0.0	3.2	4.9	4.9
B	1.1	2.3	2.8	2.5
C	1.2	0.7	2.9	3.6
D	0.0	0.0	8.3	2.0
E	0.0	11.4	5.1	1.8

figure 6

This table shows that the observed rate of damage incidents for ship types A, B and C are increasing over time, but that ship type E appears to decrease after the 1965-69 period. Interaction effects can be explicitly tested with a series of Proc GENMOD model statements, as illustrated below.

Plotting these data illustrates this point:

```
Proc gplot data=ships;
  plot rate * year = ship;
```

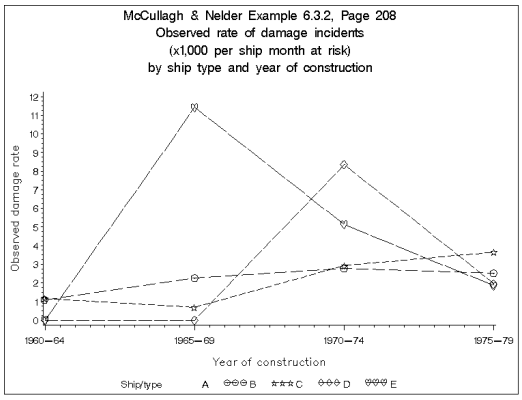

Figure 7

At first guess, it seems reasonable to assume that the frequency of damage incidents is directly proportional to the amount months in service of the ships (the amount of exposure). This is easily handled in Proc GENMOD, but should be checked! This is easy to check, but this assumption is easy to overlook!

Model Fitting

Next we will fit a sequence of models, beginning with a one-parameter, or null model, and then next examining a main effects model:

log(E(# damage incidents)) = B_0 + log(months service) + type + year + period.

The general syntax is:

```
proc GENMOD data=ships order=data;
  class ship type period;
  model damage = ship type period
          / offset=lmonths
dist=Poisson link=log;
```

The log of months service term in the model is called an *offset* term. It is equivalent to modeling the dependent variable as a rate model (count / period) = $x_1 + x_2$, or model count = offset(period) + $x_1 + x_2$. One way to check if this assumption is valid is to

explicitly model using the offset variable. One confusing matter for beginners in using an offset term is that printed GENMOD output does not directly show how the offset was used. The offset option in Proc GENMOD does provide a very flexible means for testing hypotheses about model parameters - one can easily fix parameters to desired levels using the offset term and re-estimate remaining model parameters. This is one method that can be used to fit combined additive and multiplicative model structures.

The syntax to explictly use the offset factor:

```
proc GENMOD data=ships order=data;
  class ship type period;
  model damage = lmonths ship type
   period / dist=Poisson link=log;
```

The estimated coefficient for the exposure variable (lmonths) should not be significantly different from one.

Prior to fitting the main effects model (and others), it is convenient to fit a model with no covariates, e.g. (model = / dist=poisson link=log). This is also often termed a null model or a one-parameter model. The deviance statistic from this model reports the total amount of variability that is available to be 'explained'. Then reductions in model deviance in subsequent model fits with covariates allow one to calculate the equivalent of R-squared in standard normal-theory models (Cameron and Windmeijer (1998)).

Proc GENMOD provides very flexible output - any printed table produced can be output into a SAS dataset. One example taking advantage of this is estimating the significance of the model fit. The differences in nested model deviances is a likelihood ratio test where the statistic is distributed as Chi-square with k degrees of freedom (k = number of additional parameters estimated).

Using the Proc GENMOD make option, one can output the printed table and then automatically perform a significance test with the appropriate distribution function. For additional discussion, see Tjur (1998).

```
proc GENMOD data=ships order=data;
  class ship type period;
  make 'modfit' out=modfit;
  model damage = ship type period
        / offset=lmonths dist=Poisson
         link=log;

data modfit; set modfit;
 p = 1- cdf('CHISQ', value, df);
```

After fitting a sequence of models, from a null model with no variables, to a main effects model (as above) one can investigate interaction effects by fitting models with higher order effects:

```
   proc GENMOD data=ships order=data;
     class ship type period;
     make 'modfit ' out=modfit;

   model damage = ship type period
ship*period year*period / ...

   Alternative Proc GENMOD model
statement syntax fits all two-level
interations.

model damage = ship|type|period @2;
```

The sequence of model fit statistics can be collected and reported in a common table for ease of comparision.

MODEL	MODEL2	VALUE	DIFF1	VALUEDF	R_KL
1	I	146.3283	.	4.4342	0.00000
2	Ship, Year, Period	38.6951	107.633	1.5478	0.73556
3	+ Ship*Year	14.5869	24.108	1.1221	0.90031
4	+Ship*Year + Year*Period	6.8565	7.730	0.9795	0.95314
5	All Two-level Interactions	0.0000	6.856	.	1.00000

The addition of the ship type by year interaction effects in model three does reduce the deviance quite a bit, but that involves fitting twenty additional parameters (Five ship types times 4 periods). The main effects model (model two) actually fits quite well, 'explaining' 73% of the available deviance.

Further Reading

Myers and Montgomery (1997), provide a nice introduction to GLMs and illustrate well how GLMs relate to Ordinary Least Squares. Long (1997), chapter 8 describes modeling count outcomes and discusses extensions of the Poisson model such as negative binomial and zero-inflated Poisson (ZIP) type models. Nelder (1998), discusses GLM model building techniques.

Author Contact Information


```
Matthew J. Flynn, Ph.D.
The Harford Personal Lines
400 Executive Blvd.
Southington, CT 06489

email: Matt.Flynn@TheHartford.com
phone: (860) 620-6891
fax:   (860) 276-2855
```

References

Cameron, A. Colin and Frank A. Windmeijer, (1998), An R^2 measure of goodness-of-fit for some common nonlinear regression models, Journal of Econometrics, 79, 2, 329-342.

Johnston, Gordon, (1993), SAS Software to fit the General Linear Model, SUGI 13, SAS Institute.

Long, J. Scott, (1997), Regression Models for Categorical and Limited Dependent Variables, Sage Publications.

McCullagh, Peter and J. A. Nelder, (1989), Generalized Linear Models, Chapman and Hall.

Myers, R. and D. Montgomery, (1997), A Tutorial on Generalized Linear Models, Journal of Quality Technology, 29, 3, 274-291.

Nelder, J. A., (1998). The selection of terms in response-surface models - How strong is the weak-heredity principle?, The American Statistician, 52, 4, 315-318.

Tjur, Tue, (1998), Nonlinear Regression, Quasi Likelihood, and Overdispersion in Generalized Linear Models, The American Statistician, 52, 3, 222-227.