Paper 371-2008

Count Data Models in SAS®

WenSui Liu, ChoicePoint Precision Marketing, Alpharetta, GA
Jimmy Cela, ChoicePoint Precision Marketing, Alpharetta, GA

ABSTRACT

Poisson regression has been widely used to model count data. However, it is often criticized for its restrictive assumption of equi-dispersion, meaning equality between the variance and the mean. In real-life applications, count data often exhibits over-dispersion and excess zeroes. While Negative binomial regression is able to model count data with over-dispersion, both Hurdle (Mullahy, 1986) and Zero-inflated (Lambert, 1992) regressions address the issue of excess zeroes in their own rights. Different modeling strategies for count data and various statistical tests for model evaluation are illustrated through an example of healthcare utilization. The purpose of this paper is to provide by far the most complete survey of count data modeling strategy in SAS for the user group.

KEYWORDS

Poisson regression, Negative binomial regression, Hurdle regression, Zero-Inflated regression, Overdispersion, Excess Zeroes, Vuong test.

1. INTRODUCTION

How to model count data as the dependent variable in a regression has become a popular topic in statistics, econometrics, and epidemiology. Deb and Trivedi (1997) modeled the demand for healthcare utilization by the elderly using a finite mixture negative binomial regression. Gurmu (1997) evaluated the impact of managed care program on healthcare utilization using hurdle model. Winkelmann (2004) studied the effect of healthcare reform on the number of doctor visits in Germany using a number of modified count data models. For more detailed discussions about recent development in count data models, please refer to Cameron and Trivedi (2001), Winkelmann and Zimmermann (1995), and Greene (2002).

To illustrate models covered in this paper, we use the same data analyzed by Deb and Trivedi (1997). This data is originally obtained from National Medical Expenditure Survey (NMES) conducted in 1987 and includes 4406 respondents who were aged 66 or older and covered by Medicare program. In our example, the number of hospital stays (HOSP) is used as the dependent variable and three types of measures are included in the explanatory variables, which are self-perceived health status (EXCLHLTH, POORHLTH, and NUMCHRON), demographic data (AGE and MALE), and socio-economic information (SCHOOL and PRIVINS). The summary statistics of all variables are given in Table 1.

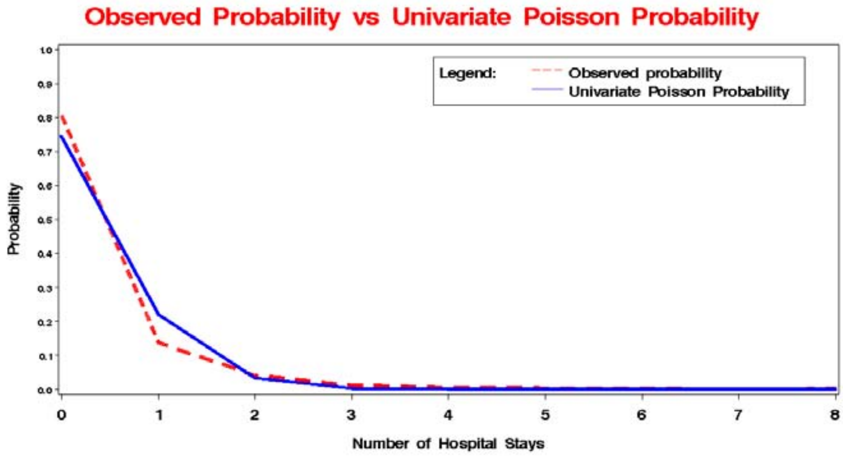
Table 1.1, Variables Used with Summary Statistics

Variable	Definition	Obs	Mean	Std. Dev.	Min	Max
HOSP	# of hospital stays	4406	0.2960	0.7464	0	8
EXCLHLTH	1 if self-perceived health is excellent	4406	0.0778	0.2680	0	1
POORHLTH	1 if self-perceived health is poor	4406	0.1257	0.3316	0	1
NUMCHRON	# of chronic conditions	4406	1.5420	1.3496	0	8
AGE	age in years (divided by 10)	4406	7.4024	0.6334	6.6	10.9
MALE	1 if the person is male	4406	0.4035	0.4907	0	1
SCHOOL	# of years of education	4406	10.2903	3.7387	0	18
PRIVINS	1 if the person is covered by private insurance	4406	0.7764	0.4167	0	1

As shown in Table 1.1, the variance of HOSP is about two times of the mean, implying the possibility of over-dispersion. A further screening on the data also shows that more than 80% of the respondents, 3541 out of 4406, have no hospital admission, indicating excess zeroes.

A good starting point of count data modeling is to compare the empirical distribution of observed counts to the univariate Poisson distribution with the mean estimated from the data. Probabilities from two distributions are plotted in Figure 1.1.

Figure 1.1, Comparison between Observed Probability and Univariate Poisson Probability



The plot in Figure 1.1 clearly shows that univariate Poisson distribution underestimates the probability at 0 and overestimates the probability at 1. Since Poisson distribution assumes the same mean across the whole sample and doesn't consider the heterogeneity in each member, it is not surprising to see that the predicted probability does not fit the observed data well. In the next section, we will allow the observed heterogeneity in the conditional mean of each sample member by including explanatory variables.

2. POISSON REGRESSION

Poisson regression is the simplest regression model for count data and assumes that each observed count Y_i is drawn from a Poisson distribution with the conditional mean u_i on a given vector X_i for case *i*. Therefore, the density function of Y_i can be expressed as

$$f(Y_i \mid X_i) = \frac{Exp(-u_i) \times u_i^{Y_i}}{Y_i!}, \text{ where } u_i = Exp(X_i \beta). \tag{2.1}$$

Given independent observations with the density function in (2.1), the log likelihood function can be obtained by

$$LL = \sum_{i=1}^{n} \left[-u_i + Y_i Log(u_i) - Log(Y_i!) \right]. \tag{2.2}$$

The maximum likelihood estimation of Poisson regression is straightforward using the log likelihood function in (2.2).

In SAS, several procedures in both STAT and ETS modules can be used to estimate Poisson regression. While GENMOD, GLIMMIX, and COUNTREG are easy to use with standard MODEL statement, NLMIXED, MODEL, NLIN provide great flexibility to model count data by specifying the log likelihood function explicitly. An illustration of both NLMIXED and COUNTREG procedures is given below. More detailed examples on how to use all mentioned procedures can be found on author's blog at statcompute.spaces.live.com.

```
/* METHOD 1: PROC NLMIXED  */
proc nlmixed data = tblNMES;
  parms b0 = 0 b1 = 0 b2 = 0 b3 = 0 b4 = 0 b5 = 0 b6 = 0 b7 = 0;
  mu = exp(b0 + b1 * EXCLHLTH + b2 * POORHLTH + b3 * NUMCHRON + b4 * AGE +
           b5 * MALE + b6 * SCHOOL + b7 * PRIVINS);
  ll = -mu + HOSP * log(mu) - log(fact(HOSP));
  model HOSP ~ general(ll);
  predict mu out = poi_out (rename = (pred = Yhat));
run;

/* METHOD 2: PROC COUNTREG */
proc countreg data = tblNMES type = poisson;
  model HOSP = EXCLHLTH POORHLTH NUMCHRON AGE MALE SCHOOL PRIVINS;
run;

/* SAMPLE OUTPUT OF PROC COUNTREG:
                               Model Fit Summary
                    Log Likelihood                      -3046
                    AIC                                  6108
                    SBC                                  6159

                              Parameter Estimates
```

```
                                       Standard                   Approx
        Parameter        Estimate         Error     t Value     Pr > |t|
        Intercept       -3.329044      0.339728       -9.80       <.0001
        exclhlth        -0.723412      0.175644       -4.12       <.0001
        poorhlth         0.626157      0.067858        9.23       <.0001
        numchron         0.264462      0.018277       14.47       <.0001
        age              0.186406      0.042014        4.44       <.0001
        male             0.103186      0.056274        1.83       0.0667
        school          -0.000206      0.007871       -0.03       0.9791
        privins          0.108652      0.069251        1.57       0.1167
*/
```

While Poisson regression is often used as a baseline model for count data, its assumption of equi-dispersion is too restrictive for many empirical applications. In practice, the variance of observed count data usually exceeds the mean, namely over-dispersion, due to the unobserved heterogeneity and/or excess zeroes. With the similar consequences of heteroskedasticity in the linear regression, over-dispersion in a Poisson regression will lead to deflated standard errors of parameter estimates and therefore inflated t-statistics. After the development of Poisson regression, it is always a sound practice to do an additional analysis for over-dispersion. In our example, we will consider two statistical tests based on the alternative Negative binomial model, which will be covered in our next section.

Cameron and Trivedi (1996) introduced a simple test for over-dispersion based on an auxiliary OLS regression without the intercept, which can be formulated as

$$\frac{(y_i - u_i)^2 - y_i}{u_i} = \alpha u_i + e_i \text{, where } u_i = Exp(X_i \beta) \text{ and } e_i \text{ is an error term.} \qquad (2.3)$$

The significance of t-statistics for the coefficient implies the existence of over-dispersion. Please note that (2.3) is specific for Negbin 2 form, the most common setting for Negative binomial. For Negbin 1 form, a different formulation of OLS regression should be used.

$$\frac{(y_i - u_i)^2 - y_i}{u_i} = \alpha + e_i \qquad (2.4)$$

The implementation in SAS with Reg procedure based on (2.3) is given below.

```
data ols_tmp;
  set poi_out;
  dep = ((HOSP - Yhat) ** 2 - HOSP) / Yhat;
run;

proc reg data = ols_tmp;
  model dep = Yhat / noint;              /* FIT A OLS REGRESSION WITHOUT INTERCEPT */
run; quit;

/* OUTPUT OF AUXILIARY OLS REGRESSION:
                              Parameter Estimates
                                    Parameter     Standard
  Variable    Label             DF    Estimate        Error  t Value  Pr > |t|
  Yhat        Predicted Value    1     1.63419      0.22609     7.23    <.0001
*/
```

The second test for over-dispersion introduced by Greene (2002) is based on the Lagrange multiplier (LM) statistics. If we consider Poisson regression a parametric restriction of Negative binomial regression with the mean equal to the variance, the LM statistics can be simply expressed as

$$LM = \frac{(e`e - n\bar{Y})^2}{2u`u} \text{, where } u = Exp(X`\beta) \text{ and } e = Y - u \qquad (2.5)$$

Under the null hypothesis of Poisson regression, the LM statistics follows the chi-squared distribution with one degree of freedom. The computation is extremely simple with SAS IML procedure or any other matrix languages.

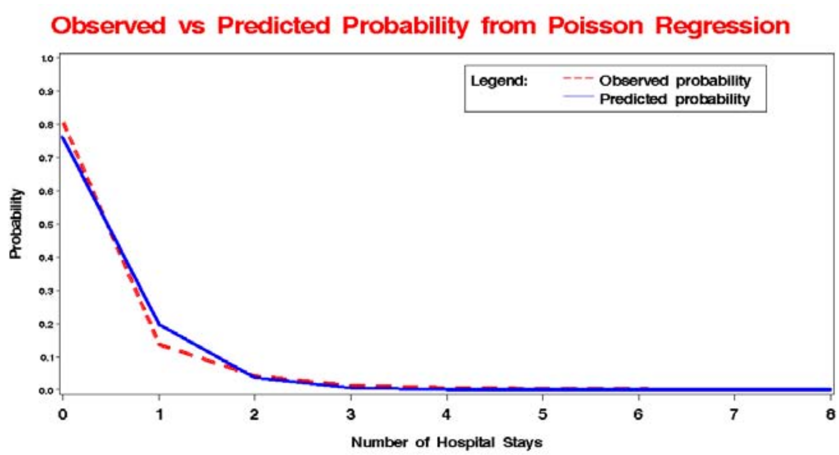
```
proc iml;
  use poi_out;
  read all var {HOSP} into y;
  read all var {Yhat} into yhat;
  close poisson_out;
  e      = (y - yhat);
  n      = nrow(y);
  ybar   = y`[, :];
  LM     = (e` * e - n * ybar) ** 2 / (2 * yhat` * yhat);
  Pvalue = 1 - probchi(LM, 1);
  print LM Pvalue;
```

```
quit;

/* OUTPUT OF LM STATISTICS:
                                       LM     PVALUE
                                794.14707          0
*/
```

Both tests come up with the same conclusion and confirm our suspicion of over-dispersion.

To evaluate the goodness-of-fit of a regression for count data, the most popular but somewhat problematic practice is to compare the predicted and observed values of the dependent variable. However, a measure of goodness-of-fit solely based upon the expected value is unable to address the improvement achieved by a model with less restrictive variance assumption. A better alternative is to compare the predicted and observed probabilities of each count outcome by taking the probability distribution into consideration. In Figure 2.1 below, a plot comparing observed probabilities to predicted probabilities side by side is given.

Figure 2.1, Comparison between Observed and Predicted Probability from Poisson Regression



Compared with Figure1.1, we can see a moderate improvement shown in Figure 2.1 after considering the observed heterogeneity in Poisson regression. However, the under-prediction at 0 and over-prediction at 1 suggest that a further improvement is still possible. In the next section, we will discuss an alternative model with less restrictive assumption, namely Negative binomial regression.

3. NEGATIVE BINOMIAL REGRESSION

As the most common alternative to Poisson regression, Negative binomial regression addresses the issue of over-dispersion by including a dispersion parameter to accommodate the unobserved heterogeneity in the count data. While there are many variants of Negative binomial, we will only focus on the Negbin 2 form in our paper.

Negative binomial regression can be considered a generalization of Poisson regression and assumes that the conditional mean u_i of Y_i is not only determined by X_i but also a heterogeneity component e_i unrelated to X_i. The formulation can be expressed as

$$u_i = Exp(X_i\beta + e_i) = Exp(X_i\beta)Exp(e_i)\text{, where } Exp(e_i) \sim Gamma(\alpha^{-1}, \alpha^{-1}) \qquad (3.1)$$

As a result, the density function of Y_i can be derived as

$$f(Y_i \mid X_i) = \frac{\Gamma(Y_i + \alpha^{-1})}{\Gamma(Y_i + 1)\Gamma(\alpha^{-1})} \left(\frac{\alpha^{-1}}{\alpha^{-1} + u_i} \right)^{\alpha^{-1}} \left(\frac{u_i}{\alpha^{-1} + u_i} \right)^{Y_i}. \qquad (3.2)$$

And the corresponding log likelihood function becomes

$$LL = \sum_{i=1}^{n} \left[Log\left(\frac{\Gamma(Y_i + \alpha^{-1})}{\Gamma(Y_i + 1)\Gamma(\alpha^{-1})} \right) - (Y_i + \alpha^{-1}) Log(1 + \alpha u_i) + Y_i Log(\alpha u_i) \right]. \qquad (3.3)$$

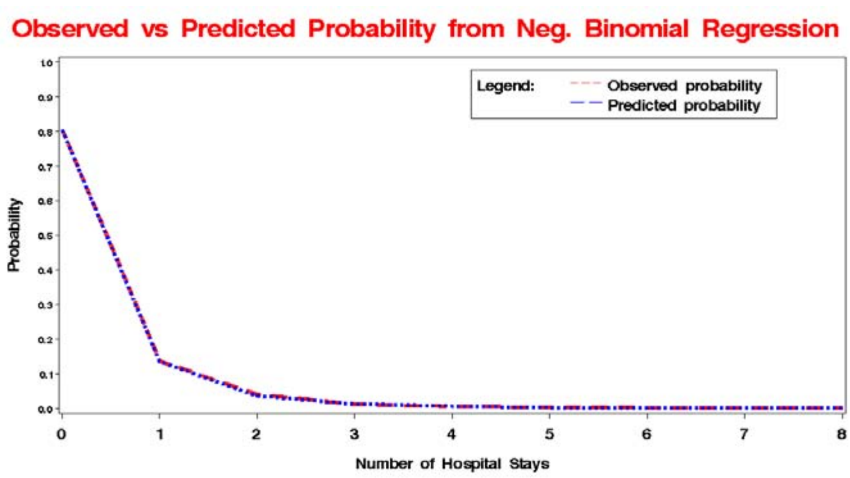
Similar to Poisson regression, Negative binomial regression can be modeled by SAS either directly with GENMOD, GLIMMIXED, and COUNTREG procedures or through log likelihood function in (3.3) with NLMIXED, MODEL, and NLIN procedures. A brief example of SAS code and related output is given below.

```
/* METHOD 1: PROC NLMIXED  */
proc nlmixed data = tblNMES;
  parms b0 = 0 b1 = 0 b2 = 0 b3 = 0 b4 = 0 b5 = 0 b6 = 0 b7 = 0;
  mu = exp(b0 + b1 * EXCLHLTH + b2 * POORHLTH + b3 * NUMCHRON + b4 * AGE +
           b5 * MALE + b6 * SCHOOL + b7 * PRIVINS);
  ll = lgamma(HOSP + 1 / alpha) - lgamma(HOSP + 1) - lgamma(1 / alpha) +
       HOSP * log(alpha * mu) -
       (HOSP + 1 / alpha) * log(1 + alpha * mu);
  model HOSP ~ general(ll);
  predict mu out = nb_out (rename = (pred = Yhat));
run;

/* METHOD 2: PROC COUNTREG     */
proc countreg data = tblNMES type = negativebinom method = qn;
  model HOSP = EXCLHLTH POORHLTH NUMCHRON AGE MALE SCHOOL PRIVINS;
run;

/* SAMPLE OUTPUT OF PROC COUNTREG:
                               Model Fit Summary
                    Log Likelihood                    -2857
                    AIC                                5731
                    SBC                                5789

                              Parameter Estimates
                                           Standard                 Approx
        Parameter         Estimate            Error    t Value    Pr > |t|
        Intercept        -3.752640         0.446835      -8.40      <.0001
        exclhlth         -0.697875         0.193318      -3.61      0.0003
        poorhlth          0.613926         0.095392       6.44      <.0001
        numchron          0.289418         0.026470      10.93      <.0001
        age               0.238444         0.055265       4.31      <.0001
        male              0.153862         0.073033       2.11      0.0351
        school           -0.002271         0.010203      -0.22      0.8238
        privins           0.093922         0.090494       1.04      0.2993
        _Alpha            1.766727         0.160492      11.01      <.0001
*/
```

Please note that Negative binomial regression is the extension of Poisson with a more liberal variance assumption and could collapsed into Poisson regression with the dispersion parameter equal to 0. This important fact provides a possibility to do the model comparison between Poisson and Negative binomial regressions. First of all, we can looked at the reported t-statistics of dispersion parameter, Alpha, to assess the significance of over-dispersion. Then a likelihood ratio (LR) test, which follows Chi-square distribution with 1 degree of freedom, between 2 regressions can be used to determine the preferred model for the data. In our example, the t-statistics of Alpha is 11.01 and the LR test is $-2(LL_{Poisson} - LL_{Negbin}) = -2[-3046 - (-2857)] = 378$, both of which are highly significant and indicate that Negative binomial regression is preferred over Poisson regression. This result is also consistent with our findings of over-dispersion in Section 2.

The goodness-of-fit of Negative binomial regression can be visualized in the similar way to Figure 2.1, as shown in Figure 3.1.

Figure 3.1, Comparison between Observed and Predicted Probability from Negative Binomial Regression



Compared with Figure 1.1 and 2.1, we can clearly see the significant improvement made by Negative binomial regression in Figure 3.1, a nearly complete overlap between predicted and observed probabilities. However, Negative binomial regression is not without criticism. The inclusion of unobserved heterogeneity will increase the probabilities of both zero counts and high counts but might not yield a good fit for the distribution of count outcome with excess zeroes. In the next two sections, we will introduce two alternative models to handle excess zeroes, which are Hurdle regression (Mullahy 1986) and Zero-inflated regression (Lambert 1992).

4. HURDLE REGRESSION

Originally developed by Mullahy (1986), Hurdle regression is also known as two-part model. Instead of assuming that count outcome comes from a single data generating process, Hurdle regression considers count outcome generated by two systematically different statistical processes, a binomial distribution determining if a count outcome is zero or nonzero and a truncated-at-zero distribution for count data governing all positive counts conditional on nonzero outcomes. The attraction of Hurdle regression is that it reflects a two-stage decision-making process in most human behaviors and therefore has an appealing interpretation. For instance, it is patient's decision whether to contact the doctor's office and to make the initial visit. However, after the patient's first visit, doctor plays a more important role in determining if the patient needs to make follow-up visits. Therefore, in a regression setting, the first decision might be reflected by a Logit or Probit regression, while the second one can be analyzed by a truncated Poisson or Negative binomial regression. Moreover, different explanatory variables are allowed to have different impacts at each decision process.

The most popular formulation of a Hurdle regression is called Logit-Poisson model, which is the combination of a Logit regression modeling zero vs. nonzero outcomes and a truncated Poisson regression modeling positive counts conditional on nonzero outcomes. Its probability density function is given as

$$f(Y_i \mid X_i) = \begin{cases} \theta_i & \text{for } Y_i = 0 \\ \dfrac{(1-\theta_i)\cdot Exp(-u_i)\cdot u_i^{Y_i}}{(1-Exp(-u_i))\cdot Y_i!} & \text{for } Y_i > 0 \end{cases}, \text{ where } \theta_i = P(Y_i = 0) \text{ and } u_i = Exp(X_i\beta) \qquad (4.1)$$

The log-likelihood function of a Logit-Poisson regression therefore can be expressed as the sum of log-likelihood functions of two components as below

$$LL = \sum_{i=1}^{n} \begin{bmatrix} I(Y_i = 0)Log(\theta_i) + \\ I(Y_i > 0)\big(Log(1-\theta_i) - u_i + Y_i Log(u_i) - Log(1 - Exp(-u_i)) - Log(Y_i!)\big) \end{bmatrix} \qquad (4.2).$$

Unlike Poisson and Negative binomial regressions, Hurdle regression can only be modeled through log-likelihood function with NLMIXED, MODEL, and NLIN procedures in SAS. For the simplicity, we use the same explanatory variables in both components of Logit-Poisson regression. However, in practice, two sets of explanatory variables do not have to coincide. An example of NLMIXED is given below.

```
/* METHOD 1: PROC NLMIXED  */
proc nlmixed data = tblNMES tech = dbldog;
  parms a0 = 0 a1 = 0 a2 = 0 a3 = 0 a4 = 0 a5 = 0 a6 = 0 a7 = 0
        b0 = 0 b1 = 0 b2 = 0 b3 = 0 b4 = 0 b5 = 0 b6 = 0 b7 = 0;
  eta0 = a0 + a1 * EXCLHLTH + a2 * POORHLTH + a3 * NUMCHRON + a4 * AGE +
         a5 * MALE + a6 * SCHOOL + a7 * PRIVINS;
  exp_eta0 = exp(eta0);
  p0 = exp_eta0 / (1 + exp_eta0);
  etap = b0 + b1 * EXCLHLTH + b2 * POORHLTH + b3 * NUMCHRON + b4 * AGE +
         b5 * MALE + b6 * SCHOOL + b7 * PRIVINS;
  exp_etap = exp(etap);
  if HOSP = 0 then ll = log(p0);
  else ll = log(1 - p0) - exp_etap + HOSP * etap - lgamma(HOSP + 1)
            - log(1 - exp(-exp_etap));
  model HOSP ~ general(ll);
  predict exp_etap out = hdl_out1 (keep = pred HOSP rename = (pred = Yhat));
  predict p0 out = hdl_out2 (keep = pred rename = (pred = p0));
run;

/* SAMPLE OUTPUT OF PROC NLMIXED:
                            Fit Statistics
               -2 Log Likelihood                    5758.4
               AIC (smaller is better)              5790.4
               AICC (smaller is better)             5790.6
               BIC (smaller is better)              5892.7

                         Parameter Estimates
```

Parameter	Estimate	Standard Error	DF	t Value	Pr > |t|	Alpha
a0	4.2311	0.4889	4406	8.65	<.0001	0.05
a1	0.5826	0.1991	4406	2.93	0.0035	0.05
a2	-0.6953	0.1073	4406	-6.48	<.0001	0.05
a3	-0.3078	0.02890	4406	-10.65	<.0001	0.05
a4	-0.2752	0.06061	4406	-4.54	<.0001	0.05
a5	-0.1948	0.08008	4406	-2.43	0.0151	0.05
a6	-0.00593	0.01126	4406	-0.53	0.5982	0.05
a7	-0.01924	0.09944	4406	-0.19	0.8466	0.05
b0	-0.4693	0.5627	4406	-0.83	0.4043	0.05
b1	-0.9422	0.4949	4406	-1.90	0.0570	0.05
b2	0.3373	0.1008	4406	3.35	0.0008	0.05
b3	0.1426	0.02967	4406	4.81	<.0001	0.05
b4	-0.01229	0.06834	4406	-0.18	0.8573	0.05
b5	-0.03854	0.09227	4406	-0.42	0.6762	0.05
b6	-0.01815	0.01290	4406	-1.41	0.1597	0.05
b7	0.2589	0.1139	4406	2.27	0.0231	0.05

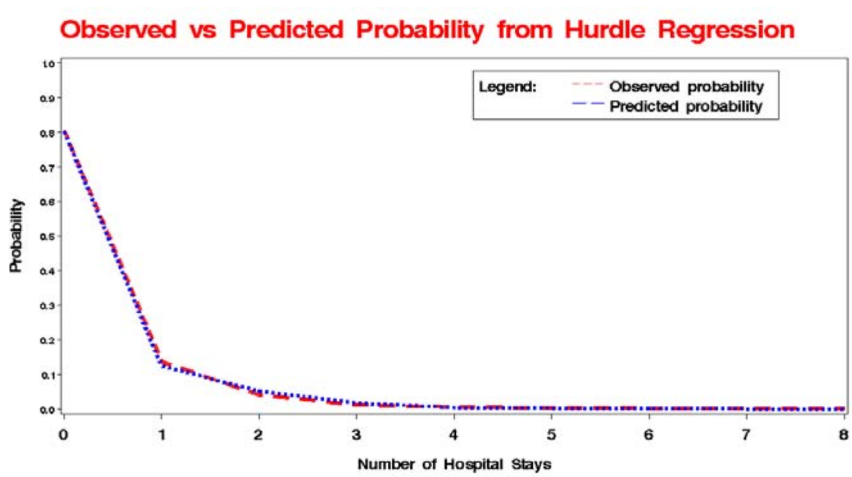
*/

Similarly to Negative binomial regression, Hurdle regression might become Poisson regression with the restriction of parameters and therefore they can be considered nested models. Thus, it is straightforward to use the Likelihood Ratio (LR) test discussed in the previous section to compare Hurdle regression and Poisson regression. In our case, the LR test is given as $-2(LL_{Poisson} - LL_{Hurdle}) = -2[-3046 - (-2879)] = 334$, which is highly significant and suggests that Hurdle regression is preferred to Poisson regression.

While Hurdle regression and Poisson regression are nested, Hurdle regression and Negative binomial regression are not. As a result, the LR test cannot be used to compare these non-nested models. In statistics literature, two methods are generally used to compare non-nested models. The first approach is to use information criteria such as AIC or BIC. However, due to its more parsimonious parameterization and higher log likelihood function, Negative binomial regression is often reported to be favored over Hurdle regression. The second one is to use Vuong test, as proposed by Greene (1994). Since more comprehensive introduction about Vuong test is given in the next section, we will skip the detailed discussion here.

Again, Figure 4.1 visualizes the goodness-of-fit of Hurdle regression, which looks as good as the one provided by Negative binomial regression.

Figure 4.1, Comparison between Observed and Predicted Probability from Hurdle Regression



5. ZERO-INFLATED REGRESSION

Introduced by Lambert (1992), Zero-inflated regression is another way to model count data with excess zeros. Similar to Hurdle regression, Zero-inflated regression can also be considered a mixture of two statistical processes, one always generating zero counts and the other generating both zero and nonzero counts. However, it is slightly different from Hurdle regression with all zero counts from a single statistical process and assumes that zero counts might come from two different sources. More specifically, in a Zero-inflated regression, a Logit model with binomial assumption is used to determine if an individual count outcome is from the always-zero or the not-always-zero group and then a model for count data, either Poisson or Negative binomial, to model outcomes in the not-always-zero group. In the paper, we will limit our discussion to Zero-inflated Poisson (ZIP) regression with formulation

$$Log\left(\frac{\omega_i}{1-\omega_i}\right) = Z_i\gamma \text{ and } Log(u_i) = X_i\beta, \tag{5.1}$$

where Z_i *and* X_i are covariate matrix. However, the same idea can be easily generalized to Zero-inflated Negative Binomial (ZINB) regression.

The density function of a ZIP model is given as

$$f(Y_i \mid X_i) = \begin{cases} \omega_i + (1-\omega_i)\cdot Exp(-u_i) & \text{for } Y_i = 0 \\ (1-\omega_i)\dfrac{Exp(-u_i)\cdot u_i^{Y_i}}{Y_i!} & \text{for } Y_i > 0 \end{cases}, \text{ where } 1-\omega_i = P(Y_i \sim Poisson(u_i)) \tag{5.2}$$

And its log-likelihood function is expressed as

$$LL = \sum_{i=1}^{n}\left[\begin{array}{l} I(Y_i = 0)Log(\omega_i + (1-\omega_i)Exp(-u_i)) + \\ I(Y_i > 0)(Log(1-\omega_i) + Y_i Log(u_i) - u_i - Log(Y_i!)) \end{array}\right] \tag{5.3}$$

In SAS, ZIP can be fitted either through log-likelihood function or directly with experimental COUNTREG procedure in ETS module.

```
/* METHOD 1: PROC COUNTREG */
proc countreg data = tblNMES type = zip;
  model HOSP = EXCLHLTH POORHLTH NUMCHRON AGE MALE SCHOOL PRIVINS
  / zi(link = logistic, var = EXCLHLTH POORHLTH NUMCHRON AGE MALE SCHOOL PRIVINS);
run;

/* METHOD 2: PROC NLMIXED  */
proc nlmixed data = tblNMES tech = dbldog;
  parms a0 = 0 a1 = 0 a2 = 0 a3 = 0 a4 = 0 a5 = 0 a6 = 0 a7 = 0
        b0 = 0 b1 = 0 b2 = 0 b3 = 0 b4 = 0 b5 = 0 b6 = 0 b7 = 0;
  eta0 = a0 + a1 * EXCLHLTH + a2 * POORHLTH + a3 * NUMCHRON + a4 * AGE +
         a5 * MALE + a6 * SCHOOL + a7 * PRIVINS;
  exp_eta0 = exp(eta0);
  p0 = exp_eta0 / (1 + exp_eta0);
  etap = b0 + b1 * EXCLHLTH + b2 * POORHLTH + b3 * NUMCHRON + b4 * AGE +
         b5 * MALE + b6 * SCHOOL + b7 * PRIVINS;
  exp_etap = exp(etap);
  if HOSP = 0 then ll = log(p0 + (1 - p0) * exp(-exp_etap));
  else ll = log(1 - p0) + HOSP * etap - exp_etap - lgamma(HOSP + 1);
  model HOSP ~ general(ll);
  predict exp_etap out = zip_out1 (keep = pred HOSP rename = (pred = Yhat));
  predict p0 out = zip_out2 (keep = pred rename = (pred = p0));
run;

/* SAMPLE OUTPUT OF PROC COUNTREG:
                          Model Fit Summary
               Log Likelihood                    -2878
               AIC                                5788
               SBC                                5890

                         Parameter Estimates
                                      Standard               Approx
 Parameter          Estimate             Error    t Value    Pr > |t|
 Intercept         -0.366506          0.572032      -0.64      0.5217
 exclhlth          -0.919990          0.458460      -2.01      0.0448
 poorhlth           0.324926          0.101157       3.21      0.0013
 numchron           0.127746          0.033867       3.77      0.0002
 age               -0.024359          0.068806      -0.35      0.7233
 male              -0.059629          0.099133      -0.60      0.5475
 school            -0.012473          0.013520      -0.92      0.3562
 privins            0.229208          0.114004       2.01      0.0444
 Inf_Intercept      4.265976          0.971218       4.39      <.0001
 Inf_exclhlth      -0.369944          0.717395      -0.52      0.6061
 Inf_poorhlth      -0.589745          0.195174      -3.02      0.0025
 Inf_numchron      -0.280116          0.062396      -4.49      <.0001
 Inf_age           -0.405962          0.119765      -3.39      0.0007
 Inf_male          -0.334773          0.162429      -2.06      0.0393
 Inf_school        -0.019390          0.022126      -0.88      0.3808
 Inf_privins        0.224859          0.196133       1.15      0.2516
*/
```

Please note that in a ZIP regression, explanatory variables used in two components do not need to be the same. However, when all covariates in both sub-models are identical, ZIP regression can become a more parsimonious model by assuming that the coefficient vector in Logit component are the product between the coefficient vector in Poisson component and a scalar parameter ***τ*** (tau), namely ZIP(*tau*) model. Its formulation can be expressed as

$$Log\left(\frac{\omega_i}{1-\omega_i}\right) = \tau X_i \beta \text{ and } Log(u_i) = X_i \beta . \qquad (5.4)$$

And the estimation of ZIP(*tau*) in SAS is straightforward with NLMIXED procedure given below.

```
/* METHOD 1: PROC NLMIXED  */
proc nlmixed data = tblNMES;
  parms b0 = 0 b1 = 0 b2 = 0 b3 = 0 b4 = 0 b5 = 0 b6 = 0 b7 = 0 tau = 1;
  eta0 = tau * (b0 + b1 * EXCLHLTH + b2 * POORHLTH + b3 * NUMCHRON + b4 * AGE +
                b5 * MALE + b6 * SCHOOL + b7 * PRIVINS);
  exp_eta0 = exp(eta0);
  p0 = exp_eta0 / (1 + exp_eta0);
  etap = b0 + b1 * EXCLHLTH + b2 * POORHLTH + b3 * NUMCHRON + b4 * AGE +
         b5 * MALE + b6 * SCHOOL + b7 * PRIVINS;
  exp_etap = exp(etap);
  if HOSP = 0 then ll = log(p0 + (1 - p0) * exp(-exp_etap));
  else ll = log(1 - p0) + HOSP * etap - exp_etap - lgamma(HOSP + 1);
  model HOSP ~ general(ll);
  predict exp_etap out = zip_out1 (keep = pred HOSP rename = (pred = Yhat));
  predict p0 out = zip_out2 (keep = pred rename = (pred = p0));
run;

/* SAMPLE OUTPUT OF PROC NLMIXED:
                           Fit Statistics
              -2 Log Likelihood                    5768.7
              AIC (smaller is better)              5786.7
              AICC (smaller is better)             5786.7
              BIC (smaller is better)              5844.2

                        Parameter Estimates
                       Standard
 Parameter   Estimate     Error    DF   t Value   Pr > |t|    Alpha
 b0           -1.3944    0.2698  4406     -5.17     <.0001     0.05
 b1           -0.2685   0.09606  4406     -2.80     0.0052     0.05
 b2            0.3223   0.05980  4406      5.39     <.0001     0.05
 b3            0.1391   0.02195  4406      6.34     <.0001     0.05
 b4            0.1040   0.02789  4406      3.73     0.0002     0.05
 b5           0.07254   0.03383  4406      2.14     0.0321     0.05
 b6          -0.00039  0.004641  4406     -0.08     0.9331     0.05
 b7           0.04292   0.04216  4406      1.02     0.3087     0.05
 tau          -1.8406    0.4585  4406     -4.01     <.0001     0.05               */
```

In Figure 5.1 and 5.2 below, plots between observed probability and predicted probability are used to visualize the goodness-of-fit of ZIP and ZIP(tau) regressions. It is clear that both models fit the observed count outcomes as well as Negative binomial and Hurdle regression.

Figure 5.1, Comparison between Observed and Predicted Probability from ZIP Regression

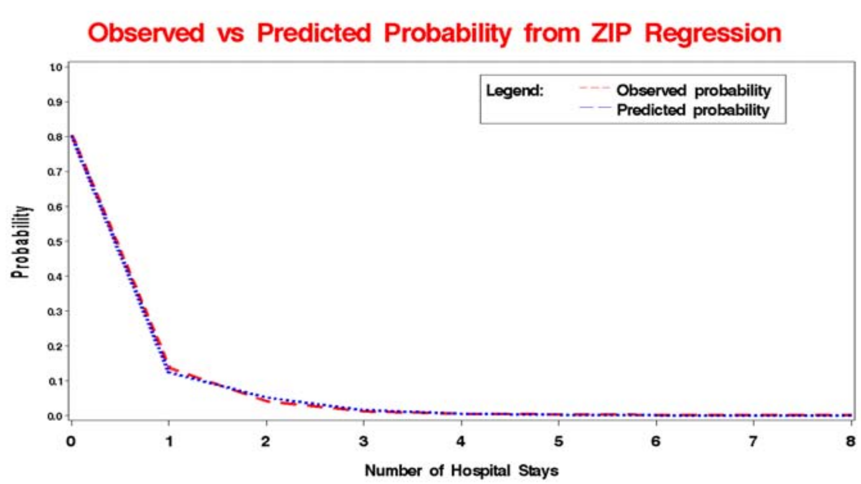
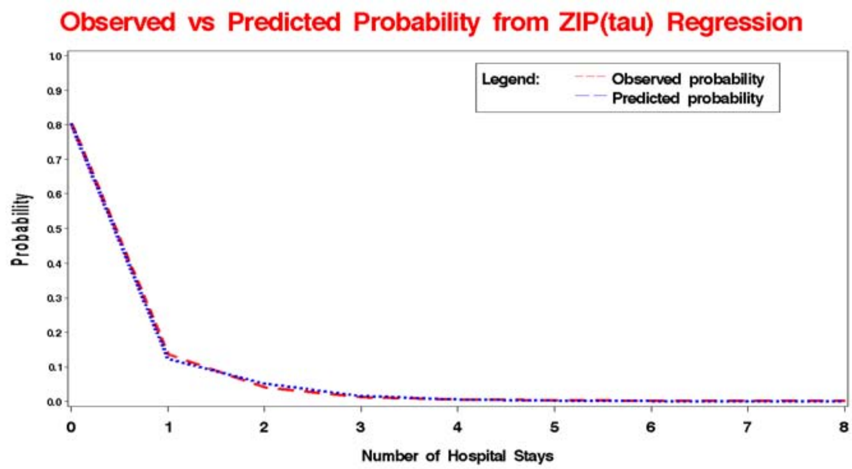


Figure 5.2, Comparison between Observed and Predicted Probability from ZIP(*τ*) Regression



While plotting the prediction can be used as an informal way to assess goodness-of-fit, Vuong test is considered a better method to compare ZIP regression to other non-nested models for count data, such as Poisson regression, Negative Binomial regression, or Hurdle regression. If we define

$$m_i = Log\left(\frac{P_1(Y_i \mid X_i)}{P_2(Y_i \mid X_i)}\right) \qquad (4.3)$$

where $P_N(Y_i|X_i)$ is the predicted probability of observed count for case *i* from model *N*, then Vuong statistic to test the hypothesis $E(m_i = 0)$ is expressed as

$$V = \frac{\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n} m_i\right)}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(m_i - \overline{m})^2}} \qquad (4.4)$$

If $V > 1.96$, the first model is preferred. If $V < -1.96$, then the second one is preferred. SAS implementation of Vuong test to compare ZIP regression to Poisson regression is given below.

```
data poi_pred (keep = poi_prob);
  set poi_out;                        /* OUTPUT FROM POISSON REGRESSION */
  do i = 0 to 8;
    poi_prob = pdf('poisson', i , Yhat);
    if hosp = i then output;
  end;
run;

data zip_pred (keep = zip_prob);
  merge zip_out1 zip_out2;            /* OUTPUT FROM ZIP REGRESSION */
  do i = 0 to 8;
    if i = 0 then zip_prob = p0 + (1 - p0) * pdf('poisson', i, Yhat);
    else zip_prob = (1 - p0) * pdf('poisson', i, Yhat);
    if hosp = i then output;
  end;
run;

data compare;
  merge poi_pred zip_pred;
  m = log(zip_prob / poi_prob);
run;

proc sql;
select
  mean(m)                          as mbar,
  std(m)                           as s,
  sqrt(count(*)) * mean(m) / std(m) as v
from
  compare;
quit;
```

```
/* RESULT OF VUONG TEST:
                              mbar         s       v
                          0.038138  0.375956  6.733444                         */
```

From the above result of Vuong test, it is clearly shown that ZIP regression fit the data better than Poisson regression.

6. MODELS EVALUATION

In previous sections, five models for count data have been built with the healthcare utilization data: Poisson regression, Negative binomial regression, Hurdle regression, ZIP regression, and ZIP(tau) regression. In practice, it is often an interest to compare these models both in statistical sense and in business sense.

Table 6.1, Estimated Coefficients of Four Models

	Poisson Regresson	Neg Bin Regresson	Hurdle		Zero-Inflated		ZIP (tau)
			Logit	Poisson	Logit	Poisson	
INTERCEPT	**-3.3290**	**-3.7526**	**4.2311**	-0.4693	**4.2660**	-0.3665	**-1.3944**
EXCLHLTH	**-0.7234**	**-0.6979**	**0.5826**	-0.9422	-0.3700	**-0.9200**	**-0.2685**
POORHLTH	**0.6262**	**0.6139**	**-0.6953**	**0.3373**	**-0.5897**	**0.3249**	**0.3223**
NUMCHRON	**0.2645**	**0.2894**	**-0.3078**	**0.1426**	**-0.2801**	**0.1277**	**0.1931**
AGE	**0.1864**	**0.2384**	**-0.2752**	-0.0123	**-0.4060**	-0.0243	**0.1040**
MALE	0.1032	**0.1539**	**-0.1948**	-0.0385	**-0.3348**	-0.0596	**0.0725**
SCHOOL	-0.0002	-0.0023	-0.0059	-0.0182	-0.0194	-0.0125	-0.0004
PRIVINS	0.1087	0.0939	-0.0192	**0.2589**	0.2249	**0.2292**	0.0429
alpha		**1.7667**					
tau							**-1.8406**
Log Likelihood	-3046	-2857		-2879		-2878	-2887
# of Parameters	8	9		16		16	9
AIC	6108	5732		5790		5788	5792
BIC	6159	5790		5892		5890	5850
Vuong Test						6.73	6.61

Highlighted Coefficients are significant at 5%.

AIC = -2 * LL + 2 * # of Parameters, BIC = -2 * LL + Log(# of cases) * # of Parameters

Estimated coefficients of all five models together with related statistics are listed in Table 6.1. While Poisson regression provides a baseline model for count data, the other four demonstrate the better fit than the basic Poisson regression. It is interesting to see that although Negative binomial and ZIP(tau) regressions have very different assumption and specification, they all indicate that the information related health status and demographic determines the frequency of hospital admissions but socio-economic variables do not. But if we take a look at Hurdle and ZIP regressions, we should see a different story from these composite models. First of all, the coefficient significance in the Logit component suggests that whether an individual is admitted in the hospital depends on the health status and demographics information. However, the Poisson component indicates that the frequencies of hospital admissions are related to both the heath status and insurance status. A potential interpretation is that if the individual with health problem is covered by the private insurance, he/she might be admitted into a hospital more often than the one without the coverage of private insurance.

If we make the justification of best model solely based on the statistical tests reported in Table 6.1, it is very tempting to conclude that Negative binomial regression out-performs its counterparts for the lower AIC and BIC. On the other hand, composite models such as Hurdle and ZIP regressions provide a greater flexibility of modeling zero outcomes and a more intuitive interpretation. The major difference between Hurdle and ZIP regressions is that the Logit component in Hurdle regression describes the probability of a zero count, while the Logit component in ZIP regression estimates the probability of a zero count from the always-zero or the not-always-zero group. In general, these two models might lead to the similar goodness-of-fit and close interpretation. However, due to the complex parameterization, they share the same limitation and would often suffer from the over-fitting.

7. CONCLUSION

In this paper, we have reviewed several modeling strategies for count data and their implementations in SAS. Basic Poisson models with and without the consideration of observed heterogeneity is a good starting point for count data modeling. For count data with the evidence of over-dispersion, Negative Binomial regression with a more liberal assumption on variance is able to provide a better solution. If the over-dispersion results from a high frequency of zero counts, advanced composite models such as Hurdle regression and ZIP regression might give more satisfactory fit to the data. An example in healthcare utilization has been used in our paper to demonstrate the usage of various models for count data and related statistical tests. However, successfully applications can also be extended to other business problems, such as database marketing, credit risk, and quality control.

REFERENCES

Cameron, A. C. and Trivedi, P. K. (1996), Count Data Models for Financial Data, Handbook of Statistics, Vol. 14, Statistical Methods in Finance, 363-392, Amsterdam, North-Holland.
Cameron, A. C. and Trivedi, P. K. (2001), Essentials of Count Data Regression, A Companion to Theoretical Econometrics, 331-348, Blackwell.
Deb, P. and Trivedi, P. (1997), Demand for Medical Care by the Elderly: A Finite Mixture Approach, Journal of Applied Econometrics, Vol. 12, No. 3, 313-336.
Greene, W. (2002), Econometric Analysis, Prentice Hall.
Greene, W. (1994), Accounting for Excess Zeros and Sample Selection in Poisson and Negative Binomial Regression Models, Working Paper, Department of Economics, New York University
Gurmu, S. (1997), Semi-Parametric Estimation of Hurdle Regression Models With an Application to Medicaid Utilization, Journal of Applied Econometrics, Vol. 12, No. 3, 225-242.
Lambert, D (1992), Zero-Inflated Poisson Regression, with an Application to Defects in Manufacturing, Technometrics, Vol. 34, No. 1, 1 – 14.
Mullahy, J. (1986), Specification and Testing of Some Modified Count Data Models, Journal of Econometrics, 33, 341-365
Winkelmann, R. and Zimmermann, K. F. (1995), Recent Developments in Count Data Modeling: Theory and Application, Theory and Applications, Journal of Economic Surveys, 9, 1-24.
Winkelmann, R. (2004), Health Care Reform and The Number of Doctor Visits – An Econometric Analysis, Journal of Applied Econometrics, 19, 455 - 472.

CONTACT INFORMATION

Wensui Liu, Statistical Project Manager
ChoicePoint Precision Marketing, Alpharetta, GA
Email: wensui.liu@choicepoint.com

Jimmy Cela, AVP
ChoicePoint Precision Marketing, Alpharetta, GA
Email: jimmy.cela@choicepoint.com