A CLASS OF PREDICTIVE MODELS FOR MULTI-LEVEL RISKS

Wensui Liu, JP Morgan Chase
Chuck Vu, Acxiom
Sandeep Kharidhi, Acxiom

ABSTRACT

In the financial service industry, discriminant analysis and its variants based upon binary outcome, such as logistic regression or neural networks, are largely used to develop predictive models. However, the two-state assumption of such models over-simplifies customers' behavioral outcomes and ignores the existence of multi-level risk. In many situations, the financial impact of a certain customer is directly related to the frequency and the severity of his/her adverse behaviors. Therefore, it is of interest to model and predict such multi-level risks. Several modeling techniques, from Poisson to Ordered Logit models, have been widely discussed in numerous research literatures about how to predict the multi-level risks. Our paper is also an attempt contributed to this end. Several modeling strategies together with their SAS implementations and related scoring scheme will be illustrated. Our purpose is to demonstrate an application of these complex statistical models with the business touch and how to implement them in a production environment.

METHODOLOGY

In retail banking and credit card industries, it is a key interest to predict the probability of customer's adverse behaviors, such as delinquencies or defaults. A widely accepted practice in the industry is to classify customers into 2 groups, the good and the bad, based upon the presence of certain adverse behaviors and then to model this binary outcome with discriminant models. For instance, a customer will be classified as the bad if he/she misses payments during a valuation horizon of one year. In SAS community, most efforts contributed to the improvement of these prediction models so far have been focusing on discovering the relationship between the outcome and predictors through either parametric or nonparametric statistical methods. Li (2006) compared discriminant analysis and logistic regression in the credit risk modeling. Liu and Cela (2007) demonstrated how to use Generalized Additive Models to capture the nonlinear relationship in a credit scoring model. However, an obvious limitation of discriminant models based upon the binary outcome is that the two-state classification over-simplifies adverse behaviors of customers. To the best of our knowledge, what financially impact a financial institute are not only the presence of a certain adverse behavior but also the frequency and the severity of such behavior. As a result, it is advantageous to differentiate different levels of the risk and evaluate the probability of each risk level.

In the definition of binary outcome, it is important to notice that customers are classified mainly based upon the adverse behaviors such as delinquencies and defaults, which can also be measured directly as frequencies or counts. Therefore, instead of modeling the binary outcome, we propose that a more sensible alternative is to model the frequency of adverse behaviors by a customer within a given valuation horizon. In the statistical content, a genuine model for count outcome is Poisson regression model with probability function

$$f(Y_i \mid X_i) = \frac{Exp(-\lambda_i) \cdot \lambda_i^{Y_i}}{Y_i!}, \text{ where } \lambda_i = Exp(X_i\beta) \qquad (1.1)$$

It is assumed that each observed count Y_i is drawn from a Poisson distribution with the conditional mean λ_i on a given covariate vector X_i for case i. In Poisson model, a strong assumption is that the mean is equal to the variance such that $E(Y_i|X_i) = Var(Y_i|X_i) = \lambda_i$, which is also known as Equi-Disperson. However, in practice, this Equi-Dispersion assumption is too restrictive for many empirical applications. In the real-world count data, the variance often exceeds the mean, namely Over-Dispersion, due to various reasons such as excess zeroes or long right tail. For instance, in a credit card portfolio, majority of cardholders should have zero delinquency at any point in time, while a few might have more than 10. With the similar consequence of heteroskedasticity in a linear regression, Over-Dispersion in a Poisson model will lead to deflated standard errors of parameter estimates and therefore inflated t-statistics. Therefore, Poisson model is often inadequate and practically unusable.

Considered a generalization of basic Poisson model, Negative Binomial model accommodates Over-Dispersion in data by including a dispersion parameter. In a Negative Binomial model, it is assumed that the conditional mean λ_i of Y_i for case i is determined not only by the observed heterogeneity explained by the covariate vector X_i but also by the unobserved heterogeneity denoted as ε_i that is independent of X_i such that

$$\lambda_i = Exp(X_i\beta + \varepsilon_i) = Exp(X_i\beta) \cdot Exp(\varepsilon_i), \text{ where } Exp(\varepsilon_i) \sim Gamma(\alpha^{-1}, \alpha^{-1}) \qquad (1.2)$$

While there are many variants of Negative Binomial model, the most common one is Negbin 2 model proposed by

Cameron and Trivedi (1966) with probability function

$$f(Y_i \mid X_i) = \frac{\Gamma\left(Y_i + \alpha^{-1}\right)}{\Gamma(Y_i + 1)\cdot\Gamma\left(\alpha^{-1}\right)}\left(\frac{\alpha^{-1}}{\alpha^{-1}+\lambda_i}\right)^{\alpha^{-1}}\left(\frac{\lambda_i}{\alpha^{-1}+\lambda_i}\right), \text{where } \alpha \text{ is dispersion parameter} \qquad (1.3)$$

For Negbin 2 model, its conditional mean $E(Y_i|X_i)$ is still λ_i, while its variance $Var(Y_i|X_i)$ becomes $\lambda_i + \alpha\,\lambda_i^2$. Since both $\lambda_i > 0$ and $\alpha > 0$, the variance must exceed the mean and therefore the issue of Over-Dispersion has been fixed.

One major limitation of standard count data models, such as Poisson and Negative Binomial model, is that all data are assumed to be generated by a single process. However, in many cases, it is more appropriate to assume that data might be governed by two or more processes. For instance, it is believed that the risk driver of the 1st time defaulter might be very different from the one of a person who had defaulted for multiple times. Therefore, we can assume that all data might be generated by multiple processes and the empirical distribution of data can be considered a mixture of multiple components. From the business standpoint, the assumption of multiple components is particularly attractive in that it provides the potential to segment the whole portfolio into two or more sub-groups based upon their adverse behaviors and personal characteristics. In the rest of this section, we would discuss 3 models under the assumption of multiple components, namely Hurdle Poisson (Mullahy 1986), Zero-Inflated Poisson (Lambert 1992), and Latent Class Poisson models (Wedel 1993).

Also known as the two-part model, Hurdle Poisson model assumes that count data come from two systematically different statistical processes, a Binomial distribution determining the probability of zero counts and a Truncated-at-Zero Poisson governing positive outcomes. The probability function can be expressed as

$$f(Y_i \mid X_i) = \begin{cases} \theta_i & \text{for } Y_i = 0 \\ \dfrac{(1-\theta_i)\cdot \mathrm{Exp}(-\lambda_i)\cdot \lambda_i^{Y_i}}{(1-\mathrm{Exp}(-\lambda_i))\cdot Y_i!} & \text{for } Y_i > 0 \end{cases}, \text{where } \theta_i = P(Y_i = 0) \text{ and } \lambda_i = \mathrm{Exp}(X_i\beta) \qquad (1.4)$$

In the modeling framework, the first process of Hurdle model can be analyzed by a Logit model and the second process can be reflected by a Truncated-at-zero Poisson model. It is interesting to notice that the traditional two-state Logit model is actually the sub-model for the 1st process in a Hurdle model. The advantage of Hurdle Poisson Model is that it is so flexible as to efficiently model both Over-Dispersed data with too many zeroes and Under-Dispersed data with too few zeroes. Another major motivation of Hurdle Poisson model is that a customer tends to behave differently after committing an adverse behavior for the first time, which is in line with our observation on human behaviors.

Alike to Hurdle Poisson model, Zero-Inflated Poisson model is another way to model count data with excess zeroes under the assumption of 2 components. However, it is slightly different from Hurdle Poisson model in the sense that it assumes zero counts coming from two different sources, one generating only zero counts and the other generating both zero and nonzero counts. Specifically, a Binomial distribution decides if an individual outcome is from the Always-Zero or the Not-Always-Zero group and then a standard Poisson distribution describes counts in the Not-always-zero group. The probability function of Zero-Inflated Poisson model is given as

$$f(Y_i \mid X_i) = \begin{cases} \omega_i + (1-\omega_i)\cdot \mathrm{Exp}(-\lambda_i) & \text{for } Y_i = 0 \\ (1-\omega_i)\dfrac{\mathrm{Exp}(-\lambda_i)\cdot \lambda_i^{Y_i}}{Y_i!} & \text{for } Y_i > 0 \end{cases}, \text{where } 1-\omega_i = P(Y_i \sim \mathrm{Poisson}(\lambda_i)) \qquad (1.5)$$

With the similar idea to Hurdle Poisson model, Zero-Inflated Poisson model can be represented jointly by two different sub-models as well. A Logit model is used to separate the Always-Zero group from the Not-Always-Zero group and a basic Poisson model is applied to individuals in the Not-always-zero group. From a business prospective, Zero-Inflated Poisson Model describes an important fact that some not-at-risk customers are well established such that they will never financial problems, while the other at-risk ones might have chances to get into troubles during the tough time. Therefore, risk exposures and underlying matrices for customers with same outcomes at zero count might still be differentiable.

In practice, a sharp dichotomization between at-risk group and not-at-risk group might not be realistic. Even a customer with the good financial condition might be exposed to risks in a certain situation. Therefore, it might make sense to split the whole portfolio into a couple segments with different levels of risk-exposure. A Latent Class Poisson model provides such flexibility by assuming that the population of interest is actually a mixture of $S \geq 2$ latent (unobservable) components and each individual is considered a draw from one of the these latent groups. The probability function of a Latent Class Poisson model with S classes can be obtained as

$$f(Y_i \mid X_i) = \sum_{s=1}^{S} p_s \frac{\mathrm{Exp}(-\lambda_{i|s})\cdot \lambda_{i|s}^{Y_i}}{Y_i!}, \text{where } \sum_{s=1}^{S} p_s = 1 \text{ and } 1 > p_s > 0 \qquad (1.6)$$

Each latent component in the mixture is assumed to have a different parameter λ, which will account for the

unobserved heterogeneity in the population. For instance, in the case of S = 2, we can assume that the whole portfolio is actually a mixture between a high risk group and a low risk one. In a Latent Class Poisson model, impacts of predictors are allowed to differ across different latent groups, providing a possibility of more informative and flexible interpretations. For instance, it has long been observed that customers with different risks would have different sensitivities to changes in income and spending.

Besides models we discussed above, it is also worth to point out that the discrete choice model, such as Logit or Probit, has also been widely used to model count data as well. However, such discrete choice model needs to be based upon sequential or ordered instead of multinomial response, namely ordered Logit, which will take the form of

$$\text{Logit}\left(\sum_{k=1}^{K} p_k\right) = \alpha_k + \beta X, \text{ where } \sum_{k=1}^{K+1} p_k = 1 \qquad (1.7)$$

APPLICATION

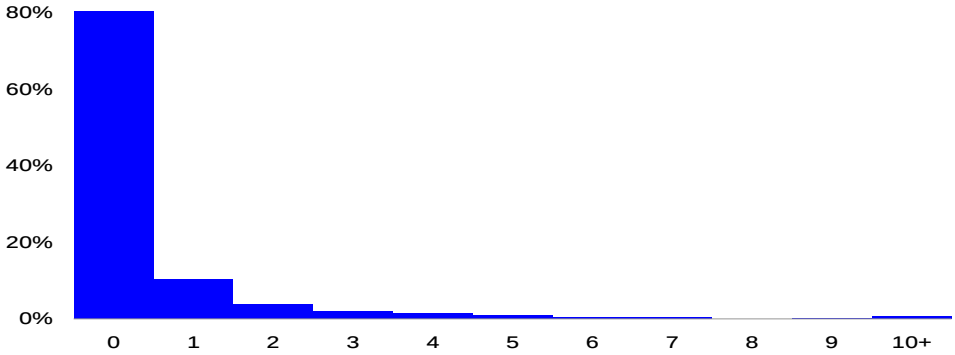
To demonstrate models discussed in the previous section, we used a credit card dataset analyzed in Econometric Analysis (Greene 1992) with the approval by Professor William Greene. This dataset is original obtained from an anonymous credit card vendor and includes 1319 accounts. The outcome variable of our primary interest in this application is the number of major derogatory reports of an individual account. A major derogatory report is defined as a 60-day delinquency in payment to the credit card account. In the model development sample, there are 11 explanatory variables. A summary of both outcome and explanatory variables is given below.

Table 3.1, Variable Summary

Variable	Desc.	Mean	Std.Dev.	Min	Max.
MAJORDRG	Major Derogatory Reports	0.46	1.35	0.00	14.00
ACTIVE	Number of active credit card accounts	7.00	6.31	0.00	46.00
AGE	Age in years as of November, 1989	33.21	10.14	0.17	83.50
AVGEXP	Average monthly credit card expense	185.06	272.22	0.00	3100.00
CUR_ADD	Number of months living at current address	55.27	66.27	0.00	540.00
DEPNDT	Number of dependents	0.99	1.25	0.00	6.00
EXP_INC	Average monthly credit card expense/Average monthly income	0.07	0.09	0.00	0.91
INC_PER	Monthly income divided by 1 + DEPNDT	2.16	1.36	0.07	11.00
INCOME	Self reported income, in $10,000s	3.37	1.69	0.21	13.50
MAJOR	Binary indicator of whether applicant has a major credit card	0.82	0.39	0.00	1.00
OWNRENT	Binary indicator of whether applicant owns their home	0.44	0.50	0.00	1.00
SELFEMPL	Binary indicator of whether the applicant is self-employed	0.07	0.25	0.00	1.00

It is clear that the variance of our dependent variable, MAJORDRG, is about four times as much as the mean, a strong indication of Over-Dispersion. In this situation, it is always helpful to do a more careful explanatory data analysis (EDA) on the dependent variable.

Figure 3.1, Distribution of Derogatory Reports for the Portfolio



In Figure 3.1, the distribution of MAJORDRG shows that more than 80% cardholders in the portfolio have zero derogatory report, while the worst accounts could have more than 10. In this case, it is evident to us that the basic Poisson Model is not able to provide a sufficient fit for the data and therefore is not worth our time for further discussion in this paper. For more details of statistical tests for Over-Dispersion in a Poisson model and related implementations in SAS, please refer to Liu and Cela (2008)

As introduced in the previous section, Negative Binomial Model is a major alternative to accommodate Over-Dispersion in count outcomes. Based upon its probability function, it is easy to derived the log likelihood function as

$$LL = \sum_{i=1}^{n}\left[\text{Log}\left(\frac{\Gamma\left(Y_i + \alpha^{-1}\right)}{\Gamma\left(Y_i + 1\right)\cdot\Gamma\left(\alpha^{-1}\right)}\right) - \left(Y_i + \alpha^{-1}\right)\cdot\text{Log}\left(1 + \alpha\lambda_i\right) + Y_i\text{Log}\left(\alpha\lambda_i\right)\right] \qquad (2.1)$$

In SAS/STAT, NLMIXED procedure provides a flexible way to develop a model using its log likelihood function directly.

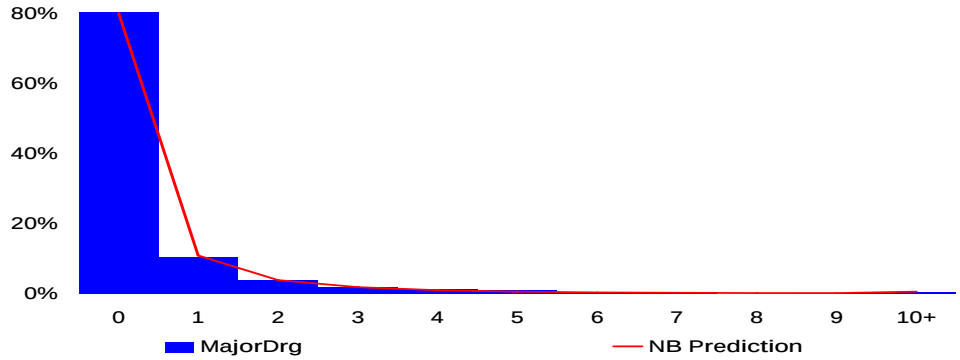
Demo 3.1, Modeling and Scoring Code of Negative Binomial Model

```
/* STEP 1: MODEL DEVELOPMENT */
proc nlmixed data = credit;
  parms B_Intercept = -.8908 B_Age     = -.0002 B_Income   = -.1269 B_Exp_inc = -16.9391
        B_Avgexp    = 0.0012 B_Ownrent = -.7568 B_Selfempl = -.0857 B_Depndt  = 0.2089
        B_Inc_per   = 0.1565 B_Cur_add = 0.0025 B_Major    = -.0021 B_Active  = 0.0776;
  mu = exp(B_Intercept + B_Age * Age + B_Income * Income + B_Exp_inc * Exp_inc +
           B_Avgexp * Avgexp + B_Ownrent * Ownrent + B_Selfempl * Selfempl +
           B_Depndt * Depndt + B_Inc_per * Inc_per + B_Cur_add * Cur_add +
           B_Major * Major + B_Active * Active);
  ll = lgamma(MajorDrg + 1 / alpha) - lgamma(MajorDrg + 1) - lgamma(1 / alpha) +
       MajorDrg * log(alpha * mu) - (MajorDrg + 1 / alpha) * log(1 + alpha * mu);
  model MajorDrg ~ general(ll);
  predict mu out = nb_out1 (rename = (pred = Yhat));
run;

/* STEP 2: CALCULATE PROBABILITY AT EACH LEVEL OF COUNT OUTCOMES */
data nb_out2;
  set nb_out1;
  do count = 0 to 14;
    prob = pdf('negbinomial', count , (1 / 3.5161) / (Yhat + (1 / 3.5161)), (1 / 3.5161));
    output;
  end;
run;

/* STEP 3: PREDICT PORTFOLIO FIT */
proc summary data = nb_out2 nway;
  class count;
  output out = nb_sum(drop = _freq_ _type_) mean(prob) =;
run;

/* STEP 4: PREDICT INDIVIDUAL SCORES */
proc sort data = nb_out2 sortsize = max; by id MajorDrg; run;

proc transpose data = nb_out2 out = nb_out3(drop = _name_) prefix = prob_;
  by id MajorDrg;
  id count; var prob;
```

```
run;

data nb_out4;
  set nb_out3;
  prob_1plus = 1 - prob_0;                      /* score for majordrg >= 1 */
  prob_2plus = 1 - sum(prob_0, prob_1);         /* score for majordrg >= 2 */
  prob_3plus = 1 - sum(prob_0, prob_1, prob_2); /* score for majordrg >= 3 */
run;
```

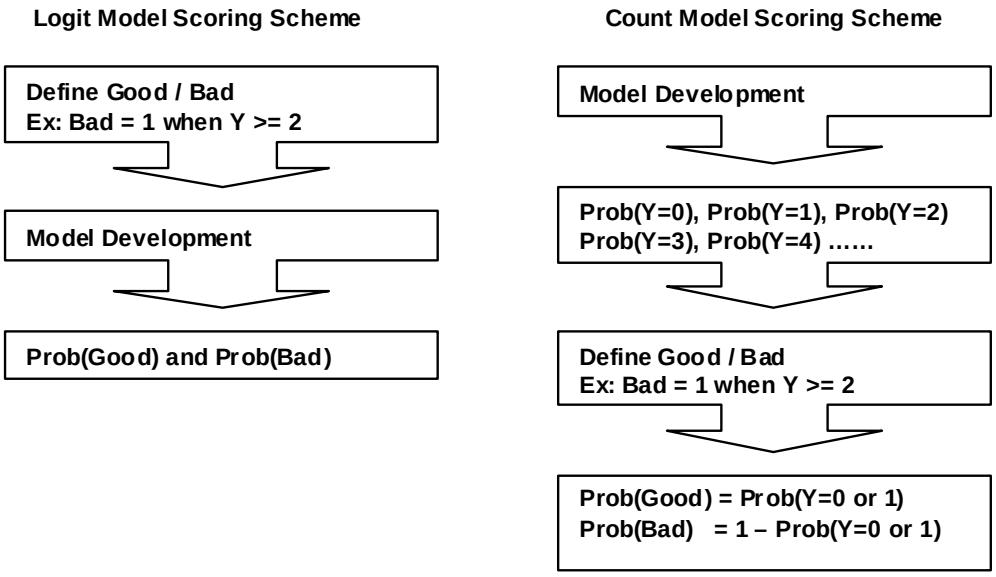
After the model development, the goodness-of-fit of a count data model should be assessed at both the portfolio level and the account level. At the portfolio level, we need to exam if the predicted count outcomes are consistent with the observed ones, as shown in Figure 3.2 below. It is clear that Negative Binomial model is able to provide a sufficient fit for data at the overall portfolio level.

Figure 3.2, Prediction of Negative Binomial Model at Portfolio Level



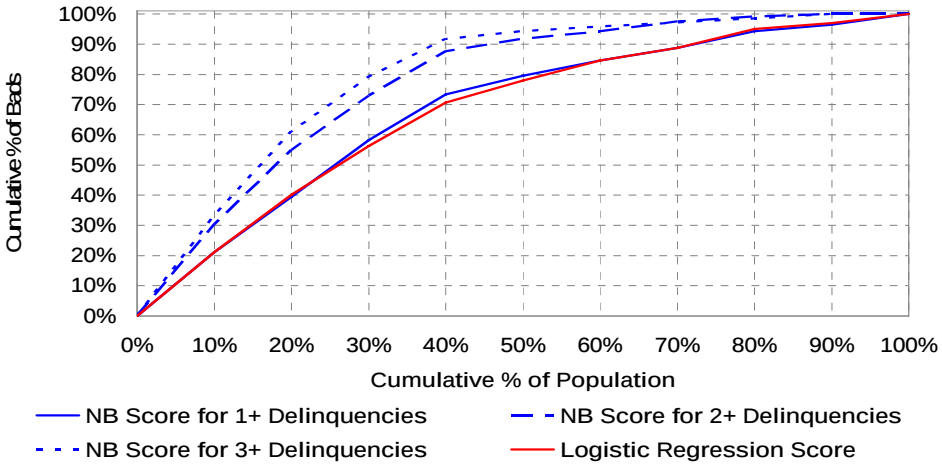
Similar to Logistic regression, Negative Binomial model is also able to provide the probability-based score(s) at the individual account level. However, the scoring scheme of a count model, such as Poisson or its variants, is different from the one of a Logistic regression. Instead of creating only a single score, a count model is can generate a set of multiple scores for different levels of risks.

Figure 3.3, Scoring Scheme of a Model for Count Data



Based upon the scoring scheme in shown Figure 3.3, we can calculate a set of multiple probability-based scores, as shown at STEP 4 in DEMO 3.1. The predictiveness of these scores can also be visualized by a standard Gain chart or Lift chart.

Figure 3.4, Gain Chart of Negative Binomial Model



In the above Gain chart, we use the score from a Logistic regression predicting the presence of delinquency (Presence = 1 for MajorDrg > 0) as the baseline line and compare it with a set of 3 scores from Negative Binomial model. It is suggested that Negative Binomial model performs comparably well relative to Logistic regression when predicting the presence of delinquency. What's more, Negative Binomial model is able to provide more information to conduct the so-called Cherry-Picking for accounts with higher risks, such as cardholders with 3 or more delinquencies.

In contrast to Negative Binomial model handling Over-Dispersion in general, a Hurdle model specifically addresses the issue of excess zero with its two-part nature and is able to do a better separation between zero and nonzero outcomes. The log likelihood function of a Hurdle model can be expressed the sum of log likelihood functions of two sub-models as below

$$LL = \sum_{i=1}^{n} \left[I(Y_i = 0) \cdot \mathrm{Log}(\theta_i) + I(Y_i > 0) \cdot \left(\mathrm{Log}(1 - \theta_i) - \lambda_i + Y_i \mathrm{Log}(\lambda_i) - \mathrm{Log}(1 - \mathrm{Exp}(-\lambda_i)) - \mathrm{Log}(Y_i!) \right) \right] \qquad (2.2)$$

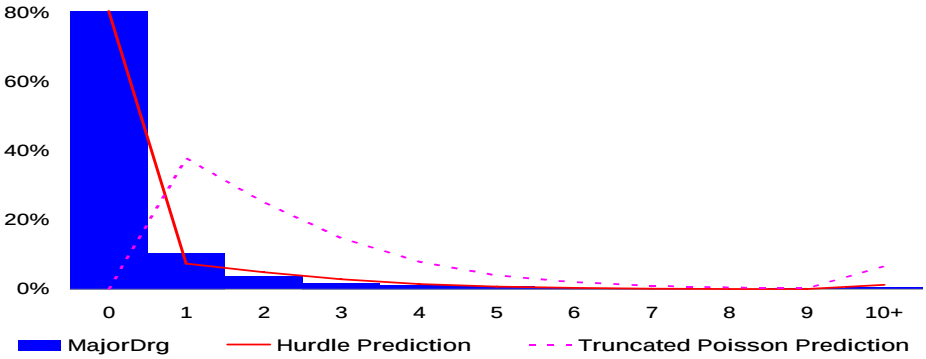
Demo 3.2, Modeling and Scoring Code of Hurdle Model

```
/* STEP 1: MODEL DEVELOPMENT */
proc nlmixed data = dcredit tech = dbldog;
  parms B1_Intercept = 1.9160 B1_Age     = -.0040 B1_Income   = 0.0053 B1_Exp_inc = 6.7049
        B1_Avgexp    = -.0004 B1_Ownrent = 0.7138 B1_Selfempl = -.0648 B1_Depndt  = -.0668
        B1_Inc_per   = -.0439 B1_Cur_add = -.0035 B1_Major    = 0.1762 B1_Active  = -.0953
        B2_Intercept = 0.6699 B2_Age     = -.0027 B2_Income   = -.1008 B2_Exp_inc = -2.5153
        B2_Avgexp    = -.0006 B2_Ownrent = -.3035 B2_Selfempl = -.0212 B2_Depndt  = 0.1744
        B2_Inc_per   = 0.1456 B2_Cur_add = 0      B2_Major    = 0.1283 B2_Active  = 0.0270;
  eta1 = B1_Intercept + B1_Age * Age + B1_Income * Income + B1_Exp_inc * Exp_inc +
         B1_Avgexp * Avgexp + B1_Ownrent * Ownrent + B1_Selfempl * Selfempl +
         B1_Depndt * Depndt + B1_Inc_per * Inc_per + B1_Cur_add * Cur_add +
         B1_Major * Major + B1_Active * Active;
  exp_eta1 = exp(eta1);
  p0 = exp_eta1 / (1 + exp_eta1);
  eta2 = B2_Intercept + B2_Age * Age + B2_Income * Income + B2_Exp_inc * Exp_inc +
         B2_Avgexp * Avgexp + B2_Ownrent * Ownrent + B2_Selfempl * Selfempl +
         B2_Depndt * Depndt + B2_Inc_per * Inc_per + B2_Cur_add * Cur_add +
         B2_Major * Major + B2_Active * Active;
  exp_eta2 = exp(eta2);
  if MajorDrg = 0 then LL = log(p0);
  else LL = log(1 - p0) - exp_eta2 + MajorDrg * eta2 - lgamma(MajorDrg + 1)
            - log(1 - exp(- exp_eta2));
  model Major ~ general(LL);
  predict exp_eta2 out = hdl1 (keep = id pred MajorDrg rename = (pred = Yhat));
  predict p0 out = hdl2 (keep = id pred rename = (pred = p_0));
run;

/* STEP 2: CALCULATE PROBABILITY AT EACH LEVEL OF COUNT OUTCOMES */
proc sort data = hdl1; by id; run;

proc sort data = hdl2; by id; run;
```

```
data hdl_out;
  merge hdl1 hdl2; by id;
  do count = 0 to 14;
    if count = 0 then prob = p_0;
    else prob = (1 - p_0) * pdf('poisson', count, Yhat) / (1 - pdf('poisson', count, Yhat));
    output;
  end;
run;

/* ... ... The rest is the same is in DEMO 3.1 */
```

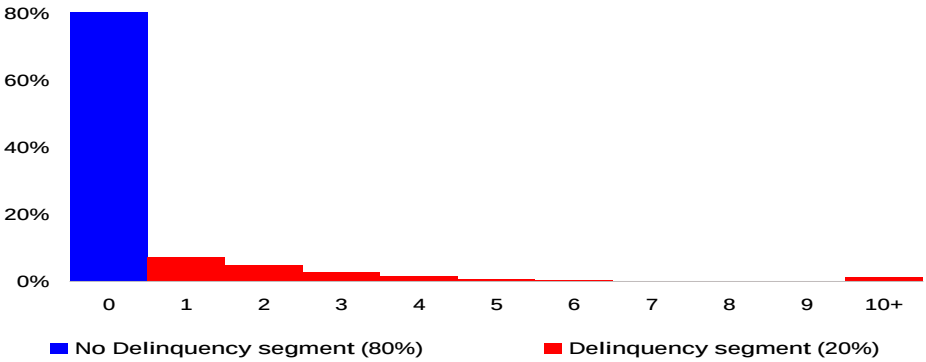
Again, in order to evaluate the goodness-of-fit of a Hurdle model, we show a graph comparing the observed outcome with the predicted at the portfolio level in Figure 3.5. A probability plot for the component of Zero-Truncated Poisson is also included, of which the probability of zero counts is equal to zero by definition. However, since accounts with positive outcomes only count for 20% of the whole portfolio, the actual probability

Figure 3.5, Prediction of Hurdle Model at Portfolio Level



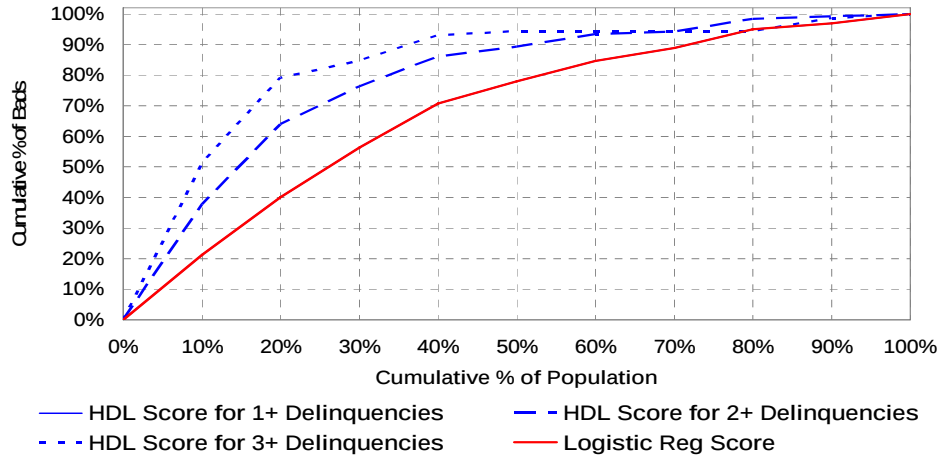
One of the most important and attractive advantages of a composite model is that it provides the possibility to segment customers based upon their behavioral outcomes and personal characteristics. In our case, Hurdle model divides the whole portfolio into two parts, 80% accounts predicted to have no delinquency and 20% at least one, which is shown in Figure 3.6. A practical consideration of such segmentation is that since the underlying motivation and risk driver of each segment could vary from each other, different sets of risk matrices might be necessary in order to explain different types of customers.

Figure 3.6, Portfolio Segmentation of Hurdle Model



With the similar idea shown in Figure 3.3 and 3.4, we can also plot the Gain chart of a Hurdle Model in Figure 3.7. As expected, Hurdle model and Logistic regression perform identically when predicting the presence of delinquency. However, it is worth noting that Hurdle model does a much better job in “cherry-picking” for high-risk accounts than Negative Binomial model.

Figure 3.7, Gain Chart of Hurdle Model



Slightly different from Hurdle model, Zero-Inflated Poisson model assumes two sources of zeroes instead of one. Its log likelihood function and related SAS code are given below.

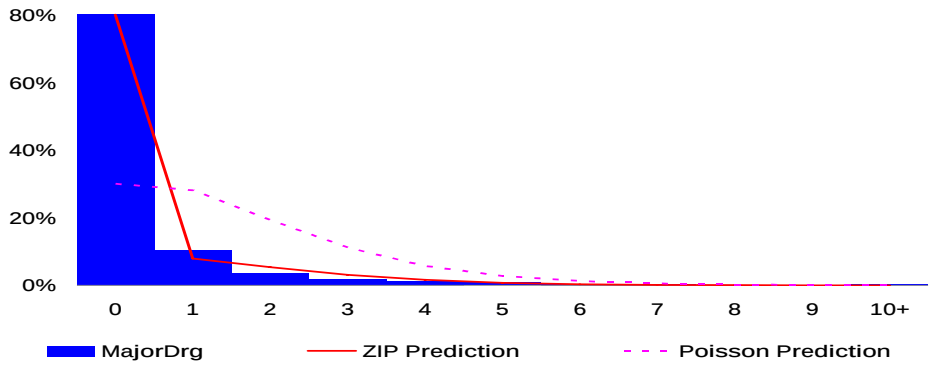
$$LL = \sum_{i=1}^{n} \left[I(Y_i = 0) \cdot Log(\omega_i + (1-\omega_i) \cdot Exp(-\lambda_i)) + I(Y_i > 0) \cdot (Log(1-\omega_i) + Y_i Log(\lambda_i) - \lambda_i - Log(Y_i!)) \right] \qquad (2.3)$$

Demo 3.3, Modeling and Scoring Code of Zero-Inflated Poisson Model

```
/* STEP 1: MODEL DEVELOPMENT */
proc nlmixed data = dcredit tech = dbldog;
  parms B1_Intercept = 1.9160 B1_Age     = -.0040 B1_Income   = 0.0053 B1_Exp_inc = 6.7049
        B1_Avgexp    = -.0004 B1_Ownrent = 0.7138 B1_Selfempl = -.0648 B1_Depndt  = -.0668
        B1_Inc_per   = -.0439 B1_Cur_add = -.0035 B1_Major    = 0.1762 B1_Active  = -.0953
        B2_Intercept = 0.6699 B2_Age     = -.0027 B2_Income   = -.1008 B2_Exp_inc = -2.5157
        B2_Avgexp    = -.0006 B2_Ownrent = -.3035 B2_Selfempl = -.0212 B2_Depndt  = 0.1744
        B2_Inc_per   = 0.1456 B2_Cur_add = 0      B2_Major    = 0.1283 B2_Active  = 0.0270;
  eta1 = B1_Intercept + B1_Age * Age + B1_Income * Income + B1_Exp_inc * Exp_inc +
         B1_Avgexp * Avgexp + B1_Ownrent * Ownrent + B1_Selfempl * Selfempl +
         B1_Depndt * Depndt + B1_Inc_per * Inc_per + B1_Cur_add * Cur_add +
         B1_Major * Major + B1_Active * Active;
  exp_eta1 = exp(eta1);
  p0 = exp_eta1 / (1 + exp_eta1);
  eta2 = B2_Intercept + B2_Age * Age + B2_Income * Income + B2_Exp_inc * Exp_inc +
         B2_Avgexp * Avgexp + B2_Ownrent * Ownrent + B2_Selfempl * Selfempl +
         B2_Depndt * Depndt + B2_Inc_per * Inc_per + B2_Cur_add * Cur_add +
         B2_Major * Major + B2_Active * Active;
  exp_eta2 = exp(eta2);
  if MajorDrg = 0 then LL = log(p0 + (1 - p0) * exp(-exp_eta2));
  else LL = log(1 - p0) + MajorDrg * eta2 - exp_eta2 - lgamma(MajorDrg + 1);
  model MajorDrg ~ general(LL);
  predict exp_eta2 out = zip1 (keep = id pred MajorDrg rename = (pred = Yhat));
  predict p0 out = zip2 (keep = id pred rename = (pred = p_0));
run;

/* STEP 2: CALCULATE PROBABILITY AT EACH LEVEL OF COUNT OUTCOMES */
proc sort data = zip1; by id; run;

proc sort data = zip2; by id; run;

data zip_out;
  merge zip1 zip2; by id;
  do count = 0 to 14;
    if count = 0 then prob = p_0 + (1 - p_0) * pdf('poisson', 0, Yhat);
    else prob = (1 - p_0) * pdf('poisson', count, Yhat);
    output;
  end;
run;

/* ... ... The rest is the same is in DEMO 3.1 */
```

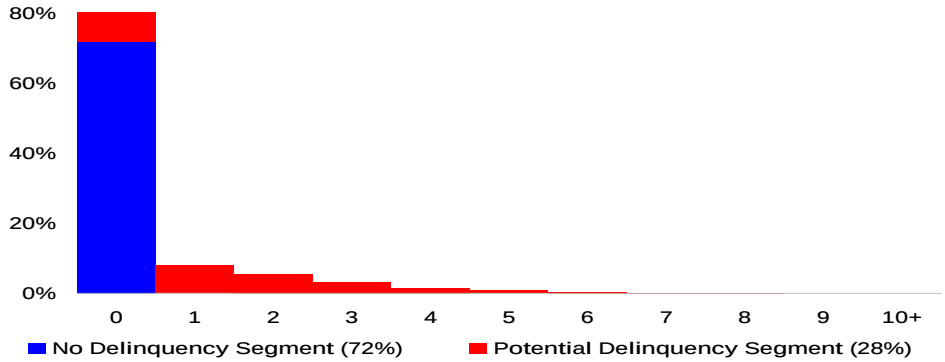
The goodness-of-fit of Zero-Inflated Poisson at the portfolio level is visualized in Figure 3.8 by comparing probabilities of observed outcomes and predicted outcomes. In addition, the probability plot for the component of standard Poisson is also provided, showing a non-zero probability of zero outcomes, which is contrary to zero probability of zero counts in Figure 3.5.

Figure 3.8, Prediction of Zero-Inflated Poisson Model at Portfolio Level



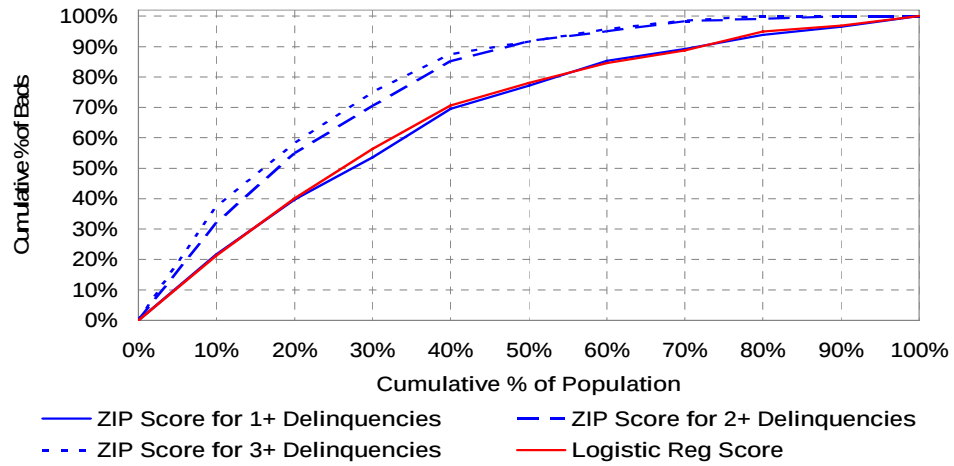
Zero-Inflated Poisson model also divides the whole portfolio into two segments in a similar way to Hurdle model but with a slight difference, 72% accounts predicted to have no risk exposure at all and 28% exposed to a certain degree of risks. In Figure 3.9, it is clear that out of 80% customers without the delinquency, 72% are fallen into the no-risk segment and 8% into the at-risk segment. A business implication is that even for customers with the same outcome of zero counts, we might still be able to conduct a further differentiation for different segments.

Figure 3.9, Portfolio Segmentation of Zero-Inflated Poisson Model



The gain chart of Zero-Inflated Poisson model is presented in Figure 3.10 below to show the “Cherry-Picking” performance at the individual account level. Similar to two previous models, Zero-Inflated Poisson model is able to deliver similar results as Logistic regression does when predicting the presence of delinquency. At higher risk levels, Zero-Inflated Poisson model performs as well as Negative Binomial model but less satisfactorily than Hurdle model.

Figure 3.10, Gain Chart of Zero-Inflated Poisson Model



Considered a more general case of Zero-Inflated Poisson model, Latent Class Poisson model assumes that all count outcomes instead of just zeroes are drawn from S ≥ 2 latent classes. Therefore, the whole population can be thought of a mixture of S Poisson components with different parameters. The log likelihood function of Latent Class Poisson model can be obtained by

$$LL = \sum_{i=1}^{n} \sum_{s=1}^{S} \left[Log(p_s) - \lambda_{i|s} + Y_i Log(\lambda_{i|s}) - Log(Y_i!) \right] \qquad (2.4)$$

While number of latent classes usually can be determined by statistics such as AIC or BIC, S = 2 is strongly supported with the lowest AIC and BIC in our example. A Latent Class Poisson model with S = 2 can be estimated with NLMIXED procedure, as shown in DEMO 3.4 below.

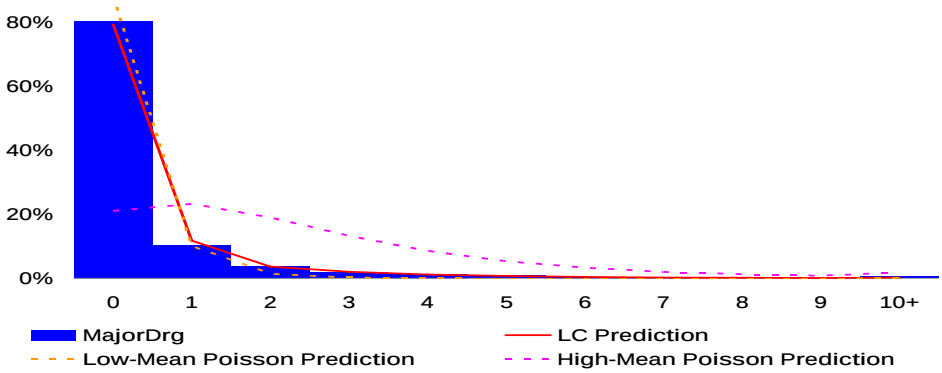
Demo 3.4, Modeling and Scoring Code of Latent Class Poisson Model

```
/* STEP 1: MODEL DEVELOPMENT */
proc nlmixed data = credit tech = dbldog;
  parms B1_Intercept = -1.2152 B1_Age     = -.0080 B1_Income   = -.2088 B1_Exp_inc = -20.4684
        B1_Avgexp    = 0.0003  B1_Ownrent = -.9261 B1_Selfempl = -.3326 B1_Depndt  = 0.1064
        B1_Inc_per   = 0.0539  B1_Cur_add = 0.0015 B1_Major    = -.1755 B1_Active  = 0.0699
        B2_Intercept = -.5664  B2_Age     = 0.0075 B2_Income   = -.0451 B2_Exp_inc = -13.4099
        B2_Avgexp    = 0.0022  B2_Ownrent = -.5874 B2_Selfempl = 0.1611 B2_Depndt  = 0.3115
        B2_Inc_per   = 0.2590  B2_Cur_add = 0.0035 B2_Major    = 0.1712 B2_Active  = 0.0853
        prior1 = 0 to 1 by 0.1;
  eta1 = B1_Intercept + B1_Age * Age + B1_Income * Income + B1_Exp_inc * Exp_inc +
         B1_Avgexp * Avgexp + B1_Ownrent * Ownrent + B1_Selfempl * Selfempl +
         B1_Depndt * Depndt + B1_Inc_per * Inc_per + B1_Cur_add * Cur_add +
         B1_Major * Major + B1_Active * Active;
  exp_eta1 = exp(eta1);
  p1 = exp(- exp_eta1) * exp_eta1 ** MajorDrg / fact(MajorDrg);
  eta2 = B2_Intercept + B2_Age * Age + B2_Income * Income + B2_Exp_inc * Exp_inc +
         B2_Avgexp * Avgexp + B2_Ownrent * Ownrent + B2_Selfempl * Selfempl +
         B2_Depndt * Depndt + B2_Inc_per * Inc_per + B2_Cur_add * Cur_add +
         B2_Major * Major + B2_Active * Active;
  exp_eta2 = exp(eta2);
  p2 = exp(- exp_eta2) * exp_eta2 ** MajorDrg / fact(MajorDrg);
  p = prior1 * p1 + (1 - prior1) * p2;
  LL = log(p);
  model MajorDrg ~ general(LL);
  predict exp_eta1 out = LC1 (keep = id pred MajorDrg rename = (pred = Yhat1));
  predict exp_eta2 out = LC2 (keep = id pred rename = (pred = Yhat2));
run;

/* STEP 2: CALCULATE PROBABILITY AT EACH LEVEL OF COUNT OUTCOMES */
proc sort data = LC1; by id; run;

proc sort data = LC2; by id; run;

data LC_out;
  merge LC1 LC2; by id;
```

```
  prior1 = 0.8698;
  do count = 0 to 14;
    prob_LC1 = pdf('poisson', count, Yhat1);
    prob_LC2 = pdf('poisson', count, Yhat2);
    prob = prob_LC1 * prior1 + prob_LC2 * (1 - prior1);
    output;
  end;
run;

/* ... ... The rest is the same is in DEMO 3.1 */
```

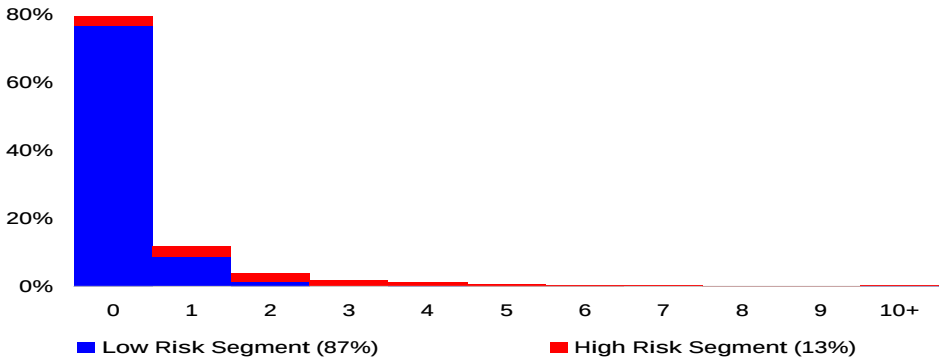
Figure 3.11 presents the goodness-of-fit of Latent Class Poisson model at the portfolio level. While the solid line is the probability plot of the mixture distribution, dotted ones are probability plots of two Poisson components, the segment with low mean and low variance and the other with high mean and high variance.

Figure 3.11, Prediction of Latent Class Poisson Model at Portfolio Level



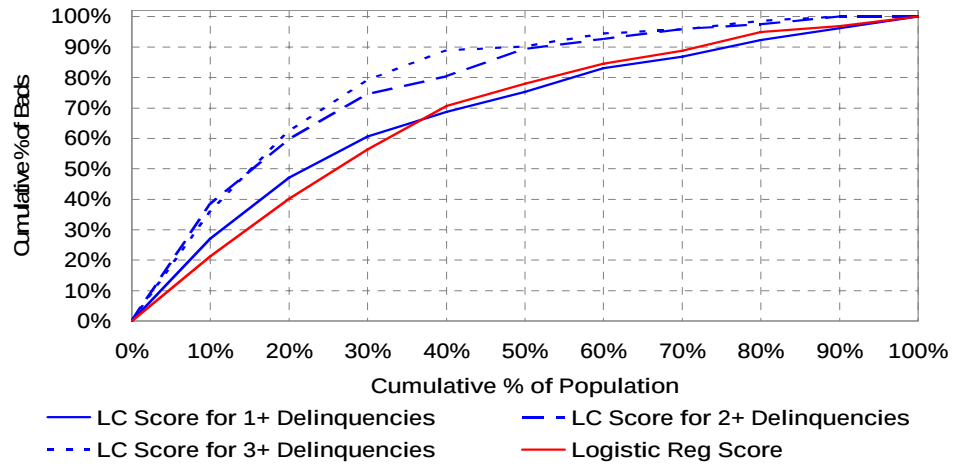
In Figure 3.12, we show how each identified latent component is presented in different level of count outcomes. It is interesting that the component in blue is overwhelming in low counts and the one in red has a higher density on the right tail. In our further analysis, we also notice that 91% of customers with 2 or less delinquencies are fallen into the segment with low mean and low variance and 91% of customers with 3 or more delinquencies into the segment with high mean and high variance. As a result, we can profile one as low-risk group and the other as high-risk group.

Figure 3.12, Portfolio Segmentation of Latent Class Poisson Model



The model predictiveness at account level is shown in Figure 3.13. Unlike 3 models discussed previously, our Latent Class Poisson model demonstrates a superior performance than the baseline Logistic regression in the first three deciles, a critical region for predictive models, when predicting the presence of delinquency.

Figure 3.13, Gain Chart of Latent Class Poisson Model



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, ZIP regression and Latent Class regression might give more satisfactory fit to the data. An example in credit risk assessment 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, healthcare utilization, 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

Your comments and questions are valued and encouraged. Contact the authors at:

Name:	Wensui Liu
Enterprise:	JP Morgan Chase
Phone:	(513) 295-4370
E-mail:	liuwensui@gmail.com
Web:	http://statcompute.spaces.live.com/

Name:	Chuck Vu
Enterprise:	Acxiom Corporation
Address:	1105 Lakewood Parkway, Suite 100
City, State ZIP:	Alpharetta, GA 30009
Phone:	(678) 537-6029
Fax:	(678) 537-6070
E-mail:	chuck.vu@acxiom.com
Web:	http://www.acxiom.com

Name:	Sandeep Kharidhi
Enterprise:	Acxiom Corporation
Address:	1105 Lakewood Parkway, Suite 100
City, State ZIP:	Alpharetta, GA 30009
Phone:	(678) 537-6013
Fax:	(678) 537-6070
E-mail:	sandeep.kharidhi@acxiom.com
Web:	http://www.acxiom.com