Paper 3600-2015

When Two Are Better Than One: Fitting Two-Part Models Using SAS®

Laura Ring Kapitula, Grand Valley State University

ABSTRACT

In many situations, an outcome of interest has a large number of zero outcomes and a group of nonzero outcomes that are discrete or highly skewed. For example, in modeling health care costs, some patients have zero costs, and the distribution of positive costs are often extremely right-skewed. When modeling charitable donations, many potential donors give nothing, and the majority of donations are relatively small with a few very large donors. In the analysis of count data, there are also times where there are more zeros than would be expected using standard methodology, or cases where the zeros might differ substantially than the non-zeros, such as number of cavities a patient has at a dentist appointment or number of children born to a mother. If data has such structure, and ordinary least squares methods are used, then predictions and estimation might be inaccurate. The two-part (two-stage) model gives us a flexible and useful modeling framework in many situations. Methods for fitting the models with SAS® software are illustrated.

INTRODUCTION

There are many situations in which an outcome of interest has a large number of zero outcomes and a group of nonzero outcomes that are discrete or highly skewed. For example, this situation may arise in modeling health care costs, modeling charitable donations and modeling counts. A useful modeling framework in such cases is a two-part model (also known as a two-stage model). In this paper we will give the basic structure of two-part models and illustrate how to fit them using the SAS software system. We note that SAS Enterprise Miner can be used to fit two-stage models but all our examples illustrate using SAS STAT.

The two-stage cost model typically uses logistic regression to model the probability of a positive cost and models the distribution of positive costs using a log-normal regression model. We will explain this model and then illustrate the use of this model with donations and with a problem involving the estimation of health care costs. Another type of two-part model is the zero inflated Poisson regression model that uses logistic regression to model the probability of a positive count and models the distribution of positive counts using a truncated Poisson distribution. An alternative model for count data with extra zeros is the zero-inflated negative binomial regression model. We will illustrate the zero-inflated Poisson and negative binomial model using data on the decayed, missing and filled teeth index collected on school children in Brazil.

THE GENERAL STRUCTURE OF THE TWO-PART MODEL

Two-part models are understandable, intuitive, flexible, useful in many settings and easy to estimate using either *SAS® STAT* or *SAS® Enterprise Miner*. The basic structure of a two-part model is that using the entire data set (or a randomly selected training data set) the probability a target is positive is estimated and then conditional on the target being positive the value of the target is modeled using another model or method. More specifically, suppose we have an outcome or target variable we will call Y. Then we can model, $E(Y)$, the mean or expected value of Y as:

$$E(Y) = P(Y > 0)E(Y|Y > 0) + P(Y = 0)E(Y|Y = 0) = P(Y > 0)E(Y|Y > 0)$$

$P(Y > 0)$ is typically modeled using a generalized linear model such as:

$$g\big(P(Y > 0)\big) = M't$$

Where g is a link function, M is a vector of covariates(predictors), and τ is a parameter vector. It is possible to predict Y being positive using other methods that are appropriate for binary outcomes such as decision trees or neural networks as well.

The positive costs, $|Y > 0 = Y_+$, is then modeled using another model or method. To model, Y_+ we need a model that is strictly positive and allows for positive skewness and possibly non-constant variance. Another generalized linear-model with potentially non-constant variance can be used for this purpose. Let

$$E(f(Y_+)) = x'\beta$$

and

$$Var\big(f(Y_+)\big) = h(w'\theta);$$

where f is a strictly positive valued link function, h is a function with strictly non-negative values, x and w are vectors of covariates(predictors), and β and θ are parameter vectors. Here the distribution of $f(Y_+)$ needs to be determined as well. Note that in these models different covariates can be used in all three parts of the model. Once a model is fit and parameter estimate obtained, we can use estimates from the model to predict Y. See McCullagh and Nelder (1989) for a more complete discussion of generalized linear models and how to fit them.

THE TWO-PART COST MODEL

Diehr et. al.(1999) discuss different methods for modeling cost and utilization data. They point out that cost data often has a sizable portion of zero costs and a skewed distribution of positive costs with non-constant variance. They suggest that a two-stage model is useful for modeling data of this type. One benefit of the two-stage model is that as well as getting an estimate of mean cost we can also estimate the probability of having any utilization at all. Lachenbruch (2001ab) compares methods for testing for group differences and gives power and sample size requirements for two-stage models.

The two-part cost model, in its most commonly used form, estimates the probability Y is positive using logistic regression and uses a log-normal model with either constant or non-constant variance to estimate the cost (or other positive Y) using a log-normal model. We can write this out as:

$$log\left(\frac{P(Y>0)}{1-P(Y>0)}\right) = M't$$

where M is a vector of covariates (predictors), and τ is a parameter vector, and given Y is positive,

$$\log(Y_+) = \mu + \sigma\epsilon, \qquad \epsilon \sim N(0,1)$$

where $\mu = x'\beta$ is the mean of Y_+ and $\log(\sigma^2) = w'\theta$ is the variance of Y_+.

Then to estimate or predict the cost we note that for a given covariate combination (x, w, m) the expected cost is:

$$E(Y|x,w,m) = P(Y>0|\tau,m)e^{x'\beta+0.5*\sigma^2(\theta,w)}$$

and for prediction or estimation we can replace (τ, β, θ) with their estimated values. Note that the index i here is left off to make these statements easier to read, but this model would be for each outcome Y_i and all covariate vectors, means and variances would be appropriately indexed as well.

Given the structure of the two-part model we can fit each part of the model separately using *SAS*® STAT. We first do some preprocessing to create an indicator variable for positive and a new variable with the log of the costs.

```
data data2;
    set data1;
    if cost ne . then ispositive=(cost>0);
    if costpositive then do;
         logcost=log(cost);
         costp=cost;
   end;
run;
```

We can use PROC LOGISTIC or PROC GENMOD to model the probability COST is positive, for example:

```
proc logistic data=data2 descending;
  class x1 x2; /*these are categorical variables */
  model ispositive = x1 x2 x3 x3*x1 ;
  output out=predlogistic pred=phat;
run;
```

or:

```
proc genmod data=data2 descending;
  class x1 x2;
  model ispositive = x1 x2 x3 x3*x1 /link=logit dist=binomial ;
  output out=predlogistic pred=phat;
run;
```

Then to estimate the log-normal part of the model with constant variance use

```
 proc genmod data=data2;
    class x2 x2 x5;
    model costp = x1 x2 x5 x3 x4 x4*x1 /dist =normal link=log;
    output out= y_hatC pred= condpred ;
 run;
```

By creating COSTP, that is equal to COST when COST is positive and missing when cost is zero, we are able to get predicted values for the whole data set. To model positive costs with non-constant variance use PROC MIXED and model log cost and then transform back to estimate the mean cost for each covariate combination in the original scale. Although we do not have repeated measures data using the repeated statement allows us to model the non-constant variance as a function of covariates. Note that all three linear components of the model can have different covariates.

```
proc mixed data=data2 asycov method=ml covtest ;
    class x1 x2 x5;
    model logcost= x1 x2 x5 x3 x4 x4*x1  /htype=1,3 outp=out1 residuals;
    repeated /local=exp(x3 x4 x6) ;
    ods output covparms=esttheta;
run;
```

With PROC MIXED some post-processing will be needed to get predictions:

```
proc transpose data=esttheta out=esttheta;
  var estimate;
  id covparm;
run;

data esttheta;
  set esttheta;
  exp_intercept=log(residual);
run;

data out2;
  if _n_ =1 then set esttheta;
  set out1;
    logsigmasq=exp_intercept+exp_x3*x3+ exp_x4*x4 +exp_x6*x6;
   /* x3 x4 x6 are covariates for variance */
  sigmasqcompute=exp(logsigmasq);
```

```
  /* use below to check that you calculated sigmasqcompute correctly,
   It gives you sigma for the positive values used in estimation*/
  sigmahat=resid/pearsonresid;
  sigmasq=sigmahat**2;/*this should agree with sigmasqcompute */
  predposcost=exp(pred+0.5*sigmasqcompute);
run;

data estimated;
  merge predpos out2;
  predcost=phat*predposcost;
run;
```

Note that PROC IML and the SAS MACRO language can be used to make the above more general, you can contact the author for code if you wish. Finding the best fitting and simplest model can take time, but the beauty is that traditional model fitting methods can be used for both parts of the model. So if you are familiar with multiple linear regression and logistic regression you can fit the two-part cost model.

EXAMPLE 1: DONATIONS

Data on yearly longitudinal giving behavior of 4712 donors were obtained from a large non-profit. There were 127 donors who gave $10,000 or more and they were not included in the analysis because they would automatically be targeted directly for gifts and this group had a different giving pattern. The goal of this analysis was to predict(estimate) future gift amounts based on past giving behavior so that ultimately overall giving behavior could be maximized through phone calls and mailings to the appropriate people. Total giving for 2012, the most current year in the data, was used as the target variable.

After holding out the very large donors(10K or larger), the analysis data set consisted of n=4585 observations and 67% of those donors did not give in 2012. In this data many individuals did not give at all and for the positive gifts the distribution of gift amounts is highly right skewed. The available covariate were gift amounts and indicator variables for whether or not a donor gave for the last nine years.

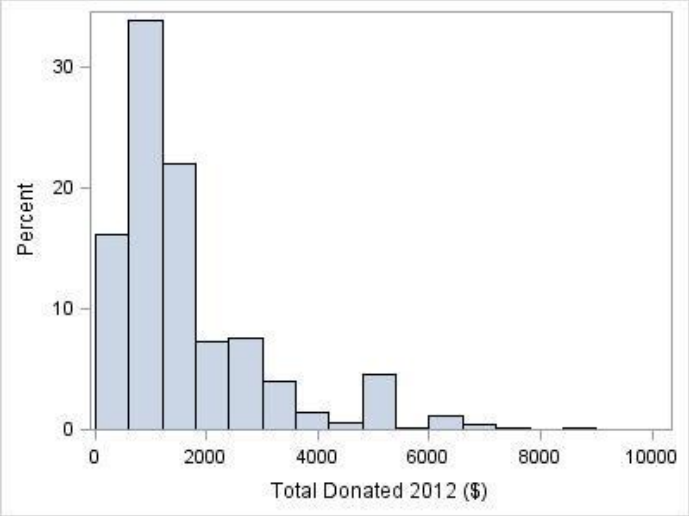


Figure 1 Distribution of Positive Gifts in 2012.

A comparison of an OLS model and the two-part cost model was done. For each iteration, the data were randomly split into a training (about 80%) and a validation set (about 20%) of the data. Then an OLS model was fit with total donated in 2012 as an outcome and the two previous years log(donated+1) and an indicator of donation as predictors. The models were fit to the training data and validated with the validation set. This was done for 100 different random splits of the data and 99% of the time the two-part model had a lower mean squared error. Furthermore, the cost-model has the benefit that it never gives

negative estimates of future donations for a set of covariates. Table 1. Mean Squared Error for **the Two-part and the Ordinary Least Squares (OLS) model for the Donation Data.** gives summaries of the differences between the mean squared error for the 100 validation data sets.

Table 1. Mean Squared Error for the Two-part and the Ordinary Least Squares (OLS) model for the Donation Data.

Model Used	Mean Square Error in the Validation Data Set (for n=100 runs)		
	Average	Minimum	Maximum
Two-Part Model	658228.7	458173.3	906344.8
OLS	840650.8	486238.8	2212692.7

The average percent drop in MSE when using the cost-model instead of the OLS model in each of the 100 different validation data sets is 18% and the median drop is 12%.

EXAMPLE 2: HEALTH CARE-COSTS

Our second example involves data collected on breast cancer costs from a large HMO. Data were obtained on a sample of women who were members of the Lovelace Health Plan, a large managed care organization in New Mexico. The data consisted of 317 cases and 951 randomly selected controls. The cases were all women, aged 20 years or older, who were diagnosed with breast cancer between January 1, 1990 and December 31, 1994 and who were in the health plan, one year prior and one year post diagnosis. Three control women were selected for each case, matched on birth year and also enrolled one year prior and one year post the date of the case woman's diagnosis. In Tollestrup, et. al. (2001) they did a matched case tobit analysis on these data.

The goal of this analysis was estimation of mean total health care costs (excluding pharmacy) in the 4th quarter post diagnosis. We wanted to estimate mean costs by stage of breast cancer adjusting for various covariates. In these patients, in a given quarter, many will have no costs and the distribution of positive costs will be highly positively skewed and have heterogeneous variance. Box-plots of total fourth quarter costs without pharmacy are given in Figure 2.

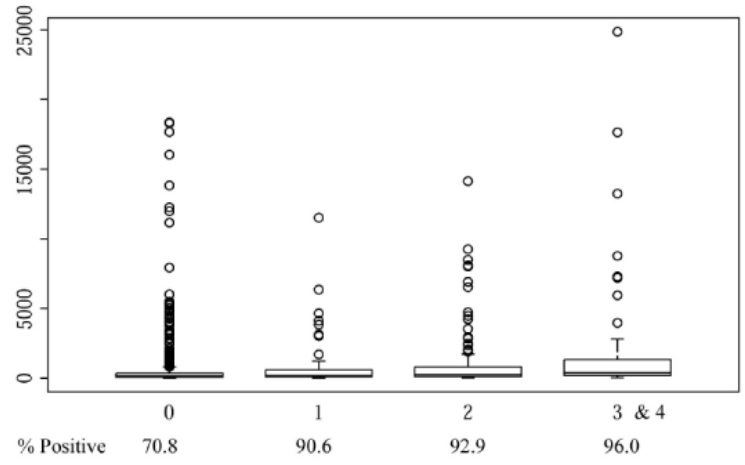


Figure 2: Box-Plots of Total Fourth Quarter Breast Cancer Costs and Percent Positive by Stage.

The predictor variables used in the final model included age, stage indicator variables, log(cost prior to diagnosis +1) and log(total cost 3rd Quarter+1) and log(total cost 2nd quarter+1). The two-stage cost model was used and after log-transform there was still signs of non-constant variance so the more general form of the model was used. The estimated model parameters are given below in Table 2.

Table 2: Parameter Estimates for the Two-Part Cost Model for Total Fourth Quarter Breast Cancer Costs Excluding Pharmacy.

Logistic Regression Model	estimate	SE
Intercept	-2.00	0.354
AGE	0.015	0.0057
stage1	0.545	0.504
stage2	0.866	0.323
stage3	1.048	0.622
log(prior total cost +1)	0.176	0.037
log(total cost 3rd quarter +1)	0.201	0.033
log(total cost 2nd quarter +1)	0.143	0.033
log(σ^2) in the Log-normal Model		
Intercept	-0.543	0.217
Age	0.00986	0.00318
stage1	0.524	0.212
stage2	-0.0158	0.123
stage3	-0.0266	0.190
log(total cost 3rd quarter +1)	0.0842	0.0194
μ in the Log-normal Model		
Intercept	3.342	0.196
AGE	0.0112	0.00299
stage1	0.165	0.247
stage2	0.309	0.112
stage3	0.743	0.173
log(prior total cost +1)	0.0486	0.0219
log(total cost 3rd quarter +1)	0.193	0.0175

The values in Table 2 illustrate that women with stage 1 breast cancer had the lowest costs and the lowest probability of having a cost compared to those with higher stages, but the women with stage 1 breast cancer also had the highest variability compared to those with higher stages of breast cancer.

ZERO-INFLATED MODELS FOR COUNTS

In many scientific areas researchers are interested in modeling count data as a function of a set of covariates. The Poisson regression model is often used for this purpose. However, the Poisson regression model may not adequately fit the data because of over-dispersion. One common type of over-dispersion occurs when there is an increase in zero counts over what would be expected under the Poisson model. A frequently employed model for this type of data is the zero-inflated Poisson regression model or ZIP model. The ZIP model was originally proposed by Lambert(1992) to model defects in manufacturing.

Böhning et. al (1999) used ZIP regression to model data collected in a prospective study on the decayed, missing and filled teeth (DMFT) index of school aged children in Brazil. We will illustrate this method here using this same data set.

In the past a ZIP model could be fit in *SAS*® using PROC LOGISTIC (or PROBIT) and then using PROC NLP to find the maximum likelihood estimates of the parameters for a truncated Poisson distribution. In

SAS® 9.2 and later the zero-inflated Poisson and zero-inflated Negative Binomial model can be fit directly in PROC GENMOD or in PROC COUNTREG.

Figure 3 gives percentages for different DMFT indexes for these data. We see that the Poisson model does not provide a good fit to these data. The ZIP model does a better job. The negative binomial model was explored as well but for simplicities sake it will not be discussed in this paper.

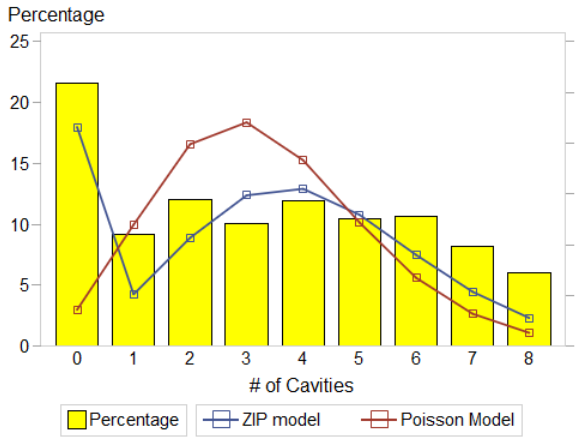


Figure 3: Percentage of Children in the Dental Health Treatment Study with the Given Number of Decayed Missing and Filled Teeth.

In this study school was the unit of randomization and six different treatment regimens were randomly assigned to the six different schools. All studied children in the same school received the same treatment. The outcome of interest was the number of Decayed Missing and Filled Teeth (DMFT) and the DMFT index was computed at the beginning and end of the study.

The final model included indicator effects for treatment, sex and ethnicity and the log of initial DMFT+0.5. Code that can be used to fit the ZIP model is given below. You can use either PROC COUNTREG or PROC GENMOD:

```
proc countreg data = dmft method = qn;
  class school ;
  model dmfte = ldmftb male othereth white school / dist= zip;
  zeromodel dmfte ~ ldmftb male othereth white school;
run;
```

or

```
proc genmod data = dmft method=qn;
  class school ;
  model dmfte = ldmftb male  othereth white school /dist=zip;
  zeromodel ldmftb male  othereth white school /link = logit ;
run;
```

The resulting estimates are given in Table 3. For these students the fluoride wash had the smallest mean increase in DMFT, although the differences were not very large.

Table 3: Estimated Mean DMFT Change at Average Value of all other Covariates

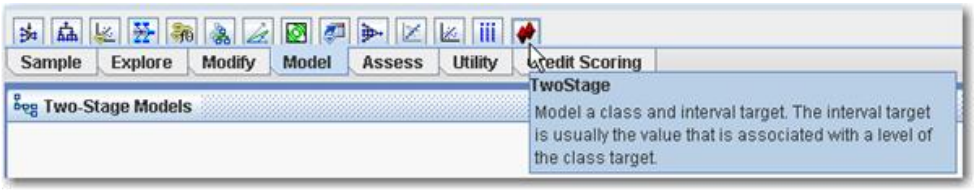
Treatment	Estimated DMFT change	Standard Error
Oral Health Ed	1.47	0.11
All together	1.29	0.11
Control	1.85	0.14
Rice-bran Diet Enrichment	1.90	0.14

Fluoride Wash	1.37	0.11
Oral hygiene instruction	1.63	0.13

See Erdman, Jackson and Sinko (2008) for more info on fitting the Zero-inflated Poisson and Negative Binomial model using PROC COUNTREG.

FITTING TWO-STAGE MODELS IN *SAS*® ENTERPRISE MINER

SAS® *Enterprise Miner* can be used to fit two part models using the two-stage modeling node. The two stage node can be found under the model tab. A discussion of using *Enterprise Miner* to fit two-part models is beyond the scope of this paper but fitting two-stage models using *Enterprise Miner* is covered in the *Advanced Predictive Modeling Using SAS® Enterprise Miner™* course.



CONCLUSION

Two-part models provide a flexible modeling framework and can be fit using SAS STAT or SAS Enterprise Miner. Two-part models can improve predictive performance over using OLS and other one-part models and should be explored as a means of improving estimation and prediction.

REFERENCES

Böhning, D., Ekkehart, D., Schlattmann, P., Mendonca, L. and Kirchner,U. 1999. "The zero-inflated poisson model and the decayed, missing and filled teeth index in dental epidemiology". *Journal of the Royal Statistical Society, Part A*, 162:195–209.

Diehr,P. ,Yanez,D. Ash, A. Hornbrook, M. & Lin, D. Y. 1999 "Methods for analyzing health care utilization and costs." *Annu. Rev. Public Health*, 20:125–44.

Erdman, D., Jackson L. and Sinko A. 2008 "Zero-Inflated Poisson and Zero-Inflated Negative Binomial Models Using the COUNTREG Procedure". *Proceedings of SAS Global Forum 2008*, paper 322-2008, retrieved from http://www2.sas.com/proceedings/forum2008/322-2008.pdf, on March 20, 2015.

Lachenbruch P. A. 2001. "Comparisons of two-part models with competitors" *Statistics in Medicine*, 20:1215–1234.

Lachenbruch P.A. 2001. "Power and sample size requirements for two-part models" *Statistics in Medicine*, 20:1235–1238.

Lambert, D. 1992 "Zero-inflated poisson regression, with an application to defects in Manufacturing". *Technometrics*, 34:1-14.

McCullagh, P. and Nelder, J. A. 1989 *Generalized Linear Models*. New York, NY :Chapman and Hall.

SAS Institute Inc. 2009 *Advanced Predictive Modeling Using SAS® Enterprise Miner™ 6.1 Course Notes.* Cary, NC, USA

Tollestrup,K., Frost, F.J., Stidely, C. A. , Bedrick, E., McMillan, G. Kunde, T. & Peterson, H.V. 2001. "The excess costs of breast cancer health care in hispanic and non-hispanic female members of a managed care organization". *Breast Cancer Research and Treatment*, 66:25–31.

ACKNOWLEDGMENTS

Thank you to Floyd Frost of Lovelace for allowing me to use the data on breast cancer treatment costs, to GVSU Center for Scholarly and Creative Excellence and the Department of Statistics at GVSU for funding my trip and to SAS Institute for awarding me a SAS Global Faculty Scholarship to attend the conference. for helping toLovelace Instithe text for the acknowledgments. This paragraph uses the PaperBody style.

CONTACT INFORMATION

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

Laura Ring Kapitula, PhD
Grand Valley State University
kapitull@gvsu.edu