Paper CC-13

Fit Discrete Distributions via SAS® Macro

Ruiwen Zhang, Cary, NC; Feng Liu, Richmond, VA

ABSTRACT

Discrete distributions like Poisson, Negative Binomial, Zero-Inflated Poisson are important distributions in modeling count data like insurance claims, frequency of diseases, and number of phone calls occurring in a given period of time. SAS/STAT provide procedures like PROC GENMOD to do model fitting, but it lacks graphic comparison between observed vs. predicted densities, also users have to run separate analysis for different types of distributions. We develop SAS macros to do sampling on large data, distribution fitting, test of goodness-of-fit, and provide final comparison tables and charts for decision making. With the help of SAS macro, users can fit different types of discrete distributions on sampled data and view both fit statistics and graphics in a few easy-to-understand options.

INTRODUCTION

Count data or event data has been increasingly common in industries like Finance and Insurance, Clinical research, Call center etc. Fitting a discrete distribution on count data becomes more critical in routine modeling job. SAS/STAT provides advanced generalized linear model fitting procedures like PROC GENMOD and PROC COUNTREG for such analysis. There are many options for each model specification and those options are hard to remember and have to be referenced many times. SAS also does not provide an overlay of empirical and fitted densities onto one chart; therefore, one has to run multiple analyses to obtain a visual comparison of model goodness-of-fit between models. We developed a SAS macro to streamline the process. This paper demonstrates the visual presentation of three discrete distributions: 1. Poisson regression 2. Negative Binomial 3. Zero Inflated Poisson. Other discrete distributions can be extended similarly.

To illustrate the problem, we use a simulated data set. There are 300 observations in the data set and we take a 50% sample, the response variable is insurance claims ("claims") happened during the day and two explanatory variables ("Var1" and "Var2").

```
data Mydata;
   input ID Claims Var1 Var2;
   datalines;
   1 16 1.160573965 2.900040003
   2 5 0.395525542 1.653306512
   3 26 1.272144607 3.140582403
   4 22 0.63364475 4.56137568
   5 8 1.32366145 0.803308752
   6 0 0.265064329 0.111638218
   7 7 0.338269017 2.24825773
   ...
   300 4 0.670061015 0.726005236
   ;
run;
```

POISSON REGRESSION

The following code fit a Poisson Regression.

```
proc genmod data = Mydata;
   model claims = var1 var2 / dist=poisson;
   odsoutput ParameterEstimates=paramvalue Modelfit=mfit;
   outputout=poissonFit1a p=pred1;
run;
```

Here DIST= option specifies Poisson distribution, LINK= option specifies log-linear regression model which is default for Poisson and can be omitted. We can also specify OFFSET= option if each object has varying length of observation time, this data set has same length. The “intercept” term is included in the regression equation by default.

We use ODS output to generate parameter estimate output as well as model fit output. The model fitted values are "Pred1".

Criteria For Assessing Goodness Of Fit

Criterion	DF	Value	Value/DF
Deviance	143	192.9445	1.3493
Scaled Deviance	143	192.9445	1.3493
Pearson Chi-Square	143	174.4825	1.2202
Scaled Pearson X2	143	174.4825	1.2202
Log Likelihood		4013.3953	
Full Log Likelihood		-402.7412	
AIC (smaller is better)		811.4825	
AICC (smaller is better)		811.6515	
BIC (smaller is better)		820.4333	

Algorithm converged.

Analysis Of Maximum Likelihood Parameter Estimates

Parameter	DF	Estimate	Standard Error	Wald 95% Confidence Limits		Wald Chi-Square	Pr>ChiSq
Intercept	1	0.9913	0.0766	0.8412	1.1414	167.59	<.0001
Var1	1	0.1970	0.0518	0.0954	0.2986	14.45	0.0001
Var2	1	0.5014	0.0183	0.4655	0.5374	748.04	<.0001
Scale	0	1.0000	0.0000	1.0000	1.0000		

Figure 1. Output for Poisson Regression

Figure 1 shows model assessment of goodness-of-fit, the deviance and Pearson Chi-Square **Value/DF** is larger than 1 and close to 1 which indicates the Poisson model is adequate to describe the counts of insurance claims. If the ratio is much larger than 1, it may indicate model misspecification or over-dispersed response variable. In such case (>>1), we might consider introducing a dispersion parameter in Poisson regression using OPTION SCALE= DEVIANCE(or DSCALE). Two variables “Var1” and “Var2” and “intercept” term are significant (P-value <0.0001) in this case.

OVERLAY EMPIRICAL WITH FITTED DISTRIBUTION

We want to overlay the discrete fitted distribution on a bar chart of original data. There are multiple ways to visualize discrete densities. From SAS 9.3 and above, we can use VBARPARM statement together with SERIES statement from PROC SGPLOT. We use PROC UNIVARITE to create binning of histograms, then use PROC SGPLOT to overlay fitted distributions.

```
proc univariate data = poissonFit1a noprint;
 histogram pred1 / midpoints = 0 to 54 by 1 vscale = count
 outhistogram= out1a;
run;

data comb1;
label _midpt_ ="Claims"
    countOrig = "Original"
    countPois = "Poisson Fitted";
merge outa(rename=(_count_=countOrig)) out1a(rename=(_count_=countPois));
by _midpt_;
run;

proc sgplot data=comb1 ; /* SAS 9.3 or above stmt */
   vbarparm category=_midpt_ response=countorig/ legendlabel = "Empirical" ;
   series x=_midpt_ y=countpois/lineattrs =(color=darkred thickness=2 pattern=dot) ;
yaxis label="Count";
title "Empirical vs Fitted Distribution for Poisson Regression";
run;
```

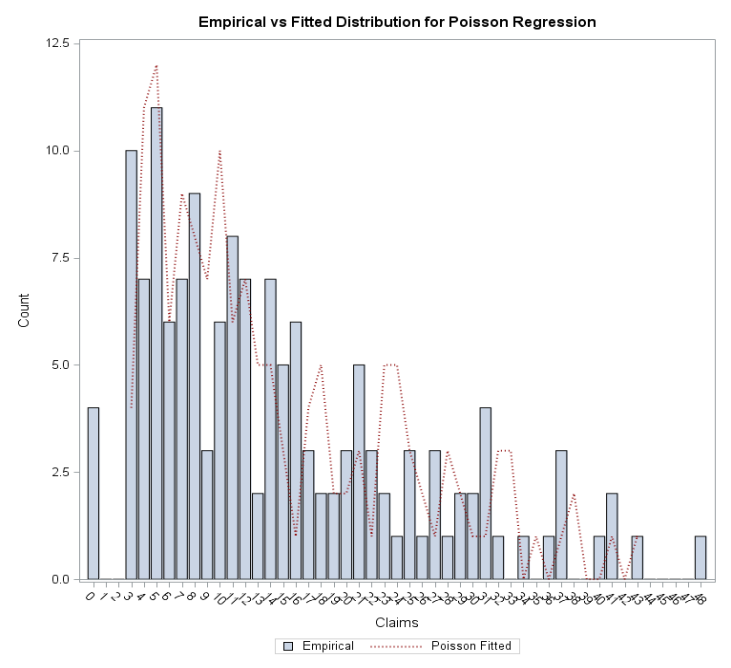

Figure 2. Overlay of Empirical vs. Fitted Poisson Regression Distribution

The dotted red line is model predictions; the Poisson Regression shows a good fit of original data.

NEGATIVE BINOMIAL REGRESSION

The following code fit a Negative Binomial Regression.

```
proc genmod data = Mydata;
   model claims = var1 var2 / dist=negbin;
   ods output ParameterEstimates=paramvalue Modelfit=mfit;
   output out=poissonFit2a p=pred1;
run;
```

The Negative Binomial Regression is another count model which addresses data with over-dispersion. In “Analysis of Maximum Likelihood Parameter Estimates”, there is one additional line called “Dispersion” parameter. If the dispersion is 0 or close to zero, then a Poisson model would be adequate for the data. Based on 95% confidence Level for the dispersion parameter, we can say the dispersion is not significantly different from 0 and we can justify our Poisson regression model.

Analysis Of Maximum Likelihood Parameter Estimates

Parameter	DF	Estimate	Standard Error	Wald 95% Confidence Limits		Wald Chi-Square	Pr>ChiSq
Intercept	1	0.9796	0.0843	0.8143	1.1449	134.94	<.0001
Var1	1	0.2043	0.0593	0.0881	0.3204	11.87	0.0006
Var2	1	0.5034	0.0205	0.4632	0.5436	603.17	<.0001
Dispersion	1	0.0157	0.0092	0.0050	0.0498		

Figure 3. Parameter Estimates for Negative Binomial Regression

Figure 4. shows the Negative Binomial Distribution fitted vs. Empirical Distribution. We can see Negative Binomial fitted distribution has a bigger or fatter tail compared to Poisson fitted.

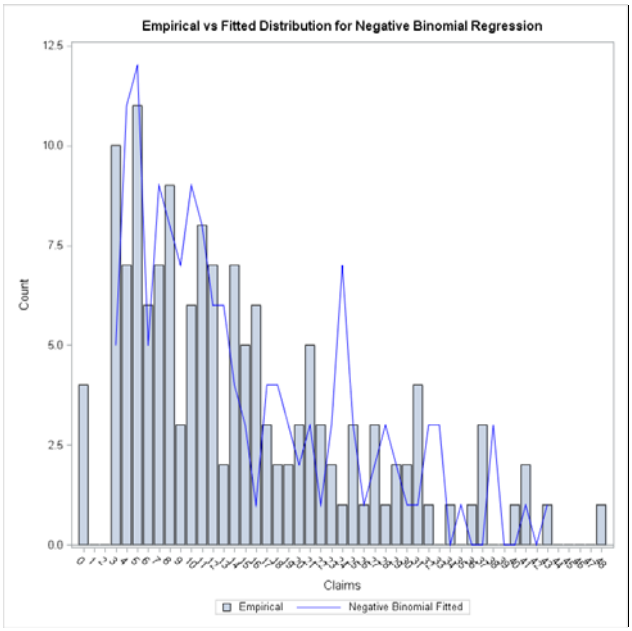

Figure 4. Fitted vs. Empirical for Negative Binomial

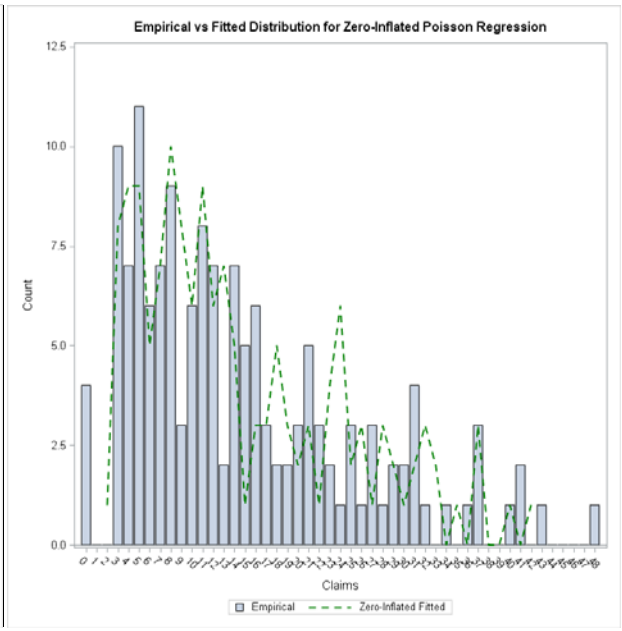

Figure 5. Fitted vs. Empirical for Zero-Inflated Poisson

ZERO-INFLATED REGRESSION

Zero-inflated regression is used to model count data with excess number of zero counts by either a Poisson or Negative Binomial model. The ZIP model (ZERO-INFLATED POISSON) has two components, one count Poisson regression model and one logit model for predicting zeros.

The following code fits a Zero-Inflated Poisson Regression.

```
proc genmod data = Mydata;
  model claims = var1 var2  / dist=zip;
  zeromodel var2 /link = logit;
  ods output ParameterEstimates=paramvalue Modelfit=mfit;
  output out=poissonFit3a p=pred1;
run;
```

The second portion of "Analysis of Parameter Estimates" in Figure 6 demonstrates model estimates for predicting excess zeros. The parameter estimates for both intercept and "Var2" are not significant (p-value > 0.05), suggesting zero-inflated model could be unnecessarily. The original data set also confirms there are not many zeros (counts) of daily claims.

Analysis Of Maximum Likelihood Parameter Estimates

Parameter	DF	Estimate	Standard Error	Wald 95% Confidence Limits		Wald Chi-Square	Pr>ChiSq
Intercept	1	1.0579	0.0787	0.9036	1.2122	180.56	<.0001
Var1	1	0.1802	0.0519	0.0785	0.2820	12.05	0.0005
Var2	1	0.4880	0.0187	0.4514	0.5246	681.84	<.0001
Scale	0	1.0000	0.0000	1.0000	1.0000		

Analysis Of Maximum Likelihood Zero Inflation Parameter Estimates

Parameter	DF	Estimate	Standard Error	Wald 95% Confidence Limits		Wald Chi-Square	Pr>ChiSq
Intercept	1	-0.7816	0.9731	-2.6889	1.1257	0.65	0.4219
Var2	1	-2.1556	1.1004	-4.3123	0.0012	3.84	0.0501

Figure 6. Parameter Estimates for Zero-Inflated Poisson

Figure 5 shows model fitted distribution for Zero-Inflated Poisson Regression.

SAS MACRO

We develop a SAS macro to do all discrete distributions and compare model results side by side. The SAS macro %FITDISCRETE is structured as follows:

1. Sample the original data set if needed.
2. Show summary statistics of original data set. PROC MEANS computes mean, standard deviation, min, and max etc for each variable.
3. Fit discrete models (POISSON, NEGATIVE BINOMIAL and ZERO-INFLATED POISSON). The explanatory variables are specified in inputs.
4. Calculate model assessment of goodness-of-fit and test statistics.
5. Provide model parameter estimates and its significance test.
6. Chart out empirical vs. model fitted discrete distributions. Three models are charted side-by-side for visual comparison.

Please contact author for %FITDISCRETE macro. We can also run SAS programs in batch and automate outputs [3]. For feature selection on large scale of data, SAS high performance node could be applied.

CONCLUSION

We provide visual comparison of SAS discrete distribution models against original count data, i.e. Poisson, Negative Binomial and Zero-Inflated Poisson regression. A SAS macro is developed to combine all test statistics and model parameter estimates. It streamlines routine modeling of count data.

REFERENCES

Gardner W, Mulvey EP, Shaw EC, 1995, “Regression Analyses of Counts and Rates: Poisson, Overdispersed Poisson, and negative binomial models” Psychol Bull. 1995 Nov;118(3):392-404

McCullagh, P. and Nelder, J.A, 1989 “Generalized Linear Models”, Second Edition, Chapman and Hall

Liu, Feng and Zhang, Ruiwen. 2013 “Using VBA to Debug and Run SAS® Program Interactively, Run Batch Jobs, Automate Output, and Build Applications”, SUGI proceedings 2013: page 411-2013

SAS Institute Inc, 2009, *SAS/STAT 9.3 User Guide*, CARY, NC SAS Institute Inc.

Zhao, Z., Zhang, R., Cox, J., Duling, D., Sarle, W. 2013 “Massively parallel feature selection: an approach based on variance preservation”, Journal of Machine Learning, July 2013, Volume 92, Issue 1, pp 195-220

Zelterman, Daniel. 2002, “Advanced Log-Linear Models Using SAS”, ISBN-13: 978-1590470800, SAS Institute

CONTACT INFORMATION

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

Ruiwen Zhang
SAS Institute
ruiwen.zhang@sas.com