Paper 8380-2016

Generalized Linear Models for Non-Normal Data

Theresa Hoang Diem Ngo, La Puente, CA

ABSTRACT

Analysts find the standard linear regression and analysis-of-variance models to be extremely convenient and useful tools. The standard linear model equation form is observations = (sum of explanatory variables) + residual with the assumptions of normality and homogeneity of variance. However, these tools are unsuitable for non-normal response variables in general. Using various transformations can stabilize the variance. However, transforming the estimates back to their original scale and interpreting the results of the analysis can be complicated. Also these transformations are often ineffective because they fail to address the skewness problem. In such cases, we reach the limits of the standard linear model. Generalized linear models have greater power to identify model effects as statistically significant when the data are not normally distributed (Stroup xvii). Unlike the standard linear model, the generalized linear model contains the distribution of the observations, the linear predictor(s), the variance function, and the link function. This paper will introduce generalized linear models using a systematic approach to adapting linear model methods on non-normal data. It will also apply different statistical tests to assess the goodness fit and identify potential problems occurring in the model. SAS®/STAT GENMOD are used to compute basic analyses.

INTRODUCTION

A generalized linear model can be constructed from a standard linear regression model with the addition of terms specifying the response variable's assumed distribution, the link function, and the variance function. The link specifies the relationship between the mean $[\mu]$ of the response variable and the linear predictor$[\ x'\beta]$. The variance function describes the relationship between the mean and the variance of the distribution of the response variable. The variance of the response variable, therefore, does not have to be a constant; it can be a function of the mean (Montgomery p. 563 – 564). Rather than the least square method used with the standard linear model, the maximum likelihood method is used to produce the parameter estimates that are most likely to occur, given the sampling data, by maximizing the likelihood function. Using the least squares method will cause the parameter estimates to be inefficient and the estimated standard errors to be biased because of the non-constant variance (Patetta p. 1-8). After formulating an appropriate generalized linear model identifying the assumed distribution and the link function, the next steps are to assess the model fit and identify and resolve potential problems occurring in the model. Table 1 shows a list of the response variable's assumed distribution that is a member of the exponential family along with its link function.

Table1. Distributions of the Exponential Family

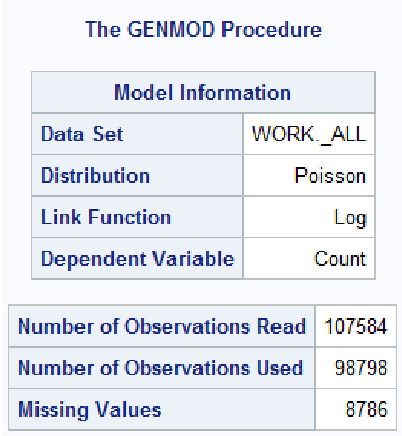
Distribution	Response Variable Type	Link Function
Normal	Continuous symmetric	μ_i (identity link)
Binomial	Categorical	$\ln\left\lvert\frac{\pi_i}{1-\pi_i}\right\rvert$ (logistic link)
Poisson	Count	$ln(\gamma)$ (log link)
Exponential	Time to event	$\frac{1}{\gamma_i}$ (reciprocal link)
Gamma	Time to event	$\frac{1}{\gamma_i}$ (reciprocal link)

GENERALIZED LINEAR MODEL

MODEL ASSESSMENT

The **Model Information** section gives the general information such as the data set, the response variable, the number of observations used, the type of model, the distribution, and the link function. A proprietary data set is used for this model output; the data set has been omitted due to confidentiality.

Output 1. Model Information



The GENMOD Procedure

Model Information	
Data Set	WORK._ALL
Distribution	Poisson
Link Function	Log
Dependent Variable	Count

Number of Observations Read	107584
Number of Observations Used	98798
Missing Values	8786

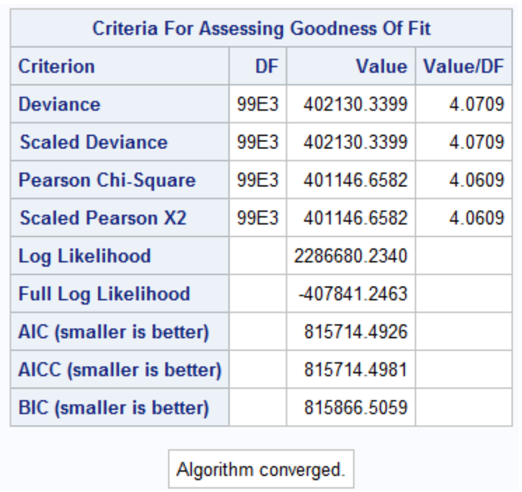
The **Criteria for Assessing Goodness of Fit** section provides useful information for the model selection and assessment: AIC, AICC, BIC, Deviance, Pearson Chi-Square, and Algorithm Convergence.

- The model with smaller AIC, AICC, and BIC values is more desirable.
- The Deviance and Pearson Chi-Square are two goodness-of-fit statistics. The Value/DF values for the Deviance and Pearson Chi-Square are approximately one indicates a good fitting model. Otherwise, it is an indication of these possible issues: model misspecification, missing data, missing other predictors, or over-dispersion of a response variable. Please note these statistics require a sufficient sample size; there must be an average of at least 10 observations in each cell level. There should also be no empty cells (no zero observation in a cell).

  Comparing the two models in Output 2, Model II is a better fit to the data than Model I because the Deviance and Pearson Chi-Square values (1.4329 and 1.2782) are much closer to one. Plus Model II has smaller AIC, AICC, and BIC values.

Output 2. Model I and II Comparison

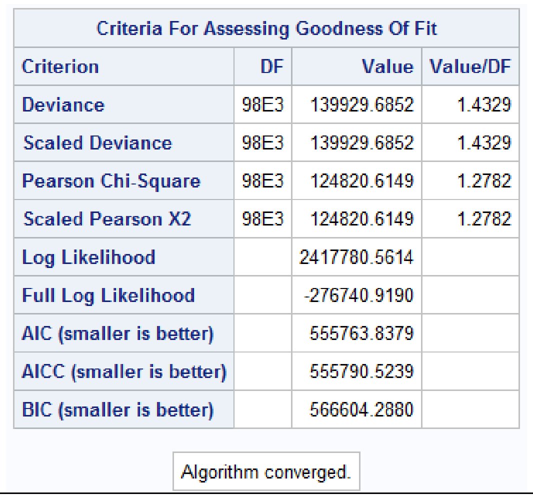
Model I



Criteria For Assessing Goodness Of Fit			
Criterion	DF	Value	Value/DF
Deviance	99E3	402130.3399	4.0709
Scaled Deviance	99E3	402130.3399	4.0709
Pearson Chi-Square	99E3	401146.6582	4.0609
Scaled Pearson X2	99E3	401146.6582	4.0609
Log Likelihood		2286680.2340	
Full Log Likelihood		-407841.2463	
AIC (smaller is better)		815714.4926	
AICC (smaller is better)		815714.4981	
BIC (smaller is better)		815866.5059	

Algorithm converged.

Model II

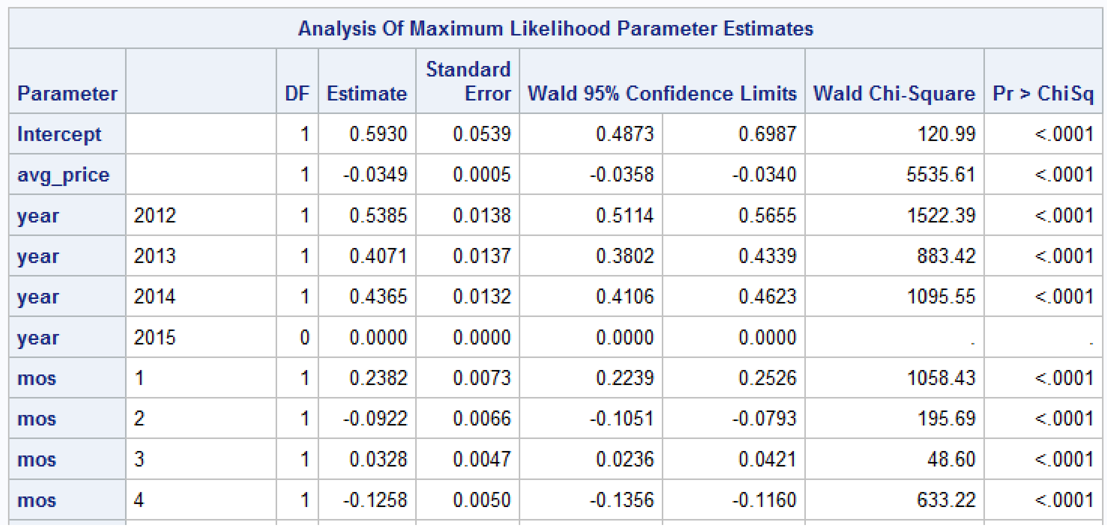


Criteria For Assessing Goodness Of Fit			
Criterion	DF	Value	Value/DF
Deviance	98E3	139929.6852	1.4329
Scaled Deviance	98E3	139929.6852	1.4329
Pearson Chi-Square	98E3	124820.6149	1.2782
Scaled Pearson X2	98E3	124820.6149	1.2782
Log Likelihood		2417780.5614	
Full Log Likelihood		-276740.9190	
AIC (smaller is better)		555763.8379	
AICC (smaller is better)		555790.5239	
BIC (smaller is better)		566604.2880	

Algorithm converged.

- The **Analysis of Parameter Estimates** section determines the significance of the predictors in the model by looking at the P-values. Given that the significant level $\alpha = 0.05$, the p-value is less than 0.05 rejecting the null hypothesis that $\beta_i = 0$. This means the predictor has a significant impact on the response variable. Besides analyzing the significance of each predictor, look at the parameter estimates to see if it makes sense in terms of the negative and positive signs. If the sign is opposite of what is expected, multicollinearity exists in the model (more details in the Potential Problems section).

Output 3. Analysis of Parameter Estimates



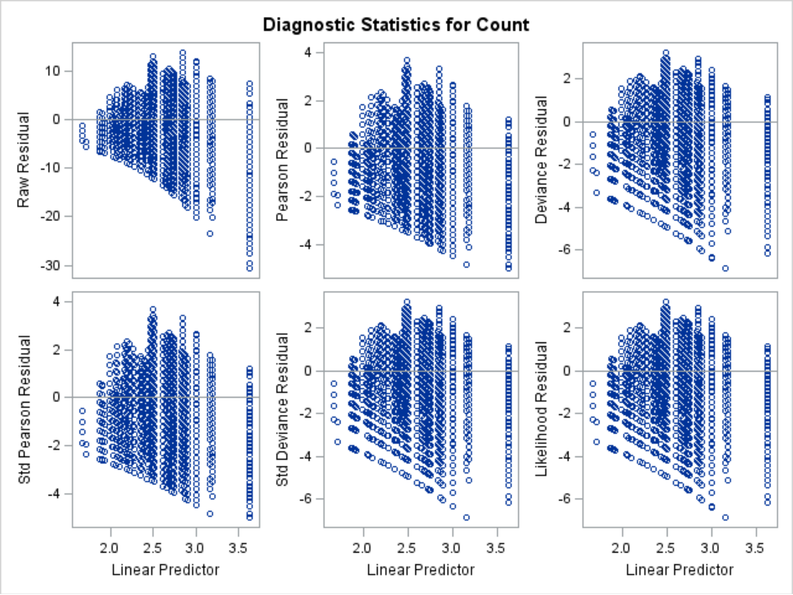
Analysis Of Maximum Likelihood Parameter Estimates								
Parameter		DF	Estimate	Standard Error	Wald 95% Confidence Limits		Wald Chi-Square	Pr > ChiSq
Intercept		1	0.5930	0.0539	0.4873	0.6987	120.99	<.0001
avg_price		1	-0.0349	0.0005	-0.0358	-0.0340	5535.61	<.0001
year	2012	1	0.5385	0.0138	0.5114	0.5655	1522.39	<.0001
year	2013	1	0.4071	0.0137	0.3802	0.4339	883.42	<.0001
year	2014	1	0.4365	0.0132	0.4106	0.4623	1095.55	<.0001
year	2015	0	0.0000	0.0000	0.0000	0.0000	.	.
mos	1	1	0.2382	0.0073	0.2239	0.2526	1058.43	<.0001
mos	2	1	-0.0922	0.0066	-0.1051	-0.0793	195.69	<.0001
mos	3	1	0.0328	0.0047	0.0236	0.0421	48.60	<.0001
mos	4	1	-0.1258	0.0050	-0.1356	-0.1160	633.22	<.0001

POTENTIAL PROBLEMS

The common problems that usually occur in the regression model are lack of fit, over-dispersion, multicollinearity, and failure to converge (Patetta p. 1-31, 2-4).

- **Lack of fit** to the data is caused by a number of possibilities – some outliers; incorrect and/or violated assumptions about the errors; an incorrect form of the predictors. The errors may not be normally distributed or the predictors may not be fixed, but random. If the residuals and partial residual plots are scattered without any patterns, it indicates a good model fit. Otherwise, there is a lack of fit and potential problems in the model. The Output 4 below shows an example of the residual plots with patterns:

Output 4. Residual Plots



- **An Over-dispersion** occurs when the variance is much larger than the mean. In another words, the Value/DF values for the Deviance and Pearson Chi-Square are greater than one. If an over-dispersion is unaccounted, the standard errors will be underestimated and test statistics will be overestimated resulting in excessive Type 1 error rates. To account for either an over-dispersion or under-dispersion, either adjust the standard errors and test statistics or choose another appropriate distribution for the model.
- **Multicollinearity** occurs when there are strong linear dependencies (correlations) among the predictors, which results in less precise parameter estimates and increases variances. A good indication of multicollinearity is when the parameter estimates have opposite signs from what is expected and/or a high correlation coefficient between each pair of predictors in the model. Removing one of the correlated predictors from the model usually minimizes the multicollinearity.
- **Failure to converge** after a number of iterations is commonly due to quasi-complete separation. Quasi-complete separation occurs when there is the zero count in the level of a categorical predictor and/or when an interaction term is created. The presence of zero count should be detected during the univariate screening and the frequency table analysis of the data. The following are solutions to this convergence problem:
    - Collapse the categories of the predictor to eliminate the zero count
    - Eliminate the category altogether
    - Treat the predictor as continuous if it is ordinal

CONCLUSION

The most widely used forecasting model is the standard linear regression, which follows a Normal distribution with mean zero and constant variance. However, the observed relationships between the response variable and the predictors are usually nonlinear. The data set, therefore, does not satisfy the assumptions of a linear regression model. In such cases, a generalized linear model is one of the viable alternatives. After formulating an appropriate generalized linear model, the next fundamental steps are to assess the model fit and identify and resolve potential problems occurring in the model.

SAS CODE

```
*** Generalized Linear Model
*** Specify the distribution of the response variable and link function in
*** the Model statement options (i.e. dist=poisson link=log);
*** Use either dscale/pscale to adjust scaled deviance/Pearson closer to 1;
*** The predictors are fixed
*** To specify repeated and/or random predictors, look at PROC GENMOD Syntax

ods graphics on;

proc genmod data= dataset plot=all;
class categorical predictor(s)
model response variable = predictor(s) / dist=poisson link=log type1 dscale;
output out=dataname pred=var1 resdev=var2 leverage=var3 cooksd=var4;
run;

ods graphics off;
```

REFERENCES

Montgomery, Douglas C. Design and Analysis of Experiments, Sixth Edition. Jon Wiley &Son, Inc., 2005.

*Patetta, Michael.*Categorical Data Analysis Using Logistic Regression Course Notes. SAS Institute Inc., 2002.

Stroup, Walter W. Generalized Linear Mixed Models: Modern Concepts, Methods and Applications. Taylor & Francis Group, 2013.

TRADEMARKS

AUTHOR CONTACT

Please contact the author for any questions and comments.

Theresa Hoang Diem Ngo

theresa.ngo1120@gmail.com