# Fitting Tweedie's Compound Poisson-Gamma Mixture Model by Using PROC HPGENSELECT

## Overview

Semicontinuous random variables are characterized by a continuous distribution that has point masses at one or more locations. One way to model semicontinuous data is to fit a generalized linear model by using a Tweedie distribution for the response variable. Tweedie distributions have been used in such diverse fields as actuarial science, economics, telecommunications, ecology, medicine, and meteorology. This example illustrates how to fit a Tweedie model to aggregate insurance claims payments data by using the HPGENSELECT procedure which is available in SAS/STAT 12.3, recently released with SAS 9.4.

The SAS source code for this example is available as an attachment in a text file. In Adobe Acrobat, right-click the icon and select **Save Embedded File to Disk**. You can also double-click the icon to open the file immediately.

source code

## Analysis

Exponential dispersion models are the response distributions for generalized linear models. Any exponential dispersion model can be characterized by its variance function $V()$, which describes the mean-variance relationship of the distribution when the dispersion is held constant. If $Y$ follows an exponential dispersion model distribution that has mean $\mu$, variance function $V()$, and dispersion $\phi$, then the variance of $Y$ can be written as

$$V(Y) = \phi V(\mu)$$

Tweedie distributions are a special case of the exponential dispersion family for which $V(\mu) = \mu^p$ and $V(Y) = \phi\mu^p$ (Dunn and Smyth 2005). The distribution is defined for all values of $p$ except values of $p$ in the open interval $(0, 1)$. Many important known distributions are a special case of Tweedie distributions including normal ($p = 0$), Poisson ($p = 1$), gamma ($p = 2$), and inverse Gaussian ($p = 3$). Apart from these special cases, the probability density function of the Tweedie distribution does not have an analytical expression. For $p > 1$, it has the form

$$f(y; \mu, \phi, p) = a(y, \phi) \exp\left[\frac{1}{\phi}\left(\frac{y\mu^{1-p}}{1-p} - \kappa(\mu, p)\right)\right]$$

where $\kappa(\mu, p) = \mu^{2-p}/(2-p)$ for $p \neq 2$ and $\kappa(\mu, p) = \log(\mu)$ for $p = 2$. The function $a(y, \phi)$ does not have an analytical expression. It is usually evaluated by using the series expansion methods that are described in Dunn and Smyth (2005).

For $1 < p < 2$, the Tweedie distribution is a compound Poisson-gamma mixture distribution, which is the distribution of $S$ defined as

$$S = \sum_{i=1}^{N} Y_i$$

where $N \sim \text{Poisson}(\lambda)$ and $Y_i \sim \text{gamma}(\alpha, \theta)$ are independently and identically distributed gamma random variables with the shape parameter $\alpha$ and the scale parameter $\theta$. At $Y = 0$, the density is a probability mass that is governed by the Poisson distribution, and for values of $Y > 0$, the density is a mixture of gamma variates with Poisson mixing probability. The parameters $\lambda$, $\alpha$, and $\theta$ are related to the natural parameters $\mu$, $\phi$, and $p$ of the Tweedie distribution as

$$\lambda = \frac{\mu^{2-p}}{\phi(2-p)}$$
$$\alpha = \frac{2-p}{p-1}$$
$$\theta = \phi(p-1)\mu^{p-1}$$

The mean of a Tweedie distribution is positive for $p > 1$.

## Example: Modeling Insurance Claims Data

When modeling aggregate payments from insurance claims, if you assume that the arrival of claims follows a Poisson distribution, that the size of individual claims are independently and identically gamma distributed, and that the arrival and sizes are independent of one another, then the aggregate payments follow a Tweedie compound Poisson-gamma mixture distribution (Frees 2010). This example uses PROC HPGENSELECT to fit a Tweedie model to the aggregate loss data from a Swedish study about third-party automobile insurance claims for 1977. The data were compiled by the Swedish Committee on the Analysis of Risk Premium in Motor Insurance (Andrews and Herzberg 1985). The following SAS statements create the data set MotorIns, and Table 1 describes the variables:

```
data motorins;
  input Kilometres Zone Bonus Make Insured Claims Payment;
  LogInsured=log(insured);
  Zeros=ifn(payment ne 0,1,0);
  datalines;
1 1 1 1 455.13 108 392491
1 1 1 2 69.17 19 46221
1 1 1 3 72.88 13 15694
```

```
1 1 1 4 1292.39 124 422201
1 1 1 5 191.01 40 119373
1 1 1 6 477.66 57 170913

   ... more lines ...

5 7 7 8 13.06 0 0
5 7 7 9 384.87 16 112252
;
run;
```

**Table 1** Example Data Set MotorIns

| Variable | Type | Description |
|---|---|---|
| Kilometres | Class | Distance traveled |
| Zone | Class | Geographical zone |
| Make | Class | Make of automobile |
| Bonus | Continuous | No-claims bonus |
| Insured | Continuous | Number of insured drivers (years × 100,000) |
| LogInsured | Continuous | Natural logarithm of Insured |
| Claims | Continuous | Number of insurance claims |
| Payment | Continuous | Sum of insurance claims payments, in Swedish kronor |
| Zeros | Binary | Indicator variable for Payment not equal to 0 |

Table 2 describes the levels of the classification variable Kilometres.

**Table 2** Values and Labels of Kilometres

| Value | Label |
|---|---|
| 1 | Less than 1,000 km per year |
| 2 | 1,000–15,000 km per year |
| 3 | 15,000–20,000 km per year |
| 4 | 20,000–25,000 km per year |
| 5 | More than 25,000 km per year |

Table 3 describes the levels of the classification variable Zone. The zones are given from a detailed investigation of 100 areas in 1972 and represent combinations of traffic intensity, state of roads, climatic differences, and so on (Andrews and Herzberg 1985).

**Table 3** Values and Labels of Zone

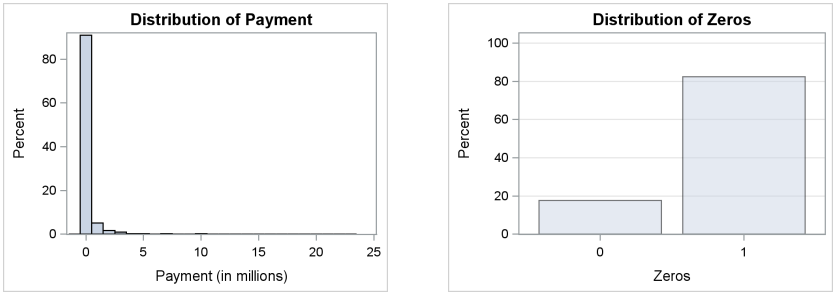
| Value | Label |
|---|---|
| 1 | Stockholm, Göteborg, Malmö with surroundings |
| 2 | Other large cities with surroundings |
| 3 | Smaller cities with surroundings in southern Sweden |
| 4 | Rural areas in southern Sweden |
| 5 | Smaller cities with surroundings in northern Sweden |
| 6 | Rural areas in northern Sweden |
| 7 | Gotland |

The models of cars are classified into 10 premium classes, but in a special investigation for 1977 eight common pure models were chosen and the rest were put in a combined class for reference (Andrews and Herzberg 1985). The levels 1–8 of the classification variable Make represent the eight pure models, and level 9 is the combined class.

The variable Bonus is a measure of individual claim history. The insured motorist starts in the class Bonus = 1. Every year that no claim is filed, the insured moves up one class (Andrews and Herzberg 1985).

Figure 1 shows a histogram of the response variable Payment and the proportion of zeros. The variable Payment exhibits a distribution that is fairly typical of semicontinuous variables: a significant density mass at zero and a continuous, right-skewed distribution elsewhere.

**Figure 1** Distribution of Payment




**Output 1** Frequency of Zeros

The FREQ Procedure

| Zeros | Frequency | Percent | Cumulative Frequency | Cumulative Percent |
|---|---|---|---|---|
| 0 | 385 | 17.64 | 385 | 17.64 |
| 1 | 1797 | 82.36 | 2182 | 100.00 |

The following SAS statements fit a Tweedie compound Poisson-gamma mixture model to the response variable Payment.

The CLASS statement specifies that the variables Kilometres, Zone, and Make are categorical variables. The SPLIT option requests that the columns of the design matrix that correspond to any effect that contains

a split classification variable be able to be selected to enter or leave a model independently of the other design columns of that effect. The PARAM= option specifies a reference cell encoding for the classification variables.

The MODEL statement specifies that the response variable have a Tweedie distribution with a log link function. The candidates for the linear predictor include the main effects and the interactions between the classification variables Kilometres, Zone, and Make and the continuous variable Bonus. The OFFSET= option specifies that the variable LogInsured be included in the linear predictor with a coefficient of 1.

The SELECTION statement requests that stepwise selection be used and that the final model be chosen based on the AICC criterion. The DETAILS=SUMMARY option requests that only a summary of the selection process be displayed rather than the details from each step of the selection process.

The OUTPUT statement requests that the selected model's prediction and residuals be saved to the SAS data set Tweedie.

The ID statement requests that the variable Payment also be included in the output data set.

```
proc hpgenselect data=motorins;
  class Kilometres Zone Make / split param=reference;
  model payment = Kilometres|Zone|Make|Bonus /
        dist=tweedie link=log offset=loginsured;
  selection method=stepwise(choose=aicc) details=summary;
  output out=tweedie P R;
  id payment;
run;
```

The "Performance Information" table in Output 2 shows that the procedure executed in single-machine mode (that is, on the server where SAS is installed). When high-performance procedures run in single-machine mode, they use concurrently scheduled threads. In this case, four threads were used.

The "Model Information" table reports that a Tweedie model was fit with a log link function.

The "Selection Information" table reports that stepwise selection was used, the selection and stopping criteria are the significance level of each individual effect, the entry significance level is the default value of 0.05, and the choose criterion is AICC.

**Output 2** Performance, Model, and Selection Information

```
The HPGENSELECT Procedure

Performance Information

Execution Mode       Single-Machine
Number of Threads    4


Model Information

Data Source                WORK.MOTORINS
Response Variable          Payment
Offset Variable            LogInsured
Class Parameterization     Reference
Distribution               Tweedie
Link Function              Log
Optimization Technique     Quasi-Newton
```

**Output 2** *continued*

Selection Information

| | |
|---|---|
| Selection Method | Stepwise |
| Select Criterion | Significance Level |
| Stop Criterion | Significance Level |
| Choose Criterion | AICC |
| Effect Hierarchy Enforced | None |
| Entry Significance Level (SLE) | 0.05 |
| Stay Significance Level (SLS) | 0.05 |
| Stop Horizon | 1 |

Output 3 shows that the sample size is 2,182, and the "Class Level Information" table shows the number of levels and the level values of the three classification variables.

**Output 3** Sample Size and Classification Variable Levels

| | |
|---|---|
| Number of Observations Read | 2182 |
| Number of Observations Used | 2182 |

Class Level Information

| Class | Levels | Reference Value | Values |
|---|---|---|---|
| Kilometres | 5 * | 5 | 1 2 3 4 5 |
| Zone | 7 * | 7 | 1 2 3 4 5 6 7 |
| Make | 9 * | 9 | 1 2 3 4 5 6 7 8 9 |

* Associated Parameters Split

The "Selection Summary" table in Output 4 reports the variables that are added at each step of the selection process. The summary shows that 30 effects plus an intercept were selected and that the selection process terminated because the sequence of effect additions and removals began cycling.

**Output 4** Effect Selection Summary

The HPGENSELECT Procedure

Selection Summary

| Step | Effect Entered | Effect Removed | Number Effects In | AICC | p Value |
|---|---|---|---|---|---|
| 0 | Intercept | | 1 | 44299.3018 | . |
| 1 | Bonus | | 2 | 43694.2249 | <.0001 |
| 2 | Zone_1 | | 3 | 43576.8235 | <.0001 |
| 3 | Kilometres_1 | | 4 | 43414.2670 | <.0001 |
| 4 | Make_4 | | 5 | 43251.5268 | <.0001 |
| 5 | Make_6 | | 6 | 43186.2263 | <.0001 |
| 6 | Bonus*Kilometres_2 | | 7 | 43128.0129 | <.0001 |
| 7 | Zone_5*Make_8 | | 8 | 43103.4200 | <.0001 |
| 8 | Bonus*Kilometres_2*Zone_1*Make_8 | | 9 | 43087.3135 | <.0001 |
| 9 | Zone_2 | | 10 | 43057.5628 | <.0001 |
| 10 | Bonus*Kilometres_3 | | 11 | 43035.8693 | <.0001 |
| 11 | Make_5 | | 12 | 43021.0275 | <.0001 |
| 12 | Kilometres_1*Make_1 | | 13 | 43010.8831 | 0.0003 |
| 13 | Make_8 | | 14 | 43000.7509 | 0.0003 |
| 14 | Make_2 | | 15 | 42990.3324 | 0.0003 |
| 15 | Bonus*Kilometres_4*Zone_6*Make_4 | | 16 | 42987.4687 | 0.0007 |
| 16 | Kilometres_1*Zone_5*Make_7 | | 17 | 42981.7743 | 0.0008 |
| 17 | Kilometres_1*Zone_6*Make_7 | | 18 | 42976.6767 | 0.0019 |
| 18 | Bonus*Zone_5 | | 19 | 42969.6789 | 0.0020 |
| 19 | Bonus*Zone_3 | | 20 | 42961.8306 | 0.0015 |
| 20 | Bonus*Kilometres_1*Make_1 | | 21 | 42954.6503 | 0.0019 |
| 21 | Kilometres_4 | | 22 | 42945.4183 | 0.0008 |
| 22 | Make_1 | | 23 | 42937.8402 | 0.0017 |
| 23 | Bonus*Make_5 | | 24 | 42931.7907 | 0.0039 |
| 24 | Bonus*Kilometres_1 | | 25 | 42925.8663 | 0.0045 |
| 25 | Zone_2*Make_6 | | 26 | 42920.2139 | 0.0075 |
| 26 | Bonus*Kilometres_1*Zone_6 | | 27 | 42915.7147 | 0.0082 |
| 27 | Bonus*Kilometres_3*Zone_5*Make_2 | | 28 | 42913.6024 | 0.0183 |
| 28 | Bonus*Kilometres_4*Zone_1*Make_5 | | 29 | 42911.4418 | 0.0190 |
| 29 | Zone_1*Make_2 | | 30 | 42908.4035 | 0.0202 |
| 30 | Bonus*Zone_1*Make_2 | | 31 | 42904.3219* | 0.0099 |
| 31 | | Kilometres_1*Zone_6*Make_7 | 30 | 42907.6359 | 0.0534 |

* Optimal Value of Criterion

Stepwise selection stopped because the sequence of effect additions and removals is cycling.

The model at step 30 is selected where AICC is 42904.32.

The "Selected Effects" note in Output 5 lists the effects that are selected for the final model. The "Dimensions" table reports that 31 effects are included in the final model and 33 parameters are estimated. The "Fit Statistics" table reports that the value of AICC for the final model is 42,904.

**Output 5** Selected Effects, Dimensions, Convergence Status, and Fit Statistics

```
Selected Effects: Intercept Kilometres_1 Kilometres_4 Zone_1 Zone_2 Make_1 Make_2 Make_4 Make_5 Make_6 Make_8
                  Kilometres_1*Make_1 Zone_1*Make_2 Zone_2*Make_6 Zone_5*Make_8 Kilometres_1*Zone_5*Make_7
                  Kilometres_1*Zone_6*Make_7 Bonus Bonus*Kilometres_1 Bonus*Kilometres_2 Bonus*Kilometres_3 Bonus*Zone_3
                  Bonus*Zone_5 Bonus*Kilometres_1*Zone_6 Bonus*Make_5 Bonus*Kilometres_1*Make_1 Bonus*Zone_1*Make_2
                  Bonus*Kilometres_2*Zone_1*Make_8 Bonus*Kilometres_3*Zone_5*Make_2 Bonus*Kilometres_4*Zone_1*Make_5
                  Bonus*Kilometres_4*Zone_6*Make_4


                                  Dimensions

                        Number of Effects                      31
                        Number of Effects after Splits         31
                        Number of Parameters                   33
                        Columns in X                           31


                                Fit Statistics

                         -2 Log Likelihood                42837
                         AIC (smaller is better)          42903
                         AICC (smaller is better)         42904
                         BIC (smaller is better)          43091
                         Pearson Chi-Square             1015375
                         Pearson Chi-Square/DF        472.04777


                    Convergence criterion (GCONV=1E-8) satisfied.
```

Output 6 displays the estimates of the model parameters. The estimate of the dispersion parameter, $\phi$, is 349.65 and the estimate of the power, $p$, is 1.36. The effect of using the SPLIT option in the CLASS statement is apparent. None of the classification variables have all their main effects or complete sets of interactions included in the model. The result is a more parsimonious model than you would achieve without enabling the design columns to enter and leave the model independently.

**Output 6** Parameter Estimates

**Parameter Estimates**

| Parameter | DF | Estimate | Standard Error | Chi-Square | Pr > ChiSq |
|---|---|---|---|---|---|
| Intercept | 1 | 6.422717 | 0.030076 | 45604.6548 | <.0001 |
| Kilometres_1 | 1 | −0.396625 | 0.050252 | 62.2944 | <.0001 |
| Kilometres_4 | 1 | −0.163970 | 0.038900 | 17.7679 | <.0001 |
| Zone_1 | 1 | 0.431027 | 0.028089 | 235.4780 | <.0001 |
| Zone_2 | 1 | 0.234752 | 0.028513 | 67.7853 | <.0001 |
| Make_1 | 1 | 0.102037 | 0.031974 | 10.1842 | 0.0014 |
| Make_2 | 1 | 0.104421 | 0.048899 | 4.5601 | 0.0327 |
| Make_4 | 1 | −0.676695 | 0.051711 | 171.2483 | <.0001 |
| Make_5 | 1 | 0.467489 | 0.091598 | 26.0480 | <.0001 |
| Make_6 | 1 | −0.200767 | 0.039391 | 25.9776 | <.0001 |
| Make_8 | 1 | 0.240254 | 0.056656 | 17.9826 | <.0001 |
| Kilometres_1*Make_1 | 1 | 0.443491 | 0.132307 | 11.2358 | 0.0008 |
| Zone_1*Make_2 | 1 | 0.728101 | 0.202768 | 12.8939 | 0.0003 |
| Zone_2*Make_6 | 1 | −0.286627 | 0.101679 | 7.9464 | 0.0048 |
| Zone_5*Make_8 | 1 | 0.520894 | 0.160295 | 10.5599 | 0.0012 |
| Kilometres_1*Zone_5*Make_7 | 1 | 0.736407 | 0.250920 | 8.6132 | 0.0033 |
| Kilometres_1*Zone_6*Make_7 | 1 | 0.453568 | 0.230985 | 3.8558 | 0.0496 |
| Bonus | 1 | −0.138215 | 0.006689 | 426.9209 | <.0001 |
| Bonus*Kilometres_1 | 1 | −0.035637 | 0.010883 | 10.7225 | 0.0011 |
| Bonus*Kilometres_2 | 1 | −0.069006 | 0.006272 | 121.0511 | <.0001 |
| Bonus*Kilometres_3 | 1 | −0.045452 | 0.006450 | 49.6621 | <.0001 |
| Bonus*Zone_3 | 1 | 0.019307 | 0.005369 | 12.9336 | 0.0003 |
| Bonus*Zone_5 | 1 | 0.027528 | 0.007469 | 13.5859 | 0.0002 |
| Bonus*Kilometres_1*Zone_6 | 1 | 0.031829 | 0.012152 | 6.8603 | 0.0088 |
| Bonus*Make_5 | 1 | −0.055964 | 0.018012 | 9.6539 | 0.0019 |
| Bonus*Kilometres_1*Make_1 | 1 | −0.063791 | 0.025037 | 6.4914 | 0.0108 |
| Bonus*Zone_1*Make_2 | 1 | −0.107665 | 0.039459 | 7.4448 | 0.0064 |
| Bonus*Kilometres_2*Zone_1*Make_8 | 1 | 0.166930 | 0.043453 | 14.7578 | 0.0001 |
| Bonus*Kilometres_3*Zone_5*Make_2 | 1 | 0.101962 | 0.045120 | 5.1068 | 0.0238 |
| Bonus*Kilometres_4*Zone_1*Make_5 | 1 | 0.098160 | 0.045496 | 4.6550 | 0.0310 |
| Bonus*Kilometres_4*Zone_6*Make_4 | 1 | 0.233105 | 0.094307 | 6.1097 | 0.0134 |
| Dispersion | 1 | 349.647997 | 28.740411 | . | . |
| Power | 1 | 1.363043 | 0.008891 | . | . |

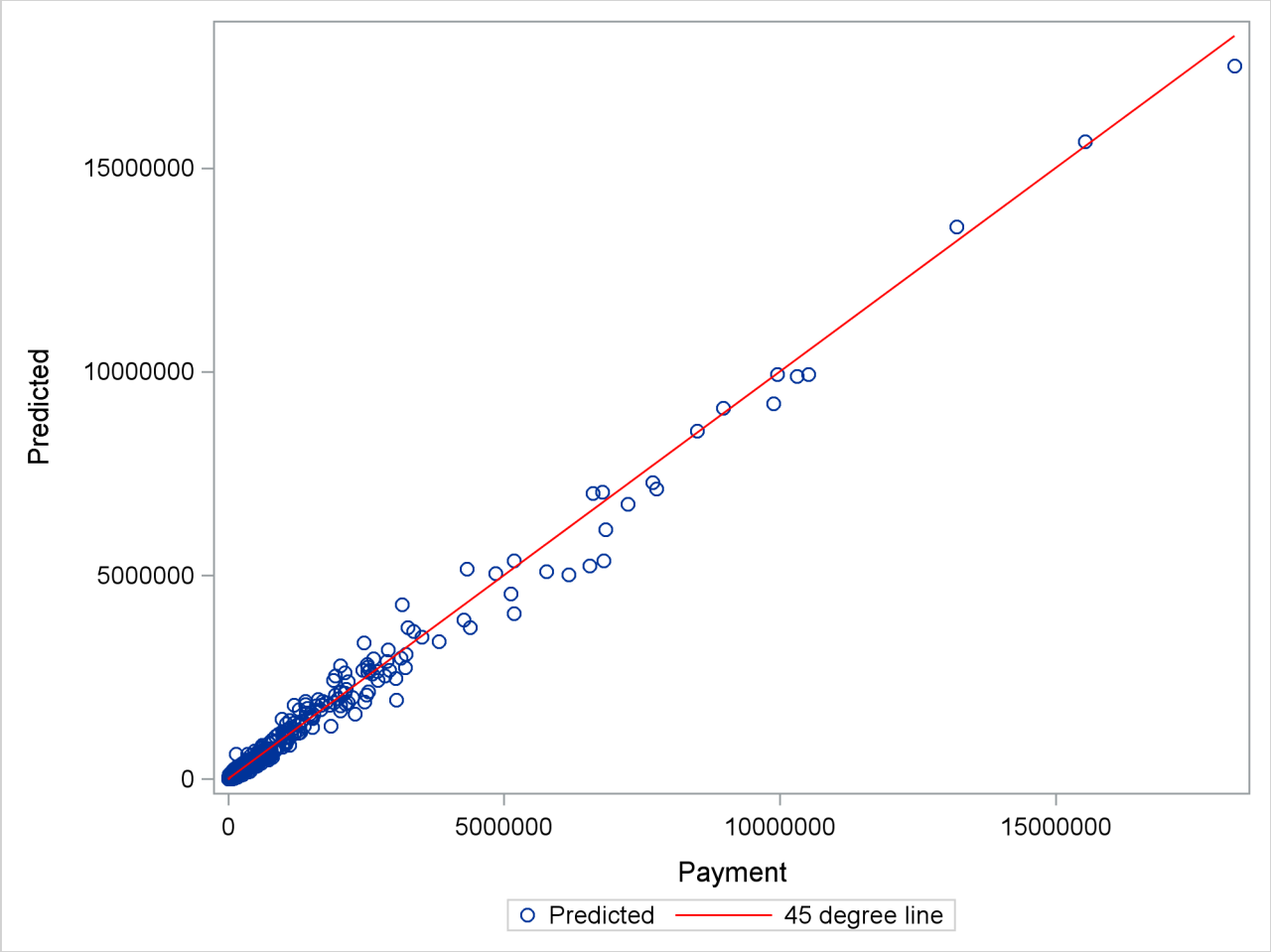
The following SAS statements generate a scatter plot that compares the model predictions with the observed values of the response variable.

```
proc sort data=tweedie out=tweedie;
  by payment;
run;


proc sgplot data=tweedie;
  scatter x=payment y=pred / legendlabel="Predicted";
  series x=payment y=payment / lineattrs=(pattern=solid color=red)
                               legendlabel="45 degree line";
  yaxis label="Predicted";
run;
```

Figure 2 shows that the predictions of the final model compare favorably with the observed responses.

**Figure 2** Scatter Plot of Predicted versus Observed Payments




## References

Andrews, D. F. and Herzberg, A. M. (1985), *A Collection of Problems from Many Fields for the Student and Research Worker*, New York: Springer-Verlag.

Dunn, P. K. and Smyth, G. K. (2005), "Series Evaluation of Tweedie Exponential Dispersion Model Densities," *Statistics and Computing*, 15, 267–280.

Frees, E. W. (2010), *Regression Modeling with Actuarial and Financial Applications*, Cambridge: Cambridge University Press.

Jørgensen, B. and Paes de Souza, M. C. (1994), "Fitting Tweedie's Compound Poisson Model to Insurance Claims Data," *Scandinavian Actuarial Journal*, 1, 69–93.