Paper 142-2012

Using PROC GAM to Forecast Claims Reserves in the Runoff Triangles

Ling Huang, Fulcrum Analytics Inc., Fairfield, CT

ABSTRACT

Runoff triangle means the two-way tabulation according to the warranty start time and duration. Forecasting adequate claims and setting up suitable reserves in the runoff triangles is an important part of an insurance company. Traditionally the claims reserves are estimated by linear regression models such as chain ladder in practice. A more accurate and flexible model can be built by examining claims counts and claims size separately, and then combine the estimates to compute the total claims reserves. This paper focuses on fitting generalized additive models (GAM) with PROC GAM to predict claims counts and claims size separately, and then calculated the total claims reserves in the runoff triangles. The final model is validated on a test dataset to check how the accuracy.

INTRODUCTION

The need to accurately forecast claims has led to many loss reserving techniques in practice. Traditionally claim reserving is estimated by linear regression models such as chain ladder. A more accurate and flexible model can be built by examining claims counts and claims size separately, and then combining the estimates into the total claims reserves.

There are many advantages of separately predicting claims counts and claims size. First, the expected claims counts change as the number of warranties change. Growth in the volume of business should be accounted directly to forecast the claims counts. Second, the effects of additional policy adjustments or general economic inflations are reflected directly in the claims size distribution. Third, the relationship between claims counts and claims size can be captured in a more transparent manner.

The paper presents an approach to fit GAM models with PROC GAM to predict claims counts and claims size separately, and then forecast claims reserves in the runoff triangles. First, the GAM model is reviewed and a runoff triangle is introduced as an example. Then GAM modeling methods and results for claims counts and claims size are presented respectively. Finally, estimates from claims counts and claims size are combined into the total claims amounts, and then results are validated to measure the accuracy.

GAM MODEL

GAM was first proposed by Hastie and Tibshirani (1986) to combine generalized linear models with additive models. Generalized linear models assume the dependent variable is generated from an exponential dispersion family, and the dependent variable is related to the linear predictors via a link function. Additive models assume the nonparametric smoothing splines for the predictors in the linear regression models. By blending these two models, generalized additive models can be used in a wide range of modeling scenarios.

Let Y be the dependent variable and $X_1, \ldots, X_p$ represents p predictors. A generalized additive model assumes Y follows exponential dispersion family with a link function g such that

$$g\left(E\left(Y|X_1, \ldots, X_p\right)\right) = \beta_0 + \sum_{i=1}^{p} \beta_i f_i(X_i) \ ,$$

where β_i are the coefficients and f_i are nonparametric smooth functions, $i = 1, \ldots, p$.

In SAS/STAT, PROC GAM provides a powerful tool for implementing GAM to identify and characterize nonlinear effects between the dependent variables and the predictors. It implements B-splines and local regression methods for univariate smoothing components, and thin-plate smoothing splines for bivariate smoothing components. Each smooth function f_i is controlled by a single smoothing parameter. By minimizing the approximated predicted errors, a generalized cross validation (GCV) function is applied in PROC GAM to automatically select the smoothing parameters. Therefore, the model scales well with the increasing predictors to avoid the curse of dimensionality.

With the flexibility of PROC GAM, it is applied mostly to visualize nonlinear effects for data exploration. One major concern of GAM for conducting forecasts is the over-fitting problem due to over-complex smooth terms. However, by applying PROC GAM with a GCV function to choose the appropriate degrees of freedom in smooth terms and rigorously validating model predictions, GAM can be used directly as a powerful predictive modeling tool.

RUNOFF TRIANGLES

Due to the variations of different products and different warranty lengths, a warranty portfolio is typically divided into disparate groups of products and durations. For each group, the warranties and related claims need to be aggregated in a table format, generally by calendar years and months. The warranties are grouped together by warranty start years and start months, which turn to rows of the table. Meanwhile, the claims counts or claims amounts are merged to related warranties, which turn to columns of the table according to when they are emerged. The data organization as the runoff triangles greatly facilitates comparison of the development history of claims counts or claims amounts by failure years and months. Our goal is to forecast the lower unobserved claims amounts of the table. A sample runoff triangle of claims amounts for two year warranty is listed below:

Start Date		Number of	Duration																							
Year	Month	Warranties	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
2008	1	257	110	0	0	0	0	0	0	0	110	0	0	330	1100	1093	220	0	613	110	220	1039	440	360	0	353	640
2008	2	306	220	110	0	0	6	0	0	0	0	110	330	110	770	0	239	0	354	702	250	462	110	110	362	220	881
2008	3	409	171	0	0	220	0	0	0	0	0	0	140	220	550	1002	1176	330	925	220	220	0	521	553	471	653	584
2008	4	290	420	0	0	0	0	0	101	0	110	110	110	220	1053	479	686	807	110	380	440	0	330	425	467	880	440
2008	5	325	241	0	0	0	0	0	0	378	110	220	0	330	347	330	836	550	212	356	330	0	194	0	440	220	220
2008	6	315	358	0	0	100	110	0	220	368	0	0	110	550	761	220	110	303	220	522	111	304	233	110	110	372	844
2008	7	261	507	0	0	0	0	0	0	0	0	0	0	110	533	0	0	0	220	0	0	110	110	846	0	0	110
2008	8	224	1037	0	0	0	0	0	0	0	0	0	220	0	797	544	0	220	220	234	243	166	236	110	444	889	695
2008	9	239	135	0	0	0	0	216	110	0	0	0	0	220	550	220	123	147	209	110	0	110	0	450	332	507	236
2008	10	197	528	0	0	0	0	0	0	110	127	0	0	220	330	622	330	415	110	0	0	443	0	0	368	421	110
2008	11	198	397	0	0	110	0	0	0	0	100	0	0	110	660	336	110	0	260	110	0	361	153	281	110	294	791
2008	12	128	0	0	272	0	0	0	0	0	0	220	0	220	0	370	110	237	467	790	180	118	0	0	143	0	0
2009	1	140	430	0	0	110	0	0	0	0	220	0	220	220	330	131	670	330	0	0	256	371	592	220	171	110	
2009	2	213	853	0	0	0	0	0	0	110	0	0	220	550	550	364	0	127	258	0	359	0	544	110	272		
2009	3	207	131	0	0	0	0	0	0	0	110	110	330	220	421	420	465	237	233	239	0	110	164	809			
2009	4	111	1364	0	0	0	0	110	0	110	0	0	0	110	440	110	310	317	358	0	110	0	110				
2009	5	150	0	333	110	0	0	110	0	0	0	220	110	220	220	352	218	220	282	233	0	110					
2009	6	154	417	0	110	0	0	0	0	0	330	0	330	0	330	330	0	0	131	0	0						
2009	7	148	68	0	0	0	110	0	0	0	0	110	0	220	110	0	510	451	110	110							
2009	8	146	220	0	123	0	0	0	110	0	0	0	110	110	366	0	0	171	0								
2009	9	136	110	0	0	0	0	0	0	0	0	0	0	0	220	343	110	110									
2009	10	123	250	0	0	0	0	0	0	110	281	119	110	479	330	0	121										
2009	11	96	0	0	0	0	0	0	0	0	0	0	0	110	220	0											
2009	12	77	110	0	0	0	0	0	0	110	0	0	110	0	289												
2010	1	48	220	0	0	0	0	0	0	0	0	0	0	0													
2010	2	149	121	0	0	0	0	110	0	0	0	0	0														
2010	3	215	220	0	0	0	0	0	114	110	110	367															
2010	4	178	110	0	0	0	151	110	0	0	0																
2010	5	187	604	110	0	0	110	114	0	0																	
2010	6	176	429	0	0	0	0	0	0																		
2010	7	168	118	133	0	0	0	0																			
2010	8	124	367	0	175	0	110																				
2010	9	182	430	0	0	0																					
2010	10	169	0	0	0																						
2010	11	140	110	110																							
2010	12	102	0																								

Figure 1. Simulated Sample Runoff Triangle of Claims Amounts for Two Year Warranty

Before using PROC GAM to predict claims reserves, claims amounts is first decomposed into claims counts and claims size, i.e., the average claims cost. Next, logarithm transformations are applied to warranty counts. Finally, all claims data is split into the training data set and the test data set by the cutoff date. Both the training data set and the test data set include start_year, start_month, log_warranties, duration, failure_year, failure_month, claim_cnt, avg_claim columns. The SAS codes for pre-processing the data are listed below:

```
%let cutoff=mdy(1,1,2011);
data train test;
set all;
*pre-process missing claim_cnt;
if failure_date <&cutoff and claim_cnt=. then
do claim_cnt=0; total_claim_amt=0;
end;
*compute average claims;
if claim_cnt=0 then avg_claim =.;
else avg_claim=total_claim_amt/claim_cnt;
*logarithm transformation of warranty counts;
log_warranties=log(warranties_cnt);
*export data to training data set and test data set respectively;
if failure_date <&cutoff then output train; else output test;
run;
```

USE PROC GAM TO FIT POISSON MODEL TO PREDICT CLAIM FREQUENCY

PROC GAM was run with claims counts as the response variable following Poisson distribution using log_warranties, duration, start_year, and failure_year as covariates. The SAS codes and outputs are as follows:

```
ods graphics on;
proc gam data=train;
model claim_cnt=param(log_warranties) spline(start_year) spline(failure_month)
spline(duration)/method=gcv dist=poisson;
output out=prediction_cnt p;
score data=test out=forecast_cnt;
run;
ods graphics off;
```

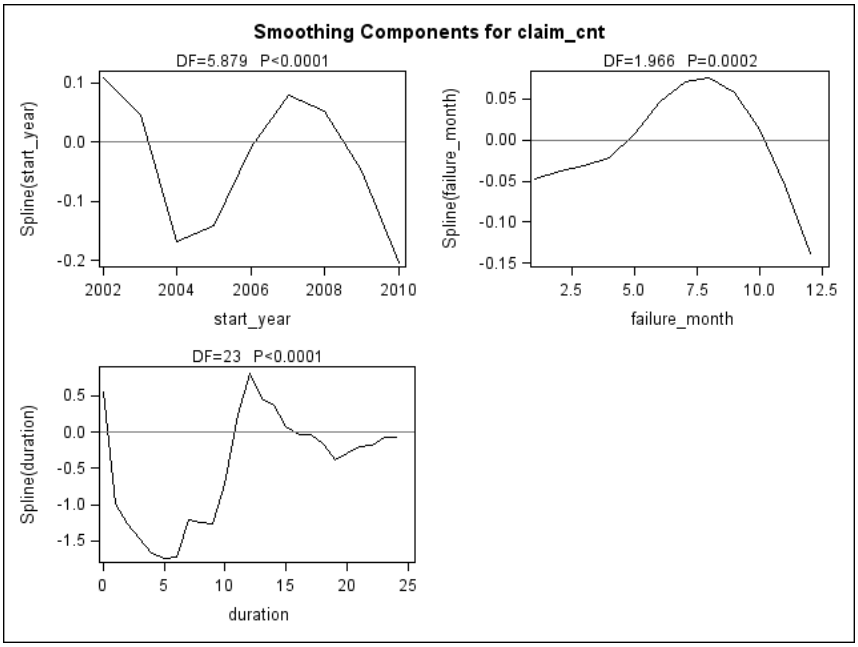
Parameter Estimates

Parameter	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	-198.17791	17.72835	-11.18	<.0001
log_warranties	0.74470	0.05873	12.68	<.0001
Linear(start_year)	0.09672	0.00879	11.01	<.0001
Linear(failure_month)	-0.00213	0.00552	-0.39	0.6998
Linear(duration)	0.03858	0.00289	13.37	<.0001

Dependent Variable: claim_cnt
Regression Model Component(s): log_warranties
Smoothing Model Component(s): spline(start_year) spline(failure_month) spline(duration)

Smoothing Model Analysis
Analysis of Deviance

Source	DF	Sum of Squares	Chi-Square	Pr > ChiSq
Spline(start_year)	5.87926	28.001091	28.0011	<.0001
Spline(failure_month)	1.96584	16.941270	16.9413	0.0002
Spline(duration)	23.00000	1002.678361	1002.6784	<.0001



Output 1. Output from PROC GAM for Predict Claim Frequency

PROC GAM fits the GAM model with Poisson responses by specifying the option DIST = Poisson in the MODEL statement. The option PARAM(log_warranties failure_year) indicates that variables inside the parentheses are estimated parametrically in the linear form. The option SPLINE(start_year), SPLINE(failure_month) and SPLINE (duration) indicate to fit univariate spline smoothers respectively. The smooth term's degrees of freedom can be determined automatically by the GCV function with the option METHOD = GCV. The OUTPUT statement exports the predicted values to the prediction_cnt data set. The SCORE statement applies the estimated GAM model to compute the forecasted claim counts in the test data set and output to the forecast_cnt data set.

The Parameter Estimates show the covariate log_warranties is statistically significant at the 1% level. The finding is expected because with all other factors being equal, more warranties will produce more claims. As shown in Analysis of Deviance, all smooth terms of start_year, failure_month and duration are significant at the 1% level respectively. Furthermore, there are clear nonlinear patterns of start_year, failure_month and duration on claim_cnt as presented in the smoothing component plots.

The deviance and the root-PMSE of the model are important statistics to check overall model fitting. Smaller are better for these statistics. Among all candidate models, both statistics achieve minimums for the model in the paper.

Deviance	Observed Average Claims Counts	Estimated Average Claims Counts	Root-PMSE
2948	1.26	1.26	1.29

Table 1. Modeling Statistics of Claim Frequency Model

The deviance of the model is 2948; the root-PMSE is 1.29. The observed average claims counts 1.26 and the estimated average claims counts 1.26 are pretty close in the training data set. All results above indicate a reasonable fit of PROC GAM model for the claims counts.

USE PROC GAM TO FIT GAMMA MODEL TO PREDICT CLAIMS SIZE

Assuming claims size as the response variable following a Gamma distribution, PROC GAM is run using duration and start_year as covariates. The SAS codes and outputs are listed below:

```
ods graphics on;
proc gam data=train;
model avg_claim=loess(start_year) loess(duration)/method=gcv dist=gamma;
output out=prediction_avg p;
score data=test out=forecast_avg;
run;
ods graphics off;
```

Parameter Estimates

Parameter	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	-0.00875	0.00021465	-40.75	<.0001

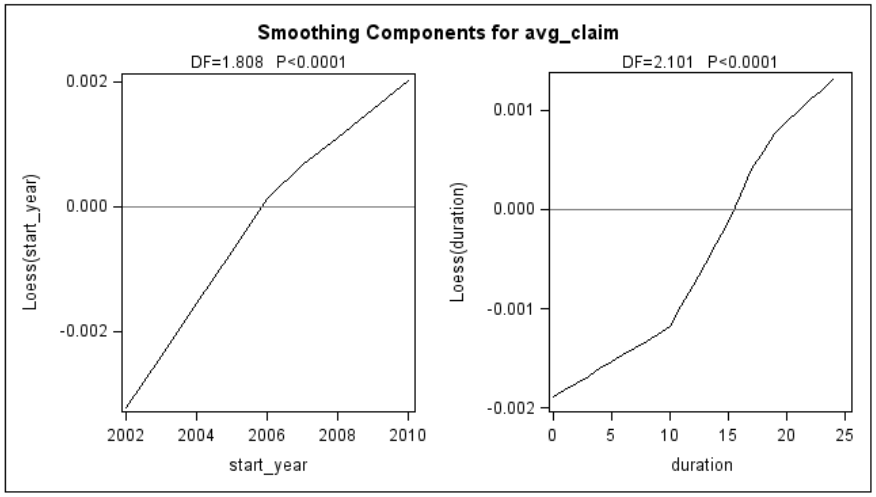
Smoothing Model Analysis
Fit Summary for Smoothing Components

Component	Smoothing Parameter	DF	GCV	Num Unique Obs
Loess(start_year)	0.765966	1.808185	0.000003293	1284
Loess(duration)	0.744938	2.100805	0.000003294	1284

Dependent Variable: avg_claim
Smoothing Model Component(s): loess(start_year) loess(duration)

Smoothing Model Analysis
Analysis of Deviance

Source	DF	Sum of Squares	Chi-Square	Pr > ChiSq
Loess(start_year)	1.80819	33.840464	45.2768	<.0001
Loess(duration)	2.10080	15.681717	20.9813	<.0001



Output 2. Output from PROC GAM for Predict Claim Size

The option DIST= Gamma in PROC GAM specifies the GAM model with Gamma distribution. The option LOESS(start_year) and LOESS(duration) indicate to apply the univariate spline smoothers for the covariates respectively. METHOD = GCV in the MODEL statement means the smooth term's degrees of freedom are determined automatically by the GCV function. The OUTPUT statement exports the predicted claims size to the prediction_avg data set. Lastly, the SCORE statement applies the estimated GAM model to compute the forecasted claims size in the test data set and output them to the forecast_avg data set.

Smooth terms of start_year and duration are significant at the 1% level as shown in Analysis of Deviance. In addition, there are clear nonlinear patterns of start_year and duration on claim_avg as shown in the smoothing component plots.

Deviance	Observed Average Claims Size	Estimated Average Claims Size	Root-PMSE
956	$110.25	$105.35	95.21

Table 2. Modeling Statistics of Claim Size Model

The deviance of the claims size model is 956; the root-PMSE is 95.21. The estimated claims size of $105.35 is close to the observed claims size $110.25 on average. And hence the PROC GAM fits the claims size data quite well.

VALIDATE THE FINAL MODEL

The total estimated claims amounts is calculated by multiplying the estimated claims counts from the Poisson model output with the estimated claims size from the Gamma model output for each group. The sum of all groups becomes the total estimated claims amounts for the entire portfolio. The following SAS codes show the combination method based on the estimated claims counts and the estimated claims size. Furthermore, the results of the estimated claims amounts for the sample runoff triangle in the Figure 1 are listed in Figure 2.

```
data prediction;
merge prediction_cnt prediction_avg;
by start_year start_month duration;
p_claim_amt=p_claim_cnt*p_avg_claim;
run;

data forecast;
merge forecast_cnt forecast_avg;
by start_year start_month duration;
p_claim_amt=p_claim_cnt*p_avg_claim;
run;
```

Start Date		Number of	Duration																							
Year	Month	Warranties	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
2008	1	257	110	0	0	0	0	0	0	0	110	0	0	330	1100	1093	220	0	613	110	220	1039	440	360	0	353	640
2008	2	306	220	110	0	0	6	0	0	0	0	110	330	110	770	0	239	0	354	702	250	462	110	110	362	220	881
2008	3	409	171	0	0	220	0	0	0	0	0	0	140	220	550	1002	1176	330	925	220	220	0	521	553	471	653	584
2008	4	290	420	0	0	0	0	0	101	0	110	110	110	220	1053	479	686	807	110	380	440	0	330	425	467	880	440
2008	5	325	241	0	0	0	0	0	0	378	110	220	0	330	347	330	836	550	212	356	330	0	194	0	440	220	220
2008	6	315	358	0	0	100	110	0	220	368	0	0	110	550	761	220	110	303	220	522	111	304	233	110	110	372	844
2008	7	261	507	0	0	0	0	0	0	0	0	0	0	110	533	0	0	0	220	0	0	110	110	846	0	0	110
2008	8	224	1037	0	0	0	0	0	0	0	0	0	220	0	797	544	0	220	220	234	243	166	236	110	444	889	695
2008	9	239	135	0	0	0	0	216	110	0	0	0	0	220	550	220	123	147	209	110	0	110	0	450	332	507	236
2008	10	197	528	0	0	0	0	0	0	110	127	0	0	220	330	622	330	415	110	0	0	443	0	0	368	421	110
2008	11	198	397	0	0	110	0	0	0	0	100	0	0	110	660	336	110	0	260	110	0	361	153	281	110	294	791
2008	12	128	0	0	272	0	0	0	0	0	0	220	0	220	0	370	110	237	467	790	180	118	0	0	143	0	0
2009	1	140	430	0	0	110	0	0	0	0	220	0	220	220	330	131	670	330	0	0	256	371	592	220	171	110	**304**
2009	2	213	853	0	0	0	0	0	0	110	0	0	220	550	550	364	0	127	258	0	359	0	544	110	272	**393**	**419**
2009	3	207	131	0	0	0	0	0	0	0	110	110	330	220	421	420	465	237	233	239	0	110	164	809	**325**	**388**	**411**
2009	4	111	1364	0	0	0	0	110	0	110	0	0	0	110	440	110	310	317	358	0	110	0	110	**189**	**206**	**245**	**261**
2009	5	150	0	333	110	0	0	110	0	0	0	220	110	220	220	352	218	220	282	233	0	110	**203**	**238**	**259**	**309**	**335**
2009	6	154	417	0	110	0	0	0	0	0	330	0	330	0	330	330	0	0	131	0	0	**179**	**208**	**244**	**266**	**323**	**355**
2009	7	148	68	0	0	0	110	0	0	0	0	110	0	220	110	0	510	451	110	110	**204**	**175**	**203**	**239**	**265**	**326**	**352**
2009	8	146	220	0	123	0	0	0	110	0	0	0	110	110	366	0	0	171	0	**216**	**204**	**174**	**202**	**243**	**273**	**330**	**349**
2009	9	136	110	0	0	0	0	0	0	0	0	0	0	0	220	343	110	110	**188**	**206**	**194**	**167**	**197**	**239**	**264**	**313**	**325**
2009	10	123	250	0	0	0	0	0	0	110	281	119	110	479	330	0	121	**181**	**176**	**192**	**181**	**159**	**190**	**227**	**246**	**285**	**288**
2009	11	96	0	0	0	0	0	0	0	0	0	0	0	110	220	0	**188**	**151**	**147**	**161**	**155**	**137**	**161**	**189**	**200**	**226**	**223**
2009	12	77	110	0	0	0	0	0	0	110	0	0	110	0	289	**163**	**160**	**129**	**126**	**140**	**136**	**119**	**137**	**157**	**162**	**179**	**174**
2010	1	48	220	0	0	0	0	0	0	0	0	0	0	0	**153**	**116**	**114**	**92**	**91**	**103**	**99**	**85**	**96**	**107**	**108**	**118**	**140**
2010	2	149	121	0	0	0	0	110	0	0	0	0	0	**189**	**358**	**270**	**266**	**219**	**221**	**246**	**231**	**194**	**213**	**232**	**231**	**307**	**328**
2010	3	215	220	0	0	0	0	0	114	110	110	367	**90**	**250**	**472**	**358**	**359**	**299**	**297**	**324**	**298**	**243**	**261**	**281**	**341**	**407**	**432**
2010	4	178	110	0	0	0	151	110	0	0	0	**42**	**78**	**218**	**414**	**320**	**324**	**265**	**258**	**276**	**247**	**196**	**208**	**274**	**298**	**354**	**379**
2010	5	187	604	110	0	0	110	114	0	0	**43**	**44**	**82**	**228**	**441**	**344**	**344**	**276**	**263**	**273**	**239**	**187**	**243**	**286**	**310**	**371**	**403**
2010	6	176	429	0	0	0	0	0	0	**40**	**42**	**43**	**79**	**224**	**438**	**336**	**329**	**258**	**240**	**243**	**210**	**201**	**233**	**274**	**299**	**364**	**400**
2010	7	168	118	133	0	0	0	0	**22**	**39**	**40**	**41**	**78**	**225**	**432**	**326**	**312**	**238**	**215**	**216**	**227**	**195**	**226**	**267**	**297**	**365**	**395**
2010	8	124	367	0	175	0	110	**17**	**18**	**31**	**32**	**34**	**65**	**183**	**346**	**255**	**237**	**177**	**158**	**193**	**182**	**156**	**182**	**219**	**246**	**298**	**316**
2010	9	182	430	0	0	0	**23**	**22**	**24**	**42**	**44**	**47**	**88**	**244**	**451**	**323**	**294**	**216**	**236**	**259**	**244**	**210**	**249**	**302**	**334**	**397**	**412**
2010	10	169	0	0	0	**25**	**22**	**21**	**23**	**41**	**44**	**45**	**83**	**227**	**407**	**285**	**256**	**230**	**225**	**246**	**233**	**204**	**244**	**292**	**317**	**368**	**372**
2010	11	140	110	110	**26**	**22**	**19**	**19**	**20**	**37**	**39**	**40**	**71**	**188**	**329**	**228**	**249**	**201**	**196**	**215**	**208**	**184**	**217**	**255**	**270**	**306**	**301**
2010	12	102	0	**25**	**20**	**17**	**15**	**15**	**17**	**30**	**31**	**31**	**53**	**138**	**239**	**202**	**198**	**159**	**156**	**175**	**170**	**149**	**172**	**197**	**204**	**225**	**219**

Figure 2. Estimated Claims Amounts in the Sample Runoff Triangle for Two Year Warranty

In order to validate the model, the estimates fitted in the training data set are explored first. Since the training data set was used to build the models, the models fitting should be well. Not surprisingly, the observed claims happened is $325,732 and the estimated claims is $336,316, a difference of less than 3.15%.

Secondly, the estimated results are validated in the test data set, which contain data not used in estimating the model parameters. The test data set contains claims data after the cutoff, approximately 15% of the portfolio. Note that the real claims happened is $20,548 and the estimated claims is $21,584 in the test data set. The difference is less than 4.80%. Comparing to the result of the predication data set, PROC GAM provides decent claims forecasts.

Finally the estimated claims amounts is aggregated by failure date and plotted in Figure 3. As mentioned in the introduction section, a potential concern of the GAM model is over-fitting the data, that is, the estimates from the training data set will not generalize to the test data set. However, as shown in the Figure 3, the predictive performance of GAM looks similar before and after the cutoff date, i.e., between the training data set and the test data set. Therefore, PROC GAM can be used as a powerful directly predictive modeling tool by scrupulous model selections and rigorous model validations.

CONCLUSION

PROC GAM is capable of incorporating the nonparametric effects simultaneously with the distributional flexibility, which helps discover the nonlinear pattern of predictors and improves predictive performance as a result. Therefore, PROC GAM is well suited to model claims counts and claims size when assuming Poisson and Gamma distributions respectively. The suggested models give satisfactory results to improve prediction performances for estimating claims in the runoff triangles.

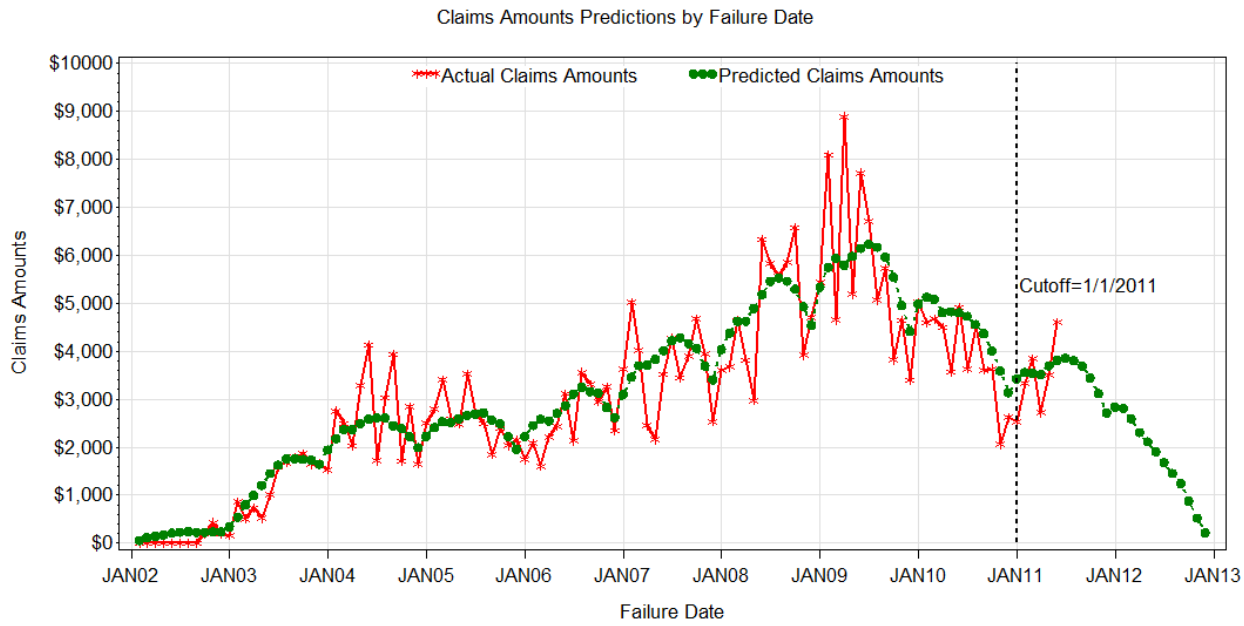


Figure 3. Claims Amounts Predictions by Failure Date in the Runoff Triangles

REFERENCES

Hastie, T. and Tibshirani, R. (1986), "Generalized Additive Models," Statistical Science, 3, 297-318.

Dong, X. (2001), "Fitting Generalized Additive Models with the GAM Procedure", SUGI Proceedings, Paper 256-26.

SAS Institute Inc. (2011), "SAS/STAT Users Guide", SAS Online Doc 9.3, Cary, NC, SAS Institute Inc.

CONTACT INFORMATION

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

Name: Ling Huang
Enterprise: Fulcrum Analytics Inc.,
Address: 55 Walls Drive, Suite 201,
City, State ZIP: Fairfield, CT 06824-5163
Work Phone: (515) 851-2445
E-mail: linghypshen@gmail.com
Web: http://www.fulcrum-mktg.com