Paper 333-2008

Using PROC GENMOD to find a fair house insurance rate for the Norwegian market

Sverre Grevskott and Sten Sture, Sparebank1 Non-life Insurance, Oslo, Norway

ABSTRACT

In this work we use PROC GENMOD from SAS/STAT® on house insurance data to identify different customer risk segments. Claim frequency and claim size are response variables, while we use customer age and the municipality where the customer lives as predictor variables. The predictor variables are categorical and we present a method on how to regroup categorical variable values if necessary. Claim frequency is assumed poisson distributed, claim size is assumed gamma distributed and the two are assumed independent. The final model is tested on data not used in the modelling from the same portfolio to determine how accurate our estimates are. Our conclusion is that PROC GENMOD is well suited to model insurance claims data and the suggested method for categorical variable value grouping works satisfactory.

INTRODUCTION

One of the core processes in an insurance company is to calculate accurate rates for the variety of risks underwritten. A fair and accurate insurance rate will ensure that each risk premium covers the expected average loss for the policy. In this work we show a method for modelling the average loss by use of PROC GENMOD [*SAS Institute Inc*]. The model is applied on aggregated house insurance policy data from the Norwegian marked, giving number of claims, total cost for the claims and number of policy years.

RISK MODEL

It is common to model the average loss as a product of two factors, the claim frequency and the claim size. The product is called pure premium, and given the number of risk years, we can calculate the total expected loss of the portfolio. However, pure premium will vary quite a lot between customer groups. We will therefore divide the customers into different risk groups and make sure that each group carries its own risk. A lot of factors may be taken into account when we calculate the rate for a particular customer group, but in this work we only use customer age and where the customer lives as rating variables. PROC GENMOD enables us to model frequency and size of the claims using a log linear regression model [*McCullagh, P. and Nelder, J.A.*]. We assume claim frequency to be poisson distributed and claim size to be gamma distributed as well as the two being independent.

INSURANCE DATA

The insurance data is available to us on an aggregated level with data on where the customers live (`Area`) and the age of the customers (`ageGroup`). In addition, we have the number of claims that the policy covers, the total cost of the claims and the number of policy years for each `Area/ageGroup` combination. The geography data is divided into 5 areas, which is a grouping of Norway's 431 municipalities based on size and urban development. The 7 age groups are in 10 year intervals from 20 to 89 years of age. `Area` and `ageGroup` will be used as categorical variables in PROC GENMOD to identify groups of customers with significant difference in claim frequency or claim size.

RESULTS

Before we run the data through GENMOD, we need to introduce some additional variables. We need the average claim and the logarithm of policy years. The logarithm of policy years is needed as an offset when running claims frequency modeling with the poisson distribution, while the average claim is used as a respons variable when modeling the claim size.

```
data insurancedata;
set insdatascaled;
AvgClaim=ClaimsCost/NbClaims;
l_exp=Log(PolicyYears);
run;
```

This gives us a table with the following columns:

Area	ageGroup	PolicyYears	NbClaims	ClaimsCost	AvgClaim	l_exp
A1	20-29	9 904	155	12 920 524	83 358	9,2007
A1	30-39	21 800	352	33 082 738	93 985	9,9897
A1	40-49	25 016	404	27 676 706	68 507	10,1273
A1	50-59	33 163	732	51 802 058	70 768	10,4092
A1	60-69	28 741	732	87 433 846	119 445	10,2661
…	…	…	…	…	…	…
A5	50-59	13 540	435	25 553 568	58 744	9,5134
A5	60-69	13 782	489	21 320 219	43 600	9,5311
A5	70-79	11 714	460	31 365 477	68 186	9,3685
A5	80-89	2 735	112	8 352 790	74 578	7,9137

The data table `insurancedata` is used as input to the GENMOD prodecure.

CLAIM FREQUENCY MODEL

We run PROC GENMOD with number of claims as response variable and Area and ageGroup as covariates:

```
Proc genmod data=insurancedata;
       Class Area     (param=ref)
             ageGroup (param=ref) ;
       Model NbClaims = Area ageGroup / dist=poisson
                                        link=log
                                        type3
                                        offset=l_exp;
       Output out=pred_poisson predicted=NbClaimsEstimate;
Run;
```

The class variables are listed with the `param=ref` option to use a reference level comparison and the design matrix is shown below in `Class Level Information`. The `Analysis Of Parameter Estimates` output will then give us significance in difference between a group level and the reference level. GENMOD chooses the last group value of the sorted group values as default reference level.

The model statement in GENMOD takes a number of options:

- `dist=poisson`, the number of claims are poisson distributed
- `link=log`, the link function is logarithmic
- `offset=l_exp`, the offset is given by the variable l_exp
- `type3`, we want a type 3 test in output

The output from GENMOD is:

```
Model Information

Data Set                     WORK.INSURANCEDATA
Distribution                            Poisson
Link Function                               Log
```

```
Dependent Variable              NbClaims
Offset Variable                    l_exp

Number of Observations Read          35
Number of Observations Used          35

                 Class Level Information
Class        Value                Design Variables

Area         A1          1      0      0      0
             A2          0      1      0      0
             A3          0      0      1      0
             A4          0      0      0      1
             A5          0      0      0      0

ageGroup     20-29       1      0      0      0      0      0
             30-39       0      1      0      0      0      0
             40-49       0      0      1      0      0      0
             50-59       0      0      0      1      0      0
             60-69       0      0      0      0      1      0
             70-79       0      0      0      0      0      1
             80-89       0      0      0      0      0      0

           Criteria For Assessing Goodness Of Fit

Criterion                 DF           Value        Value/DF

Deviance                  24         37.9292          1.5804
Scaled Deviance           24         37.9292          1.5804
Pearson Chi-Square        24         37.9962          1.5832
Scaled Pearson X2         24         37.9962          1.5832
Log Likelihood                     93644.0675

Algorithm converged.
```

The `Goodness Of Fit` listing in the output shows how well our assumption of poisson distributed frequencies fits the real data. The `Scaled Deviance` and `Scaled Pearson X2` are both close to one, indicating that the underlying distribution is indeed close to a poisson distribution.

Once we have established that we have a valid model, we observe the effect of the covariates. The result from the Type 3 test shows that both Area and ageGroup gives a significant contribution to the model and should be kept.

```
     LR Statistics For Type 3 Analysis

                          Chi-
Source           DF     Square    Pr > ChiSq

Area              4     252.83        <.0001
ageGroup          6     822.87        <.0001
```

In real life insurance rating, we use a lot more variables and then one have to decide which to keep if some variables are below the wanted significance level. In this simplified model, both variables are highly significant and are therefore kept in the model.

We also want to look at the significance level for all the covariate values listed in `Analysis Of Parameter Estimates.`

```
                              Analysis Of Parameter Estimates

                                        Standard     Wald 95% Confidence   Chi-
Parameter            DF    Estimate       Error           Limits          Square    Pr > ChiSq

Intercept             1     -3.1130      0.0357      -3.1829     -3.0431    7615.62     <.0001
Area         A1       1     -0.3825      0.0280      -0.4374     -0.3275     186.23     <.0001
```

Area	A2	1	-0.2319	0.0269	-0.2846	-0.1792	74.29	<.0001
Area	A3	1	-0.3621	0.0286	-0.4181	-0.3061	160.54	<.0001
Area	A4	1	-0.1668	0.0261	-0.2179	-0.1157	40.94	<.0001
ageGroup	20-29	1	-0.7510	0.0476	-0.8443	-0.6577	248.80	<.0001
ageGroup	30-39	1	-0.6528	0.0377	-0.7267	-0.5789	299.70	<.0001
ageGroup	40-49	1	-0.5030	0.0358	-0.5731	-0.4329	197.57	<.0001
ageGroup	50-59	1	-0.3148	0.0331	-0.3797	-0.2498	90.25	<.0001
ageGroup	60-69	1	-0.1999	0.0330	-0.2646	-0.1353	36.79	<.0001
ageGroup	70-79	1	-0.0710	0.0331	-0.1359	-0.0062	4.61	0.0318
Scale		0	1.0000	0.0000	1.0000	1.0000		

NOTE: The scale parameter was held fixed.

We observe that all values are significant with a significance level of 0.05. It is important to keep in mind that this is significance compared to the reference value. In this case the reference value for `Area` is A5 and reference value for `ageGroup` is 80-89. To reveal differences relative to other values, one should rerun the procedure with different reference level. This is done by specifying `ref=<value>` in the class statement. If we use A1 (`ref='A1'`) as reference level we get:

Analysis Of Parameter Estimates

Parameter		DF	Estimate	Standard Error	Wald 95% Confidence Limits		Chi-Square	Pr > ChiSq
Intercept		1	-3.4955	0.0329	-3.5600	-3.4309	11257.3	<.0001
Area	A2	1	0.1506	0.0235	0.1045	0.1966	41.09	<.0001
Area	A3	1	0.0204	0.0254	-0.0294	0.0701	0.64	0.4224
Area	A4	1	0.2157	0.0225	0.1715	0.2598	91.71	<.0001
Area	A5	1	0.3825	0.0280	0.3275	0.4374	186.23	<.0001
ageGroup	20-29	1	-0.7510	0.0476	-0.8443	-0.6577	248.80	<.0001
ageGroup	30-39	1	-0.6528	0.0377	-0.7267	-0.5789	299.70	<.0001
ageGroup	40-49	1	-0.5030	0.0358	-0.5731	-0.4329	197.57	<.0001
ageGroup	50-59	1	-0.3148	0.0331	-0.3797	-0.2498	90.25	<.0001
ageGroup	60-69	1	-0.1999	0.0330	-0.2646	-0.1353	36.79	<.0001
ageGroup	70-79	1	-0.0710	0.0331	-0.1359	-0.0062	4.61	0.0318
Scale		0	1.0000	0.0000	1.0000	1.0000		

NOTE: The scale parameter was held fixed.

As we observe the difference between `Area` value A3 and reference value A1 is not significant, and we therefore choose to combine these two values. If there are several values that are not significantly different, we choose to combine the two levels giving the lowest Chi Square. When we rerun the model with the new grouping of Area we get:

Criteria For Assessing Goodness Of Fit

Criterion	DF	Value	Value/DF
Deviance	25	38.5725	1.5429
Scaled Deviance	25	38.5725	1.5429
Pearson Chi-Square	25	38.6280	1.5451
Scaled Pearson X2	25	38.6280	1.5451
Log Likelihood		93643.7458	

Algorithm converged.

LR Statistics For Type 3 Analysis

Source	DF	Chi-Square	Pr > ChiSq
Area	3	252.18	<.0001
ageGroup	6	825.39	<.0001

Analysis Of Parameter Estimates

Parameter		DF	Estimate	Standard Error	Wald 95% Confidence Limits		Chi-Square	Pr > ChiSq
Intercept		1	-3.1128	0.0357	-3.1828	-3.0429	7615.19	<.0001
Area	A1A3	1	-0.3728	0.0253	-0.4224	-0.3233	217.35	<.0001
Area	A2	1	-0.2319	0.0269	-0.2846	-0.1792	74.29	<.0001
Area	A4	1	-0.1668	0.0261	-0.2179	-0.1157	40.93	<.0001
ageGroup	20-29	1	-0.7515	0.0476	-0.8448	-0.6582	249.23	<.0001
ageGroup	30-39	1	-0.6533	0.0377	-0.7272	-0.5794	300.24	<.0001
ageGroup	40-49	1	-0.5034	0.0358	-0.5735	-0.4332	197.89	<.0001
ageGroup	50-59	1	-0.3150	0.0331	-0.3799	-0.2500	90.35	<.0001
ageGroup	60-69	1	-0.2000	0.0330	-0.2646	-0.1354	36.81	<.0001
ageGroup	70-79	1	-0.0709	0.0331	-0.1358	-0.0061	4.60	0.0320
Scale		0	1.0000	0.0000	1.0000	1.0000		

NOTE: The scale parameter was held fixed.

The `Goodness Of Fit` and `Type 3 Analysis` still gives good values and `Analysis Of Parameter Estimates` shows that all values are significant. We rerun the model with all covariate values as reference value and find that no further grouping is necessary. The model for claim frequency is then established.

CLAIM SIZE MODEL

We then want to model the claim size using a gamma distribution in GENMOD:

```
Proc genmod data=insurancedata;
          Class Area     (param=ref)
                ageGroup (param=ref);
          Weight NbClaims;
          Model AvgClaim=Area ageGroup / dist=gamma
                                         link=log
                                         type3;
          Output out=pred_gamma predicted=AvgClaimEstimate;
Run;
```

Instead of an offset variable we now use the weight option `Weight` and the number of claims as weight value. The response variable is now the average claim and the assumed distribution is gamma. The output is given as:

The GENMOD Procedure

Model Information

Data Set	WORK.INSURANCEDATA
Distribution	Gamma
Link Function	Log
Dependent Variable	AvgClaim
Scale Weight Variable	NbClaims

Number of Observations Read	35
Number of Observations Used	35
Sum of Weights	17378

Class Level Information

Class	Value		Design Variables		
Area	A1	1	0	0	0

	A2	0	1	0	0	
	A3	0	0	1	0	
	A4	0	0	0	1	
	A5	0	0	0	0	
ageGroup	20-29	1	0	0	0	0	0
	30-39	0	1	0	0	0	0
	40-49	0	0	1	0	0	0
	50-59	0	0	0	1	0	0
	60-69	0	0	0	0	1	0
	70-79	0	0	0	0	0	1
	80-89	0	0	0	0	0	0

Criteria For Assessing Goodness Of Fit

Criterion	DF	Value	Value/DF
Deviance	24	326.3232	13.5968
Scaled Deviance	24	35.1752	1.4656
Pearson Chi-Square	24	336.6846	14.0285
Scaled Pearson X2	24	36.2920	1.5122
Log Likelihood		-370.0338	

Algorithm converged.

LR Statistics For Type 3 Analysis

Source	DF	Chi-Square	Pr > ChiSq
Area	4	49.46	<.0001
ageGroup	6	16.58	0.0110

The `Goodness Of Fit` and `Type 3 Analysis` output shows that we have a fairly good model with both variables contributing significantly to the model. The `Analysis Of Parameter Estimates` gives:

Analysis Of Parameter Estimates

Parameter		DF	Estimate	Standard Error	Wald 95% Confidence Limits		Chi-Square	Pr > ChiSq
Intercept		1	10.8400	0.1072	10.6298	11.0502	10216.4	<.0001
Area	A1	1	0.5674	0.0856	0.3996	0.7351	43.95	<.0001
Area	A2	1	0.1072	0.0820	-0.0536	0.2679	1.71	0.1912
Area	A3	1	0.3031	0.0871	0.1323	0.4739	12.10	0.0005
Area	A4	1	-0.0809	0.0794	-0.2366	0.0748	1.04	0.3084
ageGroup	20-29	1	-0.0558	0.1451	-0.3402	0.2286	0.15	0.7006
ageGroup	30-39	1	-0.0488	0.1148	-0.2739	0.1763	0.18	0.6709
ageGroup	40-49	1	-0.0454	0.1092	-0.2594	0.1685	0.17	0.6772
ageGroup	50-59	1	-0.0033	0.1010	-0.2012	0.1946	0.00	0.9740
ageGroup	60-69	1	0.0549	0.1004	-0.1420	0.2517	0.30	0.5848
ageGroup	70-79	1	0.2357	0.1008	0.0381	0.4332	5.47	0.0194
Scale		1	0.1078	0.0256	0.0676	0.1718		

NOTE: The scale parameter was estimated by maximum likelihood.

Contrary to the case with the number of claims modeling, most of the variable values are not significant. We rerun GENMOD with all covariate values as reference levels and find that `ageGroup` 50-59 and 80-89 are least different and should be grouped. By repeating this process, in each step combining the least different group values (for both `Area` and `ageGroup`), we end up with the model:

Class Level Information

Class	Value	Design Variables		
Area	A1	1	0	0
	A2	0	1	0
	A3	0	0	1
	A4A5	0	0	0
ageGroup	20-69,80-89	1		
	70-79	0		

Criteria For Assessing Goodness Of Fit

Criterion	DF	Value	Value/DF
Deviance	30	358.2537	11.9418
Scaled Deviance	30	35.1922	1.1731
Pearson Chi-Square	30	368.7795	12.2926
Scaled Pearson X2	30	36.2262	1.2075
Log Likelihood		-371.6760	

Algorithm converged.

LR Statistics For Type 3 Analysis

Source	DF	Chi-Square	Pr > ChiSq
Area	3	47.33	<.0001
ageGroup	1	13.95	0.0002

The 7 values of ageGroup have been reduced to only two, while only Area A4 and A5 have been combined. We observe that Goodness Of Fit and Type 3 Analysis shows slightly better values than our original model. The Analysis Of Parameter Estimates shows:

Analysis Of Parameter Estimates

Parameter		DF	Estimate	Standard Error	Wald 95% Confidence Limits		Chi-Square	Pr > ChiSq
Intercept		1	11.0192	0.0591	10.9034	11.1351	34748.2	<.0001
Area	A1	1	0.6274	0.0675	0.4950	0.7597	86.32	<.0001
Area	A2	1	0.1609	0.0627	0.0380	0.2839	6.58	0.0103
Area	A3	1	0.3613	0.0699	0.2244	0.4983	26.75	<.0001
ageGroup	20-69,80-89	1	-0.2367	0.0585	-0.3513	-0.1221	16.38	<.0001
Scale		1	0.0982	0.0234	0.0616	0.1565		

NOTE: The scale parameter was estimated by maximum likelihood.

All variable are now significantly different from each other and the model for claim size is established.

FINAL MODEL

When we have significant difference in all covariate variable values we need to calculate the average claim frequency and average claim size for each of the values. Since this is a log transformed model we need to calculate the exponential of all the estimates.

		Poisson		Gamma	
Parameter	Level1	Estimate	Exp(Estimate)	Estimate	Exp(Estimate)
Intercept		-3,113	0,044	11,019	61 037
Area	A1	-0,373	0,689	0,627	1,873
Area	A2	-0,232	0,793	0,161	1,175
Area	A3	-0,373	0,689	0,361	1,435
Area	A4	-0,167	0,846	0,000	1,000
Area	A5	0,000	1,000	0,000	1,000
ageGroup	20-29	-0,752	0,472	-0,237	0,789
ageGroup	30-39	-0,653	0,520	-0,237	0,789
ageGroup	40-49	-0,503	0,604	-0,237	0,789
ageGroup	50-59	-0,315	0,730	-0,237	0,789
ageGroup	60-69	-0,200	0,819	-0,237	0,789
ageGroup	70-79	-0,071	0,932	0,000	1,000
ageGroup	80-89	0,000	1,000	-0,237	0,789

The average claim frequency for customers in `Area` A1 and in the `ageGroup` 20-29 is then:

```
0,044 * 0,689 * 0,472 = 0,014
```

In the same way we calculate the average claim size for this group to be

```
61037 * 1,873 * 0,789 = 90211
```

The pure premium for this group is then 0,014*90211=1263. This means that customers living in area A1 and being between 20 and 29 years of age should pay a yearly premium of 1263 for their house insurance if we want the premiums to exactly cover the claims.

TESTING THE MODEL

To test the model, we fist look at how well the estimates fits the data in the dataset `insurancedata`. Since these data was used to establish the model we expect the data to fit quite well. We calculate the pure premiums for all customer groups and multiply with the number of policy years in each group. This gives us the total claims cost for each customer group, and the sum of all customer groups gives us the estimated claims cost for the entire portfolio. We then compare this with the real claims cost registered in the period, and find that the difference is less than 0,05%.

PolicyYears	Registered Claims Cost	Estimated Claims Cost
680 762	1 133 385 732	1 133 860 359

This shows that PROC GENMOD estimates the model data very well.

However, we want to test the model on data not used in estimating the model parameters. The dataset insurancedata only contains 70% of the original portfolio data, while 30% of the data was withheld to test the final model. We estimate the total claims cost for this part of the portfolio and compare with the real claims cost registered.

PolicyYears	Registered Claims Cost	Estimated Claims Cost
292 251	479 595 442	485 979 699

The difference is less than 1,4% and indicates that the model is estimating the customer risk level well.

CONCLUSION

PROC GENMOD is well suited to model claim freqeucy and claim size when assuming poisson and gamma distributions. The suggested method for grouping of variable values seems to give satisfactory results and does improve goodness of fit values.

REFERENCES

[**McCullagh, P. and Nelder, J.A**.] (1989), Generalized Linear Models, Second Edition, Longon: Chapman and Hall.

[**SAS Institute Inc**] (2007). SAS/STAT User's Guide, SAS OnlineDoc 9.1.3, Cary, NC: SAS Institute Inc.

CONTACT INFORMATION

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

Sverre Grevskott
Sparebank1 Non-life insurance
Hammersborggata 2
0106 Oslo, NORWAY
E-mail: sverre.grevskott@sparebank1.no