Name: LISUJUAN
Regression Analysis Student Project
Using GLM to estimate the average claim cost
Fall, 2015

Student Project Topic:

Using GLM to estimate the average claim cost

1. Background and Objective

The object of this project is to use GLM to estimate the average claim cost for auto. This project has 5 variables: owner age, car model, car age, number of claim, average claim costs.

2. The Initial data and Histogram

All policies are divided into 128 categories, of which there are 5 categories with No claims. The data has 8 owner age levels, 4 car model levels and 4car age levels.
(Data Source: http://www.statsci.org/data/general/carinsuk.html)

(a)The owner age levels:

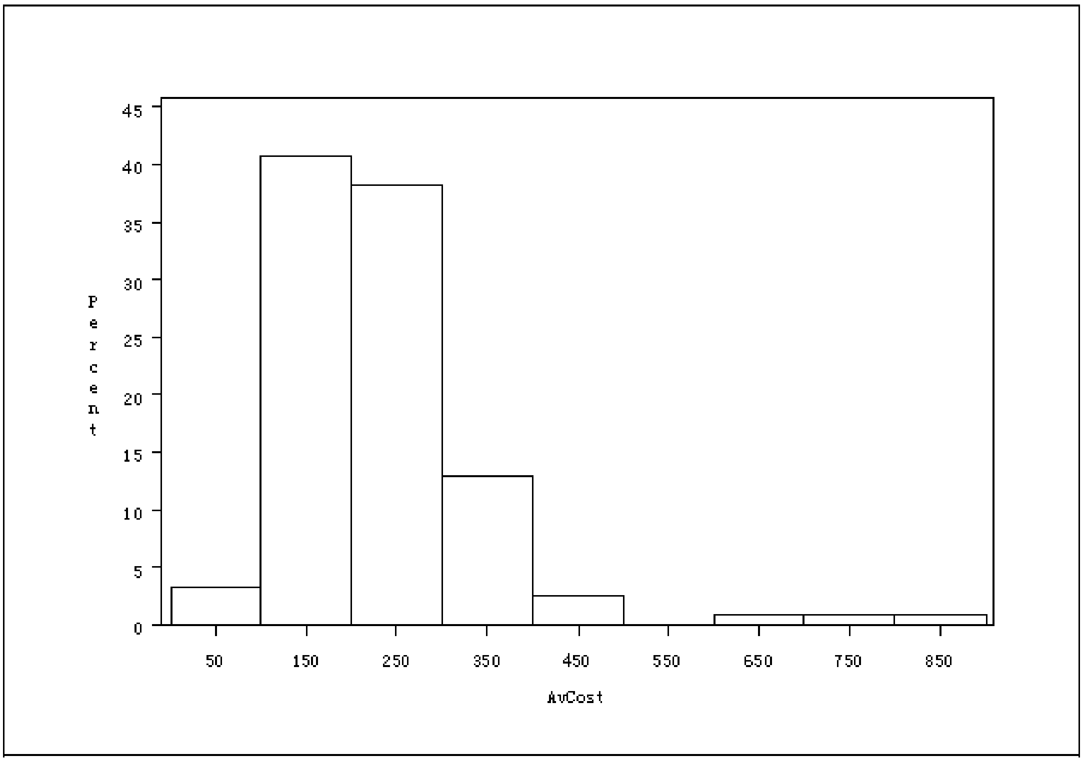
17-20, 21-24, 25-29, 30-34, 35-39, 40-49, 50-59, 60+

(b)The car model levels: A, B, C, D

(c)The car age levels: 0-3, 4-7, 8-9, 10

The Histogram as follows:

The histogram of average claim cost



From graph above, we can see that the average claim cost is difference among each risk category. Distribution significantly deviates from the normal distribution, there is a heavy tail. The largest value is 850. Also we can know that ownerage, car model, carage are all important factors to influencing average claim cost, and we can use GLM to estimate the influence degree to average claim cost.

3. GLMs

A generalized linear model (GLM) consists of three components:

(a) Random component, specifying the conditional distribution of the response variable,Y_i given the values of the explanatory variables in the model. In the initial formulation of GLMs, the distribution of the response variable of Y_i was a member of an exponential family, such as Gaussian, binomial, Poisson, Gamma, or inverse-Gaussian families of distributions.

(b) A linear predictor—that is a linear function of regressors

$$\eta_i = \alpha + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + +\beta_k X_{ik}$$

(c) Link function $g(.)$, which transforms the expectation of the response variable, $\mu_i \equiv E(Y_i)$, to the linear predictor:

$$g(\mu_i) = \eta_i = \alpha + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + +\beta_k X_{ik}$$

Since the link function is invertible, we can also write

$$\mu_i = g^{-1}(\eta_i) = g^{-1}(\alpha + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + +\beta_k X_{ik})$$

Based on the property of data, we use GLMs to estimate the average claim cost.

4. Using Gamma distribution to estimate average claim cost

In this model, we use gamma error structure with Logit link function.

To Gamma distribution, the variance $\mathrm{V(x)} = x^2$

To Logit link function, $g(\mu_i) = ln(x/(1-x))$ $\quad g^{-1}(\eta_i) = e^x/(1+e^x)$

Hence, establishing the GLMs with SAS:

(1)SAS code:

```
proc genmod data=a1;
class ownerage model carage;
weight nclaims;
model avcost= ownerage model carage/dist=gamma link=log type1 type3;
run;
```

(2) Model results

(2a)Model information

The GENMOD Procedure

Model Information

Data Set	WORK.TEST1	
Distribution	Gamma	
Link Function	Log	
Dependent Variable	avcost	avcost
Scale Weight Variable	Nclaims	Nclaims

Number of Observations Read	128
Number of Observations Used	123
Sum of Weights	8942
Missing Values	5

(2b) Class level information

Class Level Information

Class	Levels	Values
ownerage	8	17-20 21-24 25-29 30-34 35-39 40-49 50-59 60+
model	4	A B C D
carage	4	0-3 4-7 8-9 z10+

(2c) Criteria for assessing goodness of fit

From table, we can conclude that $\Pr(\chi_{109} > 125.2616) = 0.1366$, this shows that Gamma distribution can fit model well.

Criteria For Assessing Goodness Of Fit

Criterion	DF	Value	Value/DF
Deviance	109	127.1440	1.1665
Scaled Deviance	109	125.2616	1.1492
Pearson Chi-Square	109	126.5314	1.1608
Scaled Pearson X2	109	124.6581	1.1437
Log Likelihood		-623.9230	

Algorithm converged.

(2d) Analysis of Maximum Likelihood Parameter Estimates

From table below, exclude ownerage “40-49” and “50-59”, other estimates are all obvious significant, since P value for ownerage “40-49” and “50-59” are 0.7197 and 0.8536, In practice, we can merge this level with the nearest level to one level.

Analysis Of Parameter Estimates

Parameter		DF	Estimate	Standard Error	Wald 95% Confidence Limits		Chi-Square	Pr > ChiSq
Intercept		1	5.1338	0.0637	5.0090	5.2586	6499.85	<.0001
ownerage	17-20	1	0.2263	0.1107	0.0094	0.4433	4.18	0.0409
ownerage	21-24	1	0.2287	0.0598	0.1115	0.3459	14.63	0.0001
ownerage	25-29	1	0.1642	0.0438	0.0783	0.2502	14.04	0.0002
ownerage	30-34	1	0.1143	0.0420	0.0321	0.1966	7.43	0.0064
ownerage	35-39	1	-0.0877	0.0411	-0.1684	-0.0071	4.54	0.0330
ownerage	40-49	1	-0.0129	0.0358	-0.0831	0.0574	0.13	0.7197
ownerage	50-59	1	0.0069	0.0372	-0.0661	0.0799	0.03	0.8536
ownerage	60+	0	0.0000	0.0000	0.0000	0.0000	.	.
model	A	1	-0.4005	0.0429	-0.4845	-0.3165	87.29	<.0001
model	B	1	-0.4000	0.0354	-0.4694	-0.3307	127.85	<.0001
model	C	1	-0.2450	0.0364	-0.3164	-0.1735	45.20	<.0001
model	D	0	0.0000	0.0000	0.0000	0.0000	.	.
carage	0-3	1	0.6990	0.0516	0.5978	0.8002	183.32	<.0001
carage	4-7	1	0.6130	0.0516	0.5119	0.7141	141.33	<.0001
carage	8-9	1	0.3558	0.0598	0.2386	0.4730	35.43	<.0001
carage	z10+	0	0.0000	0.0000	0.0000	0.0000	.	.
Scale		1	0.9852	0.1234	0.7707	1.2594		

NOTE: The scale parameter was estimated by maximum likelihood.

(2e) Tests

From Type 1 and Type 3, we can conclude that these 3 explanatory variables are obvious significant.

The GENMOD Procedure

LR Statistics For Type 1 Analysis

Source	2*Log Likelihood	DF	Chi-Square	Pr > ChiSq
Intercept	-1456.5376			
ownerage	-1438.7844	7	17.75	0.0131
model	-1370.3694	3	68.41	<.0001
carage	-1247.8460	3	122.52	<.0001

LR Statistics For Type 3 Analysis

Source	DF	Chi-Square	Pr > ChiSq
ownerage	7	52.81	<.0001
model	3	100.54	<.0001
carage	3	122.52	<.0001

(3) Comparison

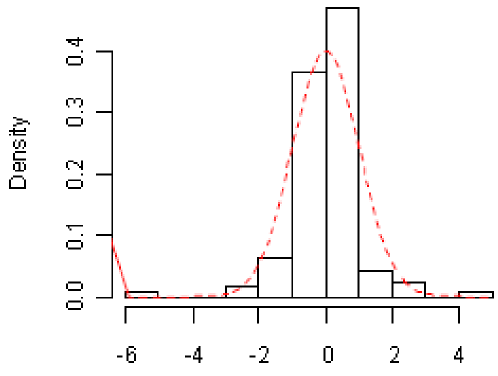
Given Logit link function, the error structure of normal distribution, Gamma distribution and Inverse Gaussian are as follows:

Parameter Estimation for GLM with 3 distributions

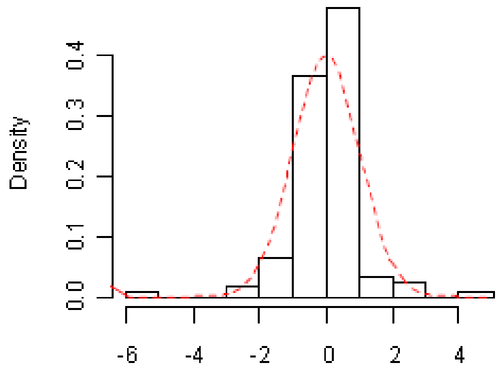
		normal		gamma		Inverse gaussian	
Parameter		Estimate	Pr>ChiSq	Estimate	Pr>ChiSq	Estimate	Pr>ChiSq
Intercept		5.1323	<.0001	5.1338	<.0001	5.1315	<.0001
OwnerAge	17-20	0.2676	0.0052	0.2263	0.0409	0.2085	0.1046
OwnerAge	21-24	0.2102	0.0002	0.2287	0.0001	0.2321	0.0008
OwnerAge	25-29	0.1381	0.0017	0.1642	0.0002	0.1796	0.0002
OwnerAge	30-34	0.1212	0.0043	0.1143	0.0064	0.1057	0.0203
OwnerAge	35-39	-0.1316	0.0055	-0.0877	0.033	-0.0685	0.1068
OwnerAge	40-49	-0.0159	0.6847	-0.0129	0.7197	-0.0117	0.7544
OwnerAge	50-59	-0.0067	0.8707	0.0069	0.8536	0.0142	0.7136
OwnerAge	60+	0	.	0	.	0	.
Model	A	-0.3893	<.0001	-0.4005	<.0001	-0.4085	<.0001
Model	B	-0.4066	<.0001	-0.4	<.0001	-0.3958	<.0001
Model	C	-0.2515	<.0001	-0.245	<.0001	-0.2431	<.0001
Model	D	0	.	0	.	0	.
CarAge	0-3	0.7253	<.0001	0.699	<.0001	0.6908	<.0001
CarAge	4-7	0.6189	<.0001	0.613	<.0001	0.6119	<.0001
CarAge	8-9	0.3528	0.002	0.3558	<.0001	0.3579	<.0001
CarAge	z10+	0	.	0	.	0	.
Scale		258.1265	0.9852	0.0718			

The Goodness of Fit for 3 distributions

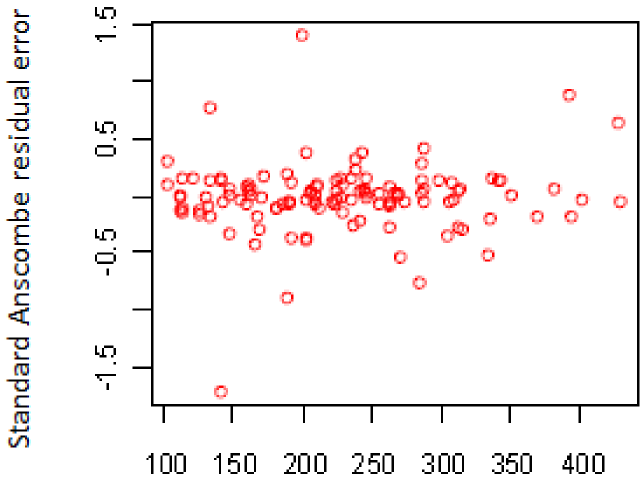
	normal	gamma	Inverse gaussian
Deviance	8195413.3	127.144	0.6342
Scaled Deviance	123.0002	125.2616	123
Pearson Chi-Square	8195413.3	126.5314	0.5629
Scaled Pearson Chi-Square	123.0002	124.6581	109.1609
Log Likelihood	-648.2247	-623.923	-624.548



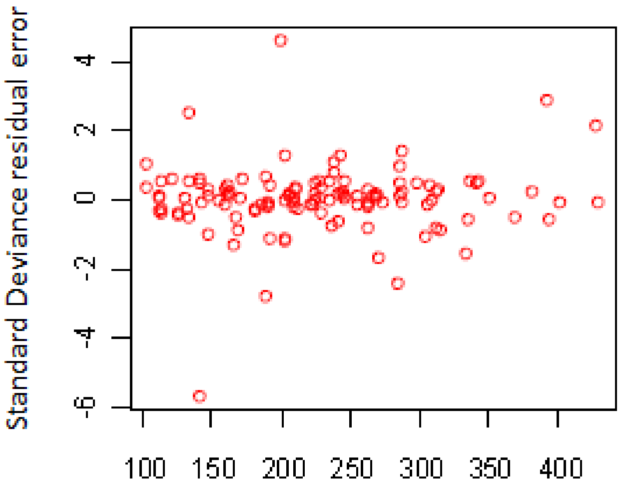
Standard Anscombe residual error



Standard Deviance residual error



The estimate of average claim cost



The estimate of average claim cost

From this index of goodness fit, we can know that the normal distribution is worse. From distribution diagram of Standard Anscombe residual error and Standard Deviance residual error, we can know that Gamma distribution is best.

(4) The estimated result

At last, the estimate of average claim cost is as follows:

Carage		0-3	4-7	8-9	10+
ownerage	model				
17-20	A	287	263	203	143
17-20	B	287	263	204	143
17-20	C	335	307	238	167
17-20	D	428	393	304	213
21-24	A	287	264	204	143
21-24	B	288	264	204	143
21-24	C	336	308	238	167
21-24	D	429	394	304	213
25-29	A	270	248	191	134
25-29	B	270	247	191	134
25-29	C	315	289	223	157
25-29	D	402	369	285	200
30-34	A	256	235	182	127
30-34	B	257	235	182	127
30-34	C	300	275	213	149
30-34	D	383	351	272	190
35-39	A	209	192	149	104
35-39	B	210	192	149	104
35-39	C	245	225	174	122

35-39	D	313	287	222	155
40-49	A	226	207	160	112
40-49	B	226	207	160	112
40-49	C	264	242	187	131
40-49	D	337	309	239	167
50-59	A	230	211	163	114
50-59	B	230	211	163	115
50-59	C	269	247	191	134
50-59	D	344	315	244	171
60+	A	229	210	162	114
60+	B	229	210	162	114
60+	C	267	245	190	133
60+	D	341	313	242	170

5. Combine the adjacent level

From above, we know ownerage “40-49” and “50-59” are not obvious significant, consider combine “40-49” , “50-59” and “60+” to “40+”,.

Using the Gamma distribution and Logit link function to establish GLMs again, the results are as follows:

The result shows that the parameter estimate and statistical significance are better.

The GENMOD Procedure

Model Information

Data Set	WORK.TEST2	
Distribution	Gamma	
Link Function	Log	
Dependent Variable	avcost	avcost
Scale Weight Variable	Mclaims	Mclaims

Number of Observations Read	128
Number of Observations Used	123
Sum of Weights	8942
Missing Values	5

Class Level Information

Class	Levels	Values
ownerage	6	17-20 21-24 25-29 30-34 35-39 40+
model	4	A B C D
carage	4	0-3 4-7 8-9 z10+

Criteria For Assessing Goodness Of Fit

Criterion	DF	Value	Value/DF
Deviance	111	127.5444	1.1490
Scaled Deviance	111	125.2682	1.1285
Pearson Chi-Square	111	127.1805	1.1458
Scaled Pearson X2	111	124.9108	1.1253
Log Likelihood		-624.1199	

Algorithm converged.

Analysis Of Parameter Estimates

Parameter		DF	Estimate	Standard Error	Wald 95% Confidence Limits		Chi-Square	Pr > ChiSq
Intercept		1	5.1300	0.0589	5.0146	5.2454	7590.26	<.0001
ownerage	17-20	1	0.2294	0.1080	0.0177	0.4411	4.51	0.0336
ownerage	21-24	1	0.2318	0.0543	0.1253	0.3383	18.20	<.0001
ownerage	25-29	1	0.1674	0.0359	0.0969	0.2378	21.68	<.0001
ownerage	30-34	1	0.1175	0.0335	0.0518	0.1833	12.29	0.0005
ownerage	35-39	1	-0.0846	0.0326	-0.1484	-0.0207	6.74	0.0094
ownerage	40+	0	0.0000	0.0000	0.0000	0.0000	.	.
model	A	1	-0.3985	0.0427	-0.4823	-0.3147	86.91	<.0001
model	B	1	-0.3987	0.0354	-0.4680	-0.3294	127.12	<.0001
model	C	1	-0.2440	0.0365	-0.3155	-0.1726	44.81	<.0001
model	D	0	0.0000	0.0000	0.0000	0.0000	.	.
carage	0-3	1	0.6986	0.0517	0.5973	0.7998	182.77	<.0001
carage	4-7	1	0.6124	0.0516	0.5113	0.7135	140.91	<.0001
carage	8-9	1	0.3557	0.0599	0.2384	0.4730	35.31	<.0001
carage	z10+	0	0.0000	0.0000	0.0000	0.0000	.	.
Scale		1	0.9822	0.1231	0.7683	1.2555		

The GENMOD Procedure

LR Statistics For Type 1 Analysis

Source	2*Log Likelihood	DF	Chi-Square	Pr > ChiSq
Intercept	-1456.5376			
ownerage	-1439.2789	5	17.26	0.0040
model	-1370.5242	3	68.75	<.0001
carage	-1248.2399	3	122.28	<.0001

LR Statistics For Type 3 Analysis

Source	DF	Chi-Square	Pr > ChiSq
ownerage	5	52.41	<.0001
model	3	100.30	<.0001
carage	3	122.28	<.0001

6. Conclusion

This project uses the GLMs with Gamma error structure and Logit link function to estimate average claim cost.

By analyzing the GLMs, this report shows that the result fit goodness.

Attachment: The initial data.

The initial data of GLMs.xls