Chapter 8: Continuous responses

8.1 *In the Enterprise Miner data set, develop a statistical model for claim size, amongst policies which had a claim.*

In the preliminary data analysis ("SAS Miner preliminary") we found that `bluebook`, `car_type` and `mvr_pts` are potential explanatory variables for `clm_amt`. Fitting the model with the Gamma response distribution and log link, the Type 3 test shows that `bluebook` (quadratic form), `mvr_pts` (banded version), `npolicy1` and `density` are significant.

All the above variables are highly significant in a multiple regression. The AIC may be used to confirm that the model below is optimal.

```
 /* Changes as detailed in "SAS miner preliminary analysis" document */
 data claims;
 set exercise.claims_sas_miner;

    bluebk = bluebook/1000-14.2; /* more stable for computation*/

    if npolicy=1 then npolicy1=1; else npolicy1=0;  /*dichotomise loyalty variable*/

* Categorize income because of zero spike;
    if income=. then incomecat=.;
    else if income=0 then incomecat=99;  /* base level*/
    else if income<=25000 then incomecat=2;
    else if income<=50000 then incomecat=3;
    else if income<=75000 then incomecat=4;
    else if income<=100000 then incomecat=5;
    else incomecat=6;

* Categorize oldclaim because of zero spike;
    if oldclaim=. then oldclaimcat=.;
    else if oldclaim=0 then oldclaimcat=99; /* base level*/
    else if oldclaim<=5000 then oldclaimcat=2;
    else if oldclaim<=10000 then oldclaimcat=3;
    else oldclaimcat=4;

* Categorize mvr_pts;
    if mvr_pts=. then mvrcat=.;
    else if mvr_pts=0 then mvrcat=99; /* base level*/
    else if mvr_pts=1 then mvrcat=1;
    else if mvr_pts=2 then mvrcat=2;
    else if mvr_pts=3 then mvrcat=3;
    else if mvr_pts<=6 then mvrcat=4;
    else mvrcat=5;

 run;
******** end of data transformations ;

 proc genmod data = claims;
 ods output obstats=claimsobs;
 ods listing exclude obstats;
 class npolicy1 mvrcat density (ref="Urban") / param=ref ;
 model clm_amt = mvrcat bluebk bluebk*bluebk npolicy1 density
    /  dist=gamma link = log type1 type3 obstats;
 where clm_flag="Yes";  /*remove non claims*/
 run;
```

```
                                The GENMOD Procedure

                                 Model Information

                Data Set              WORK.CLAIMS
                Distribution                Gamma
                Link Function                 Log
                Dependent Variable        CLM_AMT    Claim Amount

                          Number of Observations Read        2746
```

Number of Observations Used 2746

Class Level Information

Class	Value	Design Variables				
npolicy1	0	1				
	1	0				
mvrcat	1	1	0	0	0	0
	2	0	1	0	0	0
	3	0	0	1	0	0
	4	0	0	0	1	0
	5	0	0	0	0	1
	99	0	0	0	0	0
DENSITY	Highly Rural	1	0	0		
	Highly Urban	0	1	0		
	Rural	0	0	1		
	Urban	0	0	0		

Criteria For Assessing Goodness Of Fit

Criterion	DF	Value	Value/DF
Deviance	2734	1906.1893	0.6972
Scaled Deviance	2734	3024.5955	1.1063
Pearson Chi-Square	2734	4105.1339	1.5015
Scaled Pearson X2	2734	6513.7126	2.3825
Log Likelihood		-26269.9604	

Algorithm converged.

Analysis Of Parameter Estimates

Parameter		DF	Estimate	Standard Error	Wald 95% Confidence Limits		Chi-Square	Pr > ChiSq
Intercept		1	8.6621	0.0372	8.5893	8.7350	54358.1	<.0001
mvrcat	1	1	-0.0635	0.0501	-0.1616	0.0346	1.61	0.2048
mvrcat	2	1	0.1087	0.0516	0.0076	0.2097	4.44	0.0350
mvrcat	3	1	-0.0134	0.0524	-0.1161	0.0892	0.07	0.7974
mvrcat	4	1	0.1070	0.0422	0.0243	0.1898	6.42	0.0113
mvrcat	5	1	0.1138	0.0563	0.0035	0.2241	4.09	0.0432
bluebk		1	0.0251	0.0022	0.0208	0.0295	128.53	<.0001
bluebk*bluebk		1	-0.0006	0.0002	-0.0009	-0.0003	17.98	<.0001
npolicy1	0	1	0.0859	0.0308	0.0256	0.1463	7.78	0.0053
DENSITY	Highly Rural	1	0.1089	0.1436	-0.1726	0.3904	0.58	0.4483
DENSITY	Highly Urban	1	-0.1158	0.0325	-0.1796	-0.0520	12.66	0.0004
DENSITY	Rural	1	-0.1238	0.0832	-0.2869	0.0394	2.21	0.1371
Scale		1	1.5867	0.0391	1.5118	1.6653		

NOTE: The scale parameter was estimated by maximum likelihood.

LR Statistics For Type 1 Analysis

Source	2*Log Likelihood	DF	Chi-Square	Pr > ChiSq
Intercept	-52708.926			
mvrcat	-52693.912	5	15.01	0.0103
bluebk	-52578.151	1	115.76	<.0001
bluebk*bluebk	-52563.053	1	15.10	0.0001
npolicy1	-52554.864	1	8.19	0.0042
DENSITY	-52539.921	3	14.94	0.0019

LR Statistics For Type 3 Analysis

Source	DF	Chi-Square	Pr > ChiSq
mvrcat	5	18.27	0.0026
bluebk	1	122.52	<.0001
bluebk*bluebk	1	16.30	<.0001
npolicy1	1	7.79	0.0053
DENSITY	3	14.94	0.0019

Diagnostics

- **Anscombe residuals**

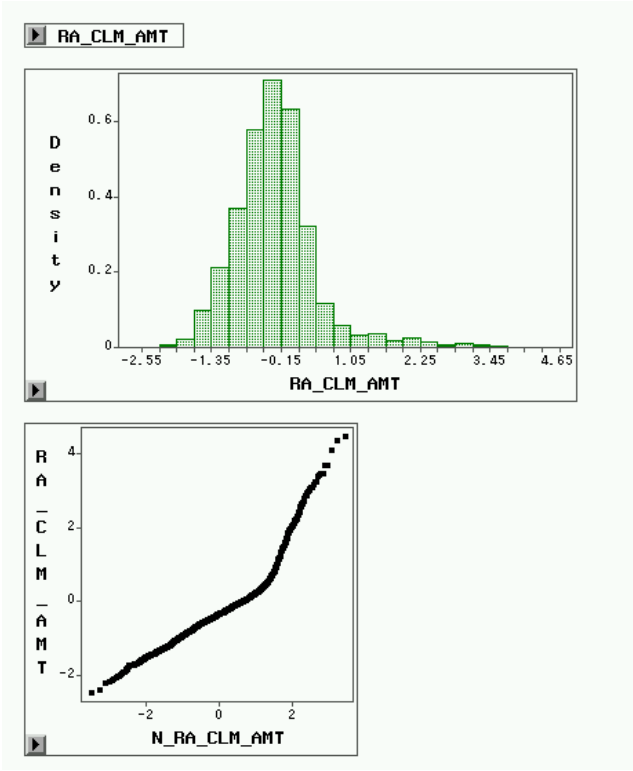

Figure 1: Anscombe residuals

Figure 1 shows that there is some evidence of non–normality in the distribution of the the Anscombe residuals but this is not severe. As Anscombe residuals are generated in SAS Insight but not by the command language, this graph has been produced in Insight.

- **Link**
  To check for the appropriateness of the link function, we need to calculate the

approximation for $g(y_i)$, as in equation (5.15). For the log link,

$$\begin{aligned} g(\mu_i) &= \ln \mu_i \\ \dot{g}(\mu_i) &= 1/\mu_i \\ g(y_i) &\approx g(\mu_i) + \dot{g}(\mu_i)(y_i - \mu_i) \\ &= \ln \mu_i + \frac{(y_i - \mu_i)}{\mu_i} \\ &= x_i'\beta + \frac{y_i}{\mu_i} - 1 \ . \end{aligned}$$

To estimate $g(y_i)$, we use the estimates $x_i'\hat{\beta}$ and $\hat{\mu}_i$, which are available in the `obstats` file as `xbeta` and `pred`, respectively:

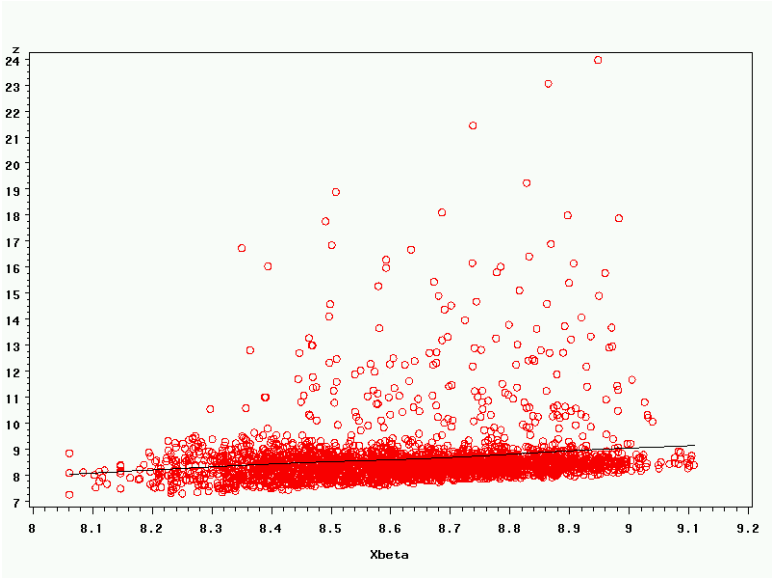

Figure 2: Link function diagnostic plot

```
data claimsobs;
set claimsobs;
z = xbeta + clm_amt/pred -1;    /*estimate of g(yi)*/
run;

symbol1 value=circle color=red;
symbol2 value=none interpol=sm60s color=black;

proc gplot data=claimsobs;
plot (z z)*xbeta / overlay;  /*plot g(yi) vs xbeta with spline overlaid*/
run;
```

As shown in Figure 2, there is an approximately linear relationship between the estimate of $g(y_i)$ and $x_i'\hat{\beta}$. Therefore, the log link function is appropriate for this model.

- **Residual deviance** The residual deviance graph looks reasonable. There are some large positive values, indicating that the fit in the upper tail of the distribution is not as good as it should be. For this reason, the inverse Gaussian is worth trying as an alternative to the gamma response distribution.
- **Hat matrix, Cook's distances** There appears to be a problem with the computation of these quantities, in SAS Insight. They are not produced by the command language.

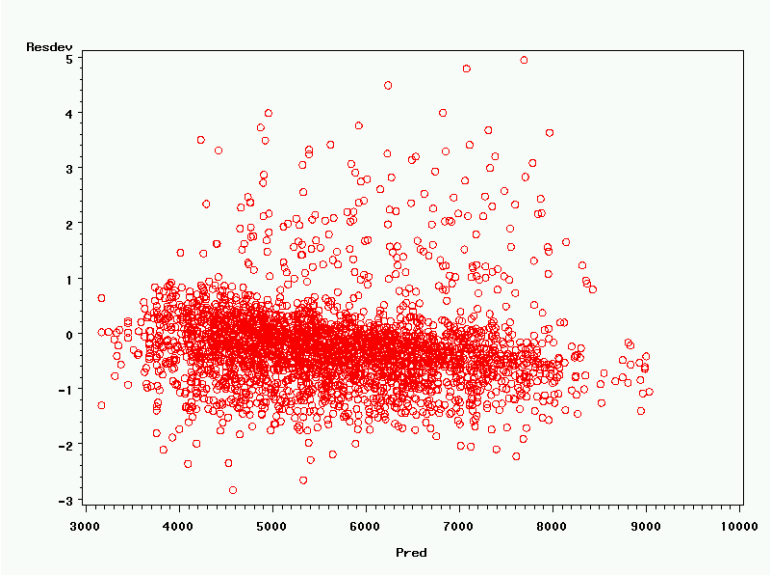

Figure 3: Residual deviance

- **Added–variable plots** The only continuous variable is `bluebk`, which is already in the model in quadratic form. We therefore do not examine its added–variable plot.

The final model

$$y \sim \mathrm{G}(\hat{\mu}, \hat{\nu} = 1.59)$$

$$\begin{aligned} \ln \hat{\mu} \quad = \quad & 8.6621 - 0.0635x_1 + 0.1087x_2 - 0.0134x_3 + 0.1070x_4 \\ & +0.1138x_5 + 0.0251x_6 - 0.0006x_6^2 + 0.0859x_7 \\ & +0.1089x_8 - 0.1158x_9 - 0.1238x_{10} \end{aligned}$$

where

$$
\begin{aligned}
x_1 &= \begin{cases} 1 & \text{if } \texttt{mvr\_pts} = 1 \\ 0 & \text{otherwise} \end{cases} \\
x_2 &= \begin{cases} 1 & \text{if } \texttt{mvr\_pts} = 2 \\ 0 & \text{otherwise} \end{cases} \\
x_3 &= \begin{cases} 1 & \text{if } \texttt{mvr\_pts} = 3 \\ 0 & \text{otherwise} \end{cases} \\
x_4 &= \begin{cases} 1 & \text{if } 4 \leq \texttt{mvr\_pts} \leq 6 \\ 0 & \text{otherwise} \end{cases} \\
x_5 &= \begin{cases} 1 & \text{if } \texttt{mvr\_pts} \geq 7 \\ 0 & \text{otherwise} \end{cases} \\
x_6 &= \text{bluebook}/1,000 - 14.2 \\
x_7 &= \begin{cases} 1 & \text{if number of policies } = 1 \\ 0 & \text{if number of policies } > 1 \end{cases} \\
x_8 &= \begin{cases} 1 & \text{if density = highly rural} \\ 0 & \text{otherwise} \end{cases} \\
x_9 &= \begin{cases} 1 & \text{if density = highly urban} \\ 0 & \text{otherwise} \end{cases} \\
x_{10} &= \begin{cases} 1 & \text{if density = rural} \\ 0 & \text{otherwise} \end{cases}
\end{aligned}
$$

Parameter interpretation

Parameter	Level	$\hat{\beta}$	$e^{\hat{\beta}}$
Intercept		8.662	5779.659
mvrcat	1	-0.064	0.938
mvrcat	2	0.109	1.115
mvrcat	3	-0.013	0.987
mvrcat	4	0.107	1.113
mvrcat	5	0.114	1.121
bluebk		0.025	1.025
bluebk*bluebk		-0.001	0.999
npolicy1	0	0.086	1.09
density	HighlyRural	0.109	1.115
density	HighlyUrban	-0.116	0.891
density	Rural	-0.124	0.884

- Base level is a customer with `mvr_pts`=0; `bluebook`=14 200; one policy with the company; living in urban area.
- Expected claim amount for the base level is \$5 779.66.
- The effect of having 7 to 12 motor vehicle points (`mvrcat`=5), compared with 0 points, is an increase in expected claim amount of 12.1%. Interpretation for other values of `mvrcat` is similar.
- The effect of having two or more policies with the company, is an increase in expected claim amount of 9%.
- The effect of living in a highly rural area, compared with urban, is an increase in expected claim amount of 11.5%. Interpretation for other values of `density` is similar.
- Because bluebook is in the model in quadratic form, the effect of an increase in bluebook depends on the starting point, and cannot be stated as above.

8.2 *In the personal injury data set, what is the impact of the level of injury on the dollar value of claims?*

Boxplots of claim size by injury code are not revealing (Figure 4); those of log claim size (Figure 5) demonstrate an increasing trend of log claim size with increasing injury code, until injury code 4, thereafter declining.

A gamma regression yields satisfactory results; inverse Gaussian regression gives an error message so is questionable.

```
proc genmod data=act.persinj;
class inj1 (ref="1") / param=ref;
model total = inj1 / dist=gamma link=log type3 ;
run;
```

```
                              The GENMOD Procedure

                               Model Information

                Data Set              ACT.PERSINJ
                Distribution                Gamma
                Link Function                 Log
                Dependent Variable          TOTAL    TOTAL
```

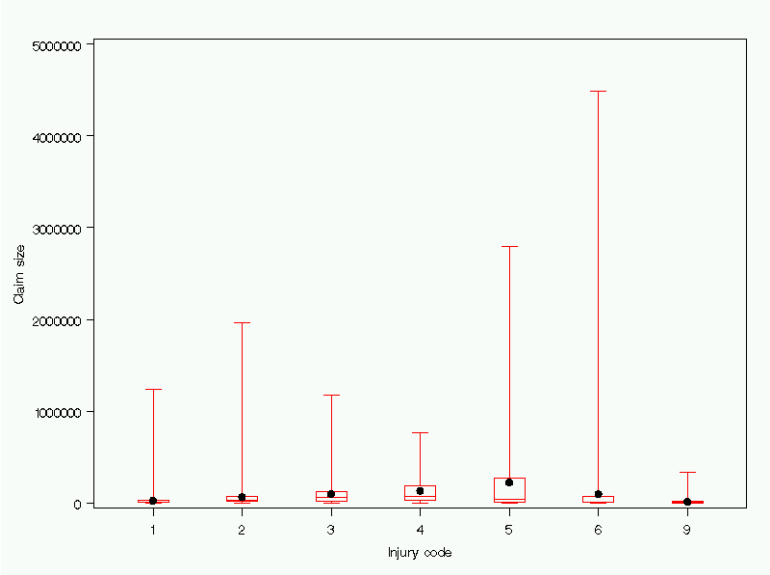

Figure 4: Claim size by injury code

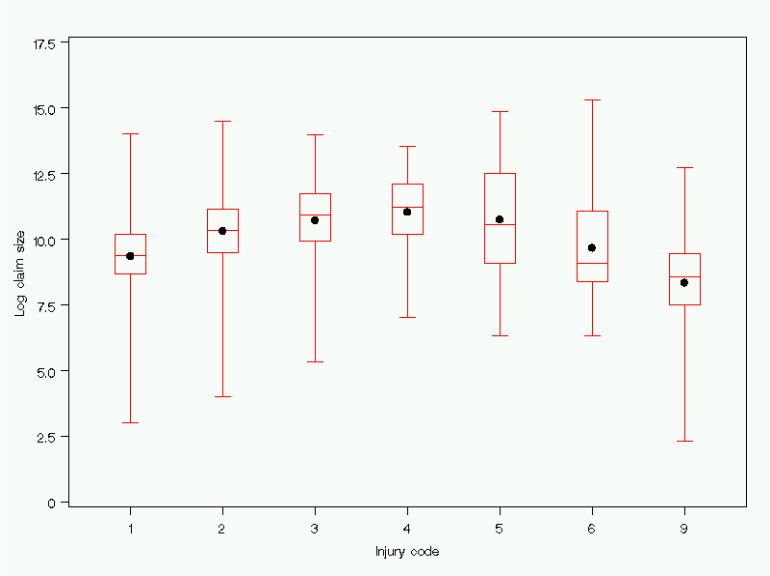

Figure 5: Log claim size by injury code

```
Number of Observations Read       22036
Number of Observations Used       22036
```

Class Level Information

Class	Value	Design Variables					
INJ1	1	0	0	0	0	0	0
	2	1	0	0	0	0	0
	3	0	1	0	0	0	0
	4	0	0	1	0	0	0
	5	0	0	0	1	0	0
	6	0	0	0	0	1	0
	9	0	0	0	0	0	1

Criteria For Assessing Goodness Of Fit

Criterion	DF	Value	Value/DF
Deviance	22E3	36360.7190	1.6506
Scaled Deviance	22E3	26488.0205	1.2024
Pearson Chi-Square	22E3	77772.9280	3.5305
Scaled Pearson X2	22E3	56655.9455	2.5719
Log Likelihood		-249974.0901	

```
Algorithm converged.
```

Analysis Of Parameter Estimates

Parameter		DF	Estimate	Standard Error	Wald 95% Confidence Limits		Chi-Square	Pr > ChiSq
Intercept		1	10.1575	0.0094	10.1391	10.1759	1175361	<.0001
INJ1	2	1	0.9404	0.0222	0.8968	0.9840	1788.68	<.0001
INJ1	3	1	1.3265	0.0360	1.2558	1.3971	1354.11	<.0001
INJ1	4	1	1.6345	0.0857	1.4664	1.8025	363.42	<.0001
INJ1	5	1	2.1624	0.0860	1.9939	2.3308	632.76	<.0001
INJ1	6	1	1.3331	0.0738	1.1884	1.4777	326.06	<.0001
INJ1	9	1	-0.6371	0.0344	-0.7045	-0.5698	343.81	<.0001
Scale		1	0.7285	0.0059	0.7169	0.7402		

```
NOTE: The scale parameter was estimated by maximum likelihood.
```

LR Statistics For Type 3 Analysis

Source	DF	Chi-Square	Pr > ChiSq
INJ1	6	5117.00	<.0001

Anscombe residuals are right–skewed but not severely non–normal. Other diagnostics do not show a lack of fit, so the gamma model appears adequate.

Parameter interpretation

Parameter	Level	$\hat{\beta}$	$e^{\hat{\beta}}$
Intercept		10.158	25783.757
`inj1`	1	0.000	1.000
`inj1`	2	0.940	2.561
`inj1`	3	1.327	3.768
`inj1`	4	1.635	5.127
`inj1`	5	2.162	8.692
`inj1`	6	1.333	3.793
`inj1`	9	-0.637	0.529

- Claimants with injury code 1 have expected claim size of $25 783.76.
- Claimants with injury code 2 have expected claim size 156.1% higher than those with injury code 1;
- Claimants with injury code 3 have expected claim size 276.8% higher than those with injury code 1;
- etc.