Chapter 9: Correlated data

9.1 *Develop a model for the number of claims, in the simulated longitudinal vehicle insurance data set.*

We try the model with driver's age, vehicle value and period.

Random intercept model

The following is the `proc nlmixed` code:

```
proc nlmixed data=act.claimslong;

bounds nu2>0;

zage=beta1age*age1+beta2age*age2+beta4age*age4+beta5age*age5+beta6age*age6;
zvalue=beta2value*value2+beta3value*value3+beta4value*value4+beta5value*value5+beta6value*value6;
zperiod=beta1period*period1+beta2period*period2;
z=alpha+beta0+zage+zvalue+zperiod;

mu=exp(z);

model numclaims ~ poisson(mu);
random alpha ~ normal(0,nu2) subject=policyid;

run;
```

This didn't converge in a reasonable time (overnight).

GEE model

We firstly try the model with unstructured correlation:

```
proc genmod data=act.claimslong;
class  agecat valuecat policyid period;
model numclaims = agecat valuecat period / dist=poisson type3;
repeated subject=policyid / type=un within=period corrw;
run;
```

This results in the estimated working correlation matrix:

Working Correlation Matrix

	Col1	Col2	Col3
Row1	1.0000	0.3425	0.3448
Row2	0.3425	1.0000	0.3618
Row3	0.3448	0.3618	1.0000

which suggests compound symmetric correlation.

```
proc genmod data=act.claimslong;
class  agecat valuecat policyid period;
model numclaims = agecat valuecat period / dist=poisson type3;
repeated subject=policyid / type=cs within=period corrw;
run;
```

The GENMOD Procedure

Model Information

Data Set	ACT.CLAIMSLONG
Distribution	Poisson
Link Function	Log
Dependent Variable	numclaims

Number of Observations Read	120000
Number of Observations Used	120000

Class Level Information

Class	Levels	Values
agecat	6	1 2 4 5 6 10
valuecat	6	2 3 4 5 6 9
policyID	40000	1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50 51 52 53 54 55 56 57 58 59 60 61 62 63 64 65 66 67 68 69 70 71 72 73 74 75 76 77 78 79 80 81 82 83 84 85 86 87 ...
period	3	1 2 3

Parameter Information

Parameter	Effect	agecat	valuecat	period
Prm1	Intercept			
Prm2	agecat	1		
Prm3	agecat	2		
Prm4	agecat	4		
Prm5	agecat	5		
Prm6	agecat	6		
Prm7	agecat	10		
Prm8	valuecat		2	
Prm9	valuecat		3	
Prm10	valuecat		4	
Prm11	valuecat		5	
Prm12	valuecat		6	
Prm13	valuecat		9	
Prm14	period			1
Prm15	period			2
Prm16	period			3

Criteria For Assessing Goodness Of Fit

Criterion	DF	Value	Value/DF
Deviance	12E4	67203.6757	0.5601
Scaled Deviance	12E4	67203.6757	0.5601
Pearson Chi-Square	12E4	105409.5109	0.8785
Scaled Pearson X2	12E4	105409.5109	0.8785
Log Likelihood		-50648.7933	

Algorithm converged.

Analysis Of Initial Parameter Estimates

Parameter		DF	Estimate	Standard Error	Wald 95% Confidence Limits		Chi-Square	Pr > ChiSq
Intercept		1	-1.8446	0.0192	-1.8822	-1.8070	9256.67	<.0001
agecat	1	1	0.1627	0.0283	0.1072	0.2182	33.03	<.0001
agecat	2	1	0.0117	0.0231	-0.0335	0.0569	0.26	0.6126
agecat	4	1	-0.0310	0.0222	-0.0745	0.0125	1.95	0.1622
agecat	5	1	-0.1685	0.0259	-0.2193	-0.1177	42.27	<.0001
agecat	6	1	-0.1289	0.0303	-0.1882	-0.0695	18.11	<.0001

agecat	10	0	0.0000	0.0000	0.0000	0.0000	.	.
valuecat	2	1	0.1385	0.0195	0.1003	0.1767	50.49	<.0001
valuecat	3	1	0.0551	0.0564	-0.0555	0.1657	0.95	0.3287
valuecat	4	1	-0.5800	0.2674	-1.1041	-0.0558	4.70	0.0301
valuecat	5	1	-0.2283	0.3537	-0.9215	0.4649	0.42	0.5186
valuecat	6	1	-0.9304	0.5001	-1.9106	0.0497	3.46	0.0628
valuecat	9	0	0.0000	0.0000	0.0000	0.0000	.	.
period	1	1	-0.1796	0.0187	-0.2161	-0.1430	92.70	<.0001
period	2	1	-0.1019	0.0183	-0.1377	-0.0661	31.14	<.0001
period	3	0	0.0000	0.0000	0.0000	0.0000	.	.
Scale		0	1.0000	0.0000	1.0000	1.0000		

NOTE: The scale parameter was held fixed.

GEE Model Information

Correlation Structure	Exchangeable
Within-Subject Effect	period (3 levels)
Subject Effect	policyID (40000 levels)
Number of Clusters	40000
Correlation Matrix Dimension	3
Maximum Cluster Size	3
Minimum Cluster Size	3

Algorithm converged.

Working Correlation Matrix

	Col1	Col2	Col3
Row1	1.0000	0.3496	0.3496
Row2	0.3496	1.0000	0.3496
Row3	0.3496	0.3496	1.0000

Exchangeable Working Correlation

Correlation	0.3496154214

Analysis Of GEE Parameter Estimates
Empirical Standard Error Estimates

Parameter		Estimate	Standard Error	95% Confidence Limits		Z	Pr > |Z|
Intercept		-1.8447	0.0213	-1.8863	-1.8030	-86.71	<.0001
agecat	1	0.1640	0.0348	0.0957	0.2322	4.71	<.0001
agecat	2	0.0104	0.0283	-0.0451	0.0659	0.37	0.7143
agecat	4	-0.0314	0.0270	-0.0843	0.0215	-1.16	0.2452
agecat	5	-0.1692	0.0314	-0.2307	-0.1077	-5.39	<.0001
agecat	6	-0.1276	0.0371	-0.2003	-0.0550	-3.44	0.0006
agecat	10	0.0000	0.0000	0.0000	0.0000	.	.
valuecat	2	0.1399	0.0239	0.0931	0.1867	5.86	<.0001
valuecat	3	0.0540	0.0689	-0.0810	0.1890	0.78	0.4329
valuecat	4	-0.5849	0.2887	-1.1507	-0.0191	-2.03	0.0427
valuecat	5	-0.2135	0.5134	-1.2198	0.7927	-0.42	0.6775
valuecat	6	-0.8814	0.5784	-2.0150	0.2522	-1.52	0.1275
valuecat	9	0.0000	0.0000	0.0000	0.0000	.	.
period	1	-0.1796	0.0142	-0.2073	-0.1518	-12.69	<.0001
period	2	-0.1019	0.0136	-0.1286	-0.0753	-7.49	<.0001
period	3	0.0000	0.0000	0.0000	0.0000	.	.

Score Statistics For Type 3 GEE Analysis

Source	DF	Chi-Square	Pr > ChiSq
agecat	5	88.09	<.0001

```
valuecat          5      44.24        <.0001
period            2     162.06        <.0001
```

- All explanatory variables are highly significant.
- Effects of explanatory variables:

Variable	Level	$\hat{\beta}$	$e^{\hat{\beta}}$
Intercept		-1.845	0.158
Driver's age	1	0.164	1.178
	2	0.010	1.010
	3	0.000	1.000
	4	-0.031	0.969
	5	-0.169	0.844
	6	-0.128	0.880
Vehicle value	1	0.000	1.000
	2	0.140	1.150
	3	0.054	1.055
	4	-0.585	0.557
	5	-0.214	0.808
	6	-0.881	0.414
Time period	1	-0.180	0.836
	2	-0.102	0.903
	3	0.000	1.000

  - The expected number of claims for a policy with driver's age level 3, vehicle value level 1, in period 3, is 0.158.
  - The effect on the expected number of claims of driver's age being level 1, compared with level 3, is an increase of 17.8%; etc.
  - The effect on the expected number of claims of vehicle value being level 3, compared with level 1, is an increase of 5.5%; etc.
  - The expected number of claims in period 1 is 83.6% of those in period 3; the expected number of claims in period 2 is 90.3% of those in period 3.