CHAPTER 12

A note on computing with SAS (Version 9):

The **SAS GENMOD** procedure is used for fitting the Poisson regression models of Chapter 12. This procedure is very general. It can also be used for the logistic regression models in Chapter 11, as well as most generalized linear models.

SAS works slightly different than the previously considered spreadsheet programs Minitab, SPSS, or EXCEL. In SAS one needs to write out a line code. The line code gets entered into a program editor, and is executed by clicking the SAS "run" and "submit" tabs. Here we list an example of the line code, with a detailed discussion of important options. Many more options are available, and they can be reviewed by looking at the on-line help pages within SAS.

We list the input for Exercise 12.1:

```
data exer12n1;
```
specifies the file name for data set
```
input type year period ms nudamage;
```
specifies the input variables
```
lnms=log(ms);
```
specifies a transformation; here the natural log transformation
```
datalines;
1     1     1     127    0
1     1     2     63     0
1     2     1     1095   3
1     2     2     1095   4
1     3     1     1512   6
1     3     2     3353   18
1     4     2     2244   11
2     1     1     44882  39
2     1     2     17176  29
2     2     1     28609  58
2     2     2     20370  53
2     3     1     7064   12
2     3     2     13099  44
2     4     2     7117   18
3     1     1     1179   1
3     1     2     552    1
3     2     1     781    0
3     2     2     676    1
3     3     1     783    6
3     3     2     1948   2
3     4     2     274    1
```

```
4     1     1     251    0
4     1     2     105    0
4     2     1     288    0
4     2     2     192    0
4     3     1     349    2
4     3     2     1208   11
4     4     2     2051   4
5     1     1     45     0
5     2     1     789    7
5     2     2     437    7
5     3     1     1157   5
5     3     2     2161   12
5     4     2     542    1
;
proc genmod data=exer12n1;
```

PROC GENMOD is called

```
class type / param=ref ref=first;
class year / param=ref ref=first;
class period / param=ref ref=first;
```

specifies that type, year, and period are class (factor) variables; SAS creates the appropriate indicator variables automatically. The first numeric value is taken as the base for comparisons.

```
model nudamage=type year period lnms / d=poisson obstats
covb corrb lrci type3;
```

Here the model gets specified. The response is nudamage. The first three variables on the right hand side of the equal sign are factors. The last variable (lnms) is a covariate (not a factor). Options are listed after the slash.

d=Poisson: Poisson link.

Covb, Corrb: Covariance and correlation matrices of the parameter estimates are displayed.

Obstats: results in detailed output (fitted values, residuals, …)

Lrci requests that two-sided confidence intervals for all model parameters are computed based on the profile likelihood function. This is sometimes called the partially maximized likelihood function. Two-sided Wald confidence intervals are calculated, if lrci is not specified.

Likelihood ratio-based confidence intervals, also known as profile likelihood confidence intervals, of parameter estimates in generalized linear models can be explained as follows. Suppose that the parameter vector is $\boldsymbol{\beta} = (\beta_0, \beta_1, ..., \beta_p)'$ and one wants a confidence interval for β_i.The profile likelihood function for β_i is defined as $l^*(\beta_i) = \max_{\tilde{\boldsymbol{\beta}}} l(\boldsymbol{\beta})$, where $\tilde{\boldsymbol{\beta}}$ is the vector $\boldsymbol{\beta}$ with the ith element fixed at β_i and $l = l(\boldsymbol{\beta})$ is the log likelihood function. Let $l = l(\hat{\boldsymbol{\beta}})$ be the log likelihood evaluated at the maximum likelihood estimate $\hat{\boldsymbol{\beta}}$. Under the assumption that β_i is the true parameter value, $2(l - l^*(\beta_i))$ has a limiting chi-square distribution with one degree of freedom. A $100(1-\alpha)$ percent confidence interval for β_i is

$$\{\beta_i : l^*(\beta_i) \geq l - 0.5\chi^2(1-\alpha;1)\}$$

where $\chi^2(1-\alpha;1)$ is the $100(1-\alpha)$ percentile of the chi-square distribution with one degree of freedom. The endpoints of the confidence interval can be found by solving numerically for values of β_i that satisfy the equality in the preceding relation.

Type 3: requests that statistics for Type 3 contrasts be computed for each class variable (factor) specified in the MODEL statement. This means that likelihood-ratio tests are calculated for the contrasts of the class variables.Type 3 means that these are partial tests, comparing the full model with the restricted model that lacks the indicated class variable (factor).

OFFSET = lnms: specifies a variable in the input data set (here lnms) to be used as an offset variable. This variable cannot be a CLASS variable. In our example it seems reasonable to suppose that the number of damage incidents is directly proportional to MS, the months of service, and one can expect that the coefficient in the Poisson regression model that corresponds to ln(MS) is one. OFFSET = lnms restricts this parameter to one.

Scale = deviance: Overdispersion is a phenomenon that sometimes occurs in data that are modeled with the Poisson (and also binomial - see Chapter 11) distributions. If the estimate of dispersion after fitting, as measured by the deviance or Pearson's chi-square divided by the degrees of freedom, is not near 1, then the data may be overdispersed if the dispersion estimate is greater than 1, or underdispersed if the dispersion estimate is less than 1. A simple way to model this situation is to allow the variance function of the Poisson distribution to have a multplicative overdispersion factor, $\mathrm{Var}(\mu) = \phi\mu$ (or $\mathrm{Var}(\mu) = \phi\mu(1-\mu)$ for the binomial link).

The models are fit in the usual way. The parameter estimates are not affected by the value of ϕ. The covariance matrix, however, is multiplied by ϕ, and the scaled deviance and log likelihoods used in likelihood ratio tests are divided by ϕ.

The SCALE= option in the MODEL statement enables you to specify a value of ϕ for the Poisson (and also binomial) distributions. If you specify the SCALE=DEVIANCE option in the MODEL statement, the procedure uses the deviance divided by the degrees of freedom as an estimate of ϕ, and all statistics are adjusted appropriately. You can use Pearson's chi-square instead of the deviance by specifying the SCALE=PEARSON option.

```
run;
```

Executes the program

Many other options are available. See the SAS on-line help for further discussion and examples.

12.1

(a) We use SAS GENMOD to estimate the Poisson regression model with link

$$\ln\mu = \beta_0 + \beta_1 \ln(\mathrm{MS}) + \beta_2 \mathrm{X2} + ... + \beta_5 \mathrm{X5} + \beta_6 \mathrm{Z2} + ... + \beta_8 \mathrm{Z4} + \beta_9 \mathrm{W2}$$

Here X1 through X5 are the indicator variables for the type of ship (a class variable with five possibilities), Z1 through Z4 are the indicator variables for the year of construction (a class variable with four possibilities), and W1 and W2 are the indicator variables for the period of operation (a class variable with two possibilities). SAS GENMOD creates the associated indicator variables for the specified class variables automatically. The first outcome is declared as the reference.

The (type 3) test statistics at the end of the program output test the significance of the class variables. For example, the test statistic for "type" is obtained by comparing the log-likelihood of the full model (768.4585) with the log-likelihood of the restricted model that is missing that factor (the model with year, period, and ln(MS)). The log-likelihood of the restricted model is 762.1757. Hence the log-likelihood statistic is 2(768.4582 - 762.1757) = 12.57. Comparing this value to a chi-square with 4 degrees of freedom (since there are 4 restrictions), leads to the probability value

$P(\chi^2(4) \geq 12.57) = 0.0136$. These are the values given at the end of the output. The tests for the other factors can be obtained similarly. They indicate that one can not simplify the model. All three factors are needed to explain the number of damage claims.

Ships of type 3 report the smallest number of damage incidents. Ships constructed in years 2 (1965-1969) and 3 (1970-1974) experience the highest number of reported damage incidents. The second period of operation (1975-79) is associated with a higher number of reported damage incidents.

Fitting results for the full model:

```
                         The GENMOD Procedure

                          Model Information

                 Data Set              WORK.EXER12N1
                 Distribution                Poisson
                 Link Function                   Log
                 Dependent Variable         nudamage
                 Observations Used                34
                       Class Level Information

             Class      Value         Design Variables

             type       1          0      0      0      0
                        2          1      0      0      0
                        3          0      1      0      0
                        4          0      0      1      0
                        5          0      0      0      1

             year       1          0      0      0
                        2          1      0      0
                        3          0      1      0
                        4          0      0      1

             period     1          0
                        2          1

                        Parameter Information

         Parameter       Effect       type    year    period

         Prm1            Intercept
         Prm2            lnms
         Prm3            type         2
         Prm4            type         3
         Prm5            type         4
         Prm6            type         5
         Prm7            year                 2
         Prm8            year                 3
         Prm9            year                 4
         Prm10           period                       2
```

Criteria For Assessing Goodness Of Fit

Criterion	DF	Value	Value/DF
Deviance	24	37.8043	1.5752
Scaled Deviance	24	37.8043	1.5752
Pearson Chi-Square	24	39.4494	1.6437
Scaled Pearson X2	24	39.4494	1.6437
Log Likelihood		768.4585	

Algorithm converged.

Estimated Correlation Matrix

	Prm1	Prm2	Prm3	Prm4	Prm5	Prm6	Prm7	Prm8	Prm9	Prm10
Prm1	1.0000	-0.9688	0.6048	-0.3172	-0.3046	-0.3304	-0.3405	-0.4538	-0.4298	-0.1729
Prm2	-0.9688	1.0000	-0.7587	0.2328	0.2200	0.2234	0.2291	0.3364	0.3495	0.1216
Prm3	0.6048	-0.7587	1.0000	0.0990	0.1226	0.1958	-0.1165	-0.0967	-0.1341	-0.0768
Prm4	-0.3172	0.2328	0.0990	1.0000	0.2798	0.3483	0.0899	0.1225	0.1660	0.0258
Prm5	-0.3046	0.2200	0.1226	0.2798	1.0000	0.3706	0.0788	0.1001	0.0024	0.0225
Prm6	-0.3304	0.2234	0.1958	0.3483	0.3706	1.0000	0.0466	0.0428	0.1200	0.0522
Prm7	-0.3405	0.2291	-0.1165	0.0899	0.0788	0.0466	1.0000	0.6612	0.5146	-0.0770
Prm8	-0.4538	0.3364	-0.0967	0.1225	0.1001	0.0428	0.6612	1.0000	0.5938	-0.1854
Prm9	-0.4298	0.3495	-0.1341	0.1660	0.0024	0.1200	0.5146	0.5938	1.0000	-0.2444
Prm10	-0.1729	0.1216	-0.0768	0.0258	0.0225	0.0522	-0.0770	-0.1854	-0.2444	1.0000

Analysis Of Parameter Estimates

Parameter		DF	Estimate	Standard Error	Wald 95% Confidence Limits		Chi-Square	Pr > ChiSq
Intercept		1	-5.5940	0.8724	-7.3038	-3.8841	41.12	<.0001
lnms		1	0.9027	0.1018	0.7032	1.1022	78.63	<.0001
type	2	1	-0.3499	0.2702	-0.8795	0.1797	1.68	0.1954
type	3	1	-0.7631	0.3382	-1.4259	-0.1003	5.09	0.0240
type	4	1	-0.1355	0.2971	-0.7178	0.4469	0.21	0.6484
type	5	1	0.2739	0.2418	-0.1999	0.7478	1.28	0.2572
year	2	1	0.6625	0.1536	0.3614	0.9637	18.60	<.0001
year	3	1	0.7597	0.1777	0.4115	1.1079	18.29	<.0001
year	4	1	0.3697	0.2458	-0.1121	0.8516	2.26	0.1326
period	2	1	0.3703	0.1181	0.1387	0.6018	9.82	0.0017
Scale		0	1.0000	0.0000	1.0000	1.0000		

LR Statistics For Type 3 Analysis

Source	DF	Chi-Square	Pr > ChiSq
lnms	1	101.28	<.0001
type	4	12.57	0.0136
year	3	27.20	<.0001
period	1	9.97	0.0016

Fitting results for the restricted model without type of ship:

The GENMOD Procedure

Model Information

Data Set	WORK.EXER12N1
Distribution	Poisson
Link Function	Log
Dependent Variable	nudamage
Observations Used	34

Class Level Information

Class	Value	Design Variables		
year	1	0	0	0
	2	1	0	0
	3	0	1	0
	4	0	0	1
period	1	0		
	2	1		

Criteria For Assessing Goodness Of Fit

Criterion	DF	Value	Value/DF
Deviance	28	50.3699	1.7989
Scaled Deviance	28	50.3699	1.7989
Pearson Chi-Square	28	46.7116	1.6683
Scaled Pearson X2	28	46.7116	1.6683
Log Likelihood		762.1757	

Algorithm converged.

Analysis Of Parameter Estimates

Parameter		DF	Estimate	Standard Error	Wald 95% Confidence Limits		Chi-Square	Pr > ChiSq
Intercept		1	-5.2229	0.4826	-6.1688	-4.2771	117.12	<.0001
lnms		1	0.8311	0.0460	0.7409	0.9213	326.13	<.0001
year	2	1	0.6735	0.1503	0.3790	0.9681	20.08	<.0001
year	3	1	0.7967	0.1702	0.4631	1.1303	21.91	<.0001
year	4	1	0.3978	0.2337	-0.0603	0.8560	2.90	0.0887
period	2	1	0.3546	0.1168	0.1256	0.5837	9.21	0.0024
Scale		0	1.0000	0.0000	1.0000	1.0000		

NOTE: The scale parameter was held fixed.

(b) It seems reasonable to suppose that the number of damage incidents is directly proportional to MS, the months of service, and one can expect that the coefficient β_1 is one. The literature refers to the term $\ln(\mathrm{MS})$ as an "offset." Let us test for the offset, and test whether $\beta_1 = 1$. The estimate is $\hat{\beta}_1 = 0.9027$, and the 95 percent Wald confidence interval is given by $0.9027 \pm (1.96)(0.1018)$, 0.90 ± 0.20, or $0.70 \le \beta_1 \le 1.10$. The interval includes one, which makes the off-set interpretation plausible.

(c) We assume an "offset" for aggregate months of service (that is, we impose the restriction $\beta_1 = 1$) and estimate the model with link

$$\ln \mu = \beta_0 + \ln(\mathrm{MS}) + \beta_2 \mathrm{X2} + ... + \beta_5 \mathrm{X5} + \beta_6 \mathrm{Z2} + ... + \beta_8 \mathrm{Z4} + \beta_9 \mathrm{W2}$$

The results of the estimation are similar to the ones of the full model in (a).

Fitting results for the model with an offset:

The GENMOD Procedure

Model Information

Data Set	WORK.EXER12N1
Distribution	Poisson
Link Function	Log
Dependent Variable	nudamage
Offset Variable	lnms
Observations Used	34

Class Level Information

Class	Value	Design Variables			
type	1	0	0	0	0
	2	1	0	0	0
	3	0	1	0	0
	4	0	0	1	0
	5	0	0	0	1
year	1	0	0	0	
	2	1	0	0	
	3	0	1	0	
	4	0	0	1	
period	1	0			
	2	1			

Parameter Information

Parameter	Effect	type	year	period
Prm1	Intercept			
Prm2	type	2		
Prm3	type	3		
Prm4	type	4		
Prm5	type	5		
Prm6	year		2	
Prm7	year		3	
Prm8	year		4	
Prm9	period			2

Criteria For Assessing Goodness Of Fit

Criterion	DF	Value	Value/DF
Deviance	25	38.6951	1.5478
Scaled Deviance	25	38.6951	1.5478
Pearson Chi-Square	25	42.2753	1.6910
Scaled Pearson X2	25	42.2753	1.6910
Log Likelihood		768.0131	

Algorithm converged.

Estimated Correlation Matrix

	Prm1	Prm2	Prm3	Prm4	Prm5	Prm6	Prm7	Prm8	Prm9
Prm1	1.0000	-0.8114	-0.3784	-0.3706	-0.4699	-0.4843	-0.5501	-0.4015	-0.2161
Prm2	-0.8114	1.0000	0.4332	0.4468	0.5707	0.0856	0.2714	0.2285	0.0254
Prm3	-0.3784	0.4332	1.0000	0.2375	0.3136	0.0358	0.0455	0.0971	-0.0031
Prm4	-0.3706	0.4468	0.2375	1.0000	0.3338	0.0277	0.0286	-0.0966	-0.0047
Prm5	-0.4699	0.5707	0.3136	0.3338	1.0000	-0.0041	-0.0371	0.0528	0.0269

Prm6	-0.4843	0.0856	0.0358	0.0277	-0.0041	1.0000	0.6335	0.4755	-0.1201
Prm7	-0.5501	0.2714	0.0455	0.0286	-0.0371	0.6335	1.0000	0.5482	-0.2636
Prm8	-0.4015	0.2285	0.0971	-0.0966	0.0528	0.4755	0.5482	1.0000	-0.3154
Prm9	-0.2161	0.0254	-0.0031	-0.0047	0.0269	-0.1201	-0.2636	-0.3154	1.0000

Analysis Of Parameter Estimates

Parameter		DF	Estimate	Standard Error	Wald 95% Confidence Limits		Chi-Square	Pr > ChiSq
Intercept		1	-6.4059	0.2174	-6.8321	-5.9797	867.89	<.0001
type	2	1	-0.5433	0.1776	-0.8914	-0.1953	9.36	0.0022
type	3	1	-0.6874	0.3290	-1.3323	-0.0425	4.36	0.0367
type	4	1	-0.0760	0.2906	-0.6455	0.4936	0.07	0.7938
type	5	1	0.3256	0.2359	-0.1367	0.7879	1.91	0.1675
year	2	1	0.6971	0.1496	0.4038	0.9904	21.70	<.0001
year	3	1	0.8184	0.1698	0.4857	1.1512	23.24	<.0001
year	4	1	0.4534	0.2332	-0.0036	0.9104	3.78	0.0518
period	2	1	0.3845	0.1183	0.1527	0.6163	10.57	0.0012
Scale		0	1.0000	0.0000	1.0000	1.0000		

NOTE: The scale parameter was held fixed.

LR Statistics For Type 3 Analysis

Source	DF	Chi-Square	Pr > ChiSq
type	4	23.67	<.0001
year	3	31.41	<.0001
period	1	10.66	0.0011

(d) Let us look at the deviance goodness-of-fit statistics. Comparing the deviance D = 37.8043 to a chi-square with 24 degrees of freedom, leads to the probability value $P(\chi^2(24) \geq 37.80) = 1 - 0.9637 = 0.0363$. The deviance exceeds the 95th percentile and the probability value is slightly smaller than 0.05. This is a sign of overdispersion. We adjust the analysis for overdispersion by allowing the variance function of the Poisson distribution to have a multplicative overdispersion factor, $\text{Var}(\mu) = \phi\mu$. The model is fit in the usual way, and the parameter estimates are not affected by the value of ϕ. The covariance matrix, however, is multiplied by ϕ, and the scaled deviance and log likelihoods used in likelihood ratio tests are divided by ϕ. The SCALE=DEVIANCE option in the MODEL statement enables us to specify a value of ϕ for the Poisson distribution. The procedure uses the deviance divided by the degrees of freedom as an estimate of ϕ, and all statistics are adjusted appropriately.

The results are basically unchanged. The test statistics indicate that all three factors are statistically significant. Ships of types 2 and 3 experience the smallest numbers of reported damage incidents. Ships constructed in years 2 (1965-1969) and 3 (1970-1974) experience the largest numbers of reported damage incidents. The second period of operation (1975-79) is associated with a higher number of reported damage incidents.

Fitting results for the model with scale adjustment:

The GENMOD Procedure

Model Information

Data Set	WORK.EXER12N1
Distribution	Poisson
Link Function	Log
Dependent Variable	nudamage
Offset Variable	lnms
Observations Used	34

Class Level Information

Class	Value	Design Variables			
type	1	0	0	0	0
	2	1	0	0	0
	3	0	1	0	0
	4	0	0	1	0
	5	0	0	0	1
year	1	0	0	0	
	2	1	0	0	
	3	0	1	0	
	4	0	0	1	
period	1	0			
	2	1			

Parameter Information

Parameter	Effect	type	year	period
Prm1	Intercept			
Prm2	type	2		
Prm3	type	3		
Prm4	type	4		
Prm5	type	5		
Prm6	year		2	
Prm7	year		3	
Prm8	year		4	
Prm9	period			2

Criteria For Assessing Goodness Of Fit

Criterion	DF	Value	Value/DF
Deviance	25	38.6951	1.5478
Scaled Deviance	25	25.0000	1.0000
Pearson Chi-Square	25	42.2753	1.6910
Scaled Pearson X2	25	27.3131	1.0925
Log Likelihood		496.1960	

Algorithm converged.

Estimated Correlation Matrix

	Prm1	Prm2	Prm3	Prm4	Prm5	Prm6	Prm7	Prm8	Prm9
Prm1	1.0000	-0.8114	-0.3784	-0.3706	-0.4699	-0.4843	-0.5501	-0.4015	-0.2161
Prm2	-0.8114	1.0000	0.4332	0.4468	0.5707	0.0856	0.2714	0.2285	0.0254
Prm3	-0.3784	0.4332	1.0000	0.2375	0.3136	0.0358	0.0455	0.0971	-0.0031
Prm4	-0.3706	0.4468	0.2375	1.0000	0.3338	0.0277	0.0286	-0.0966	-0.0047
Prm5	-0.4699	0.5707	0.3136	0.3338	1.0000	-0.0041	-0.0371	0.0528	0.0269

Prm6	-0.4843	0.0856	0.0358	0.0277	-0.0041	1.0000	0.6335	0.4755	-0.1201
Prm7	-0.5501	0.2714	0.0455	0.0286	-0.0371	0.6335	1.0000	0.5482	-0.2636
Prm8	-0.4015	0.2285	0.0971	-0.0966	0.0528	0.4755	0.5482	1.0000	-0.3154
Prm9	-0.2161	0.0254	-0.0031	-0.0047	0.0269	-0.1201	-0.2636	-0.3154	1.0000

Analysis Of Parameter Estimates

Parameter		DF	Estimate	Standard Error	Wald 95% Confidence Limits		Chi-Square	Pr > ChiSq
Intercept		1	-6.4059	0.2705	-6.9361	-5.8757	560.72	<.0001
type	2	1	-0.5433	0.2209	-0.9764	-0.1103	6.05	0.0139
type	3	1	-0.6874	0.4094	-1.4898	0.1149	2.82	0.0931
type	4	1	-0.0760	0.3615	-0.7845	0.6326	0.04	0.8336
type	5	1	0.3256	0.2935	-0.2496	0.9007	1.23	0.2672
year	2	1	0.6971	0.1862	0.3323	1.0620	14.02	0.0002
year	3	1	0.8184	0.2112	0.4044	1.2324	15.01	0.0001
year	4	1	0.4534	0.2901	-0.1151	1.0220	2.44	0.1180
period	2	1	0.3845	0.1471	0.0961	0.6729	6.83	0.0090
Scale		0	1.2441	0.0000	1.2441	1.2441		

NOTE: The scale parameter was estimated by the square root of DEVIANCE/DOF.

LR Statistics For Type 3 Analysis

Source	Num DF	Den DF	F Value	Pr > F	Chi-Square	Pr > ChiSq
type	4	25	3.82	0.0147	15.29	0.0041
year	3	25	6.76	0.0017	20.29	0.0001
period	1	25	6.89	0.0146	6.89	0.0087

(e) A model with every possible two-factor interaction contains
1 (const) + 4 + 3 + 1 (main effects) + 4*3 + 4*1 + 3*1 (2-factor interactions) = 28 parameters. This is a highly non-parsimonious model, considering that there are only 34 observations. The number of parameters in the fully saturated model (with the 3-factor interaction added) exceeds the number of observations.

Here we enter each two-factor interaction one at-a-time. The type 3 test results for the models with the type by period interaction (4 additional parameters) and the year by period interaction (3 additional parameters) are given below. The model with the type by year interaction (12 additional parameters) experienced convergence problems, probably due to the large number of additional parameters and the sparseness of the data. The results indicate that interaction components are not needed. Note that type 3 LR test statistics are partial tests, always testing whether the factor in question is significant when added last to the model. The period effect is insignificant when adding it to the model with type, year, and the type by period interaction. However, it becomes significant when the type by period interaction is omitted.

<u>Fitting results for the model with interaction:</u>

LR Statistics For Type 3 Analysis

Source	DF	Chi-Square	Pr > ChiSq

type	4	12.13	0.0164
year	3	30.70	<.0001
period	1	1.57	0.2105
type*period	4	4.94	0.2936

LR Statistics For Type 3 Analysis

Source	DF	Chi-Square	Pr > ChiSq
type	4	23.71	<.0001
year	3	25.26	<.0001
period	1	7.29	0.0069
year*period	3	4.00	0.2613

(f) See parts (a) – (e)

12.2 PROC GENMOD is used to estimate the Poisson regression model with link

$$\ln\mu = \beta_0 + \alpha \ln(\mathrm{H}) + \beta_1 \mathrm{A}_2 + \beta_2 \mathrm{T}_2 + \beta_3 \mathrm{T}_3$$

where *H* is the number of policies and A_1, A_2 and $\mathrm{T}_1, \mathrm{T}_2, \mathrm{T}_3$ are the corresponding indicator variables for the two age groups and three car types.

The type 3 test statistics at the end of the program output are tests of the significance of the class variables. For example, the test statistic for "age" is obtained by comparing the log-likelihood of the full model (838.1594) with the log-likelihood of the restricted model (the model with type and ln(H); log-likelihood is 817.8596). The log-likelihood statistic is 40.60. Comparing this values to a chi-square with 1 degree of freedom (since there is only restrictions), leads to the probability value $\mathrm{P}(\chi^2(1) \geq 40.60) = 0.0000$.

The type 3 test statistics indicate that both age and type are highly significant. Both factors are needed to explain the number of claims. Looking at the individual parameter estimates, we see that the second age group experiences more claims than the first. The second and third car type experience fewer claims than the first, and the third car type experiences fewer claims than the second.

It seems reasonable to suppose that the number of claims is directly proportional to the number of policies, and that one can expect the coefficient β_1 to be one. Let us test whether $\beta_1 = 1$. The estimate is $\hat{\beta}_1 = 0.6189$, and the 95 percent Wald confidence interval is given by $0.6189 \pm (1.96)(0.3113)$, 0.62 ± 0.61, or $0.01 \leq \beta_1 \leq 1.23$. The interval is quite wide because there are only very few (six) observations. However, it includes one, which makes the off-set interpretation plausible.

Note that this run also asked for an additional table of statistics to be displayed. For each observation, the following items are displayed: the value of the response variable y_i, the values of the regressor variables, the predicted mean $\hat{\mu}_i = \exp(\boldsymbol{x}_i'\hat{\boldsymbol{\beta}})$, the standard error in the linear predictor $\boldsymbol{x}_i'\hat{\boldsymbol{\beta}}$, the value of the Hessian weight at the final iteration (diagonal elements of the matrix in equation (12.12)), lower and upper confidence limits of the predicted value of the mean (see equation (12.19), the raw residual, the Pearson residual (equation (12.23)), the standardized Pearson residual, the deviance residual (equation (12.22)), the standardized deviance residual, and the likelihood residual. Most of these statistics are explained in Chapter 12.

Fitting results for the full model:

```
                 The GENMOD Procedure

                  Model Information

          Data Set                WORK.EXER12N2
          Distribution                  Poisson
          Link Function                     Log
          Dependent Variable           nuclaims
          Observations Used                   6

               Class Level Information

                                   Design
             Class     Value     Variables

             age       1          0
                       2          1

             car       1          0      0
                       2          1      0
                       3          0      1

                Parameter Information

        Parameter       Effect       age    car

        Prm1            Intercept
        Prm2            lnnupol
        Prm3            age          2
        Prm4            car                 2
        Prm5            car                 3

         Criteria For Assessing Goodness Of Fit

Criterion                DF           Value        Value/DF

Deviance                  1          1.4084          1.4084
Scaled Deviance           1          1.4084          1.4084
Pearson Chi-Square        1          1.2742          1.2742
Scaled Pearson X2         1          1.2742          1.2742
Log Likelihood                     838.1594

Algorithm converged.
```

Estimated Correlation Matrix

	Prm1	Prm2	Prm3	Prm4	Prm5
Prm1	1.0000	-0.9979	-0.7731	0.7040	-0.4578
Prm2	-0.9979	1.0000	0.7416	-0.7275	0.4500
Prm3	-0.7731	0.7416	1.0000	-0.4975	0.2953
Prm4	0.7040	-0.7275	-0.4975	1.0000	-0.2073
Prm5	-0.4578	0.4500	0.2953	-0.2073	1.0000

Analysis Of Parameter Estimates

Parameter		DF	Estimate	Standard Error	Wald 95% Confidence Limits		Chi-Square	Pr > ChiSq
Intercept		1	-0.1920	1.9964	-4.1048	3.7208	0.01	0.9234
lnnupol		1	0.6189	0.3113	0.0089	1.2290	3.95	0.0468
age	2	1	1.1313	0.2005	0.7383	1.5244	31.83	<.0001
car	2	1	-0.5266	0.1856	-0.8904	-0.1628	8.05	0.0046
car	3	1	-1.9130	0.3045	-2.5099	-1.3161	39.46	<.0001
Scale		0	1.0000	0.0000	1.0000	1.0000		

NOTE: The scale parameter was held fixed.

LR Statistics For Type 3 Analysis

Source	DF	Chi-Square	Pr > ChiSq
lnnupol	1	4.34	0.0372
age	1	40.60	<.0001
car	2	74.23	<.0001

Observation Statistics (need to read across)

Observation	nuclaims	lnnupol / HessWgt / Resdev	age	car	Pred / Upper / StReschi	Xbeta / Resraw / Reslik	Std / Reschi
			Lower / StResdev				
1	42	6.2146081	1	1	38.649518	3.6545343	0.1413353
		38.649518	29.298086		50.985762	3.350482	0.5389336
		0.5314155	1.1130493		1.128796	1.125226	
2	37	7.0900768	1	2	39.243343	3.6697818	0.1513849
		39.243343	29.168017		52.798927	-2.243343	-0.358107
		-0.361603	-1.139821		-1.128802	-1.129916	
3	1	4.6051702	1	3	2.1071828	0.7453519	0.5078463
		2.1071828	0.7787941		5.7014036	-1.107183	-0.762725
		-0.850683	-1.259006		-1.12883	-1.190029	
4	101	5.9914645	2	1	104.3505	4.6477554	0.0936696
		104.3505	86.848591		125.37942	-3.350497	-0.327991
		-0.32977	-1.134924		-1.128801	-1.129319	
5	73	6.2146081	2	2	70.756662	4.2592467	0.1155164
		70.756662	56.420843		88.73503	2.2433384	0.2666927
		0.2653017	1.1229119		1.1287992	1.1284714	
6	14	5.7037825	2	3	12.89285	2.5566729	0.2679085
		12.89285	7.6261405		21.796816	1.10715	0.3083415
		0.3040793	1.1131926		1.1287961	1.1276393	

Next, we assume an "offset" for the number of policies (that is, we impose the restriction $\beta_1 = 1$) and estimate the model with link

$$\ln \mu = \beta_0 + \ln(\mathrm{H}) + \beta_1 \mathrm{A}_2 + \beta_2 \mathrm{T}_2 + \beta_3 \mathrm{T}_3 .$$

The results are given below. The interpretation of the earlier model is largely unchanged. Both age and type are highly significant. The second age group experiences more claims than the first, the second and third car type experience fewer claims than the first, and the third car type experiences fewer claims than the second.

Goodness-of-fit statistics: Comparing the deviance D = 2.82 (in the model with the offset) to a chi-square with 2 degrees of freedom, leads to the probability value $P(\chi^2(2) \geq 2.82) = 1 - 0.7559 = 0.2441$. The deviance does not exceed the critical 95th percentile (5.99) and the probability value is larger than 0.05. Hence there is no sign of overdispersion and there is no need to adjust the analysis.

<u>Fitting results for the model with an offset:</u>

```
                         The GENMOD Procedure

                          Model Information

                  Data Set                WORK.EXER12N2
                  Distribution                  Poisson
                  Link Function                     Log
                  Dependent Variable           nuclaims
                  Offset Variable               lnnupol
                  Observations Used                   6

                       Class Level Information

                                        Design
                     Class     Value     Variables

                     age       1          0
                               2          1

                     car       1          0      0
                               2          1      0
                               3          0      1

                        Parameter Information

                Parameter       Effect       age    car

                Prm1            Intercept
                Prm2            age          2
                Prm3            car                 2
                Prm4            car                 3

               Criteria For Assessing Goodness Of Fit

      Criterion                 DF           Value        Value/DF

      Deviance                   2          2.8207          1.4103
      Scaled Deviance            2          2.8207          1.4103
      Pearson Chi-Square         2          2.8416          1.4208
      Scaled Pearson X2          2          2.8416          1.4208
      Log Likelihood                      837.4533

Algorithm converged.


                      Estimated Correlation Matrix
```

	Prm1	Prm2	Prm3	Prm4
Prm1	1.0000	-0.7729	-0.5286	-0.1298
Prm2	-0.7729	1.0000	0.1487	-0.0841
Prm3	-0.5286	0.1487	1.0000	0.1877
Prm4	-0.1298	-0.0841	0.1877	1.0000

Analysis Of Parameter Estimates

Parameter		DF	Estimate	Standard Error	Wald 95% Confidence Limits		Chi-Square	Pr > ChiSq
Intercept		1	-2.6367	0.1318	-2.8950	-2.3784	400.20	<.0001
age	2	1	1.3199	0.1359	1.0536	1.5863	94.34	<.0001
car	2	1	-0.6928	0.1282	-0.9441	-0.4414	29.18	<.0001
car	3	1	-1.7643	0.2724	-2.2981	-1.2304	41.96	<.0001
Scale		0	1.0000	0.0000	1.0000	1.0000		

NOTE: The scale parameter was held fixed.

LR Statistics For Type 3 Analysis

Source	DF	Chi-Square	Pr > ChiSq
age	1	104.64	<.0001
car	2	72.82	<.0001

Finally, we estimate the model with an interaction term. This is a saturated model with the same number of parameters as observations. The output is given below. The type 3 analysis indicates that the interaction is not needed. Now you may wonder why it is possible to test for an interaction term in a saturated model. In the usual (normal) linear model this would not be possible as the saturated model leaves no degrees of freedom for the error term. With a Poisson link, however, the variance is the same as the mean and there is no extra parameter (variance or dispersion parameter) that needs to be estimated; the program indicates this fact when it says that the scale parameter was held fixed. Hence we can compare the log-likelihood of the full (saturated) model (838.8636) with the log-likelihood of the model without the interaction (837.4533) and compute the log- likelihood ratio test statistic 2(838.8636-837.4533)=2.82. Since its probability value $\mathrm{P}(\chi^2(2) \geq 2.82) = 1 - 0.7559 = 0.2441$ exceeds 0.05, the interaction is insignificant and we can use the simpler model without interaction. Note that the likelihood ratio test statistic for the interaction in the saturated model is identical to the deviance in the model without the interaction component.

<u>Fitting results for the model with interaction:</u>

The GENMOD Procedure

Model Information

Data Set	WORK.EXER12N2
Distribution	Poisson
Link Function	Log
Dependent Variable	nuclaims
Offset Variable	lnnupol
Observations Used	6

Class Level Information

Class	Value	Design Variables	
age	1	0	
	2	1	
car	1	0	0
	2	1	0
	3	0	1

Parameter Information

Parameter	Effect	age	car
Prm1	Intercept		
Prm2	age	2	
Prm3	car		2
Prm4	car		3
Prm5	age*car	2	2
Prm6	age*car	2	3

Criteria For Assessing Goodness Of Fit

Criterion	DF	Value	Value/DF
Deviance	0	0.0000	.
Scaled Deviance	0	0.0000	.
Pearson Chi-Square	0	0.0000	.
Scaled Pearson X2	0	0.0000	.
Log Likelihood		838.8636	

Algorithm converged.

The GENMOD Procedure

Estimated Correlation Matrix

	Prm1	Prm2	Prm3	Prm4	Prm5	Prm6
Prm1	1.0000	-0.8404	-0.6844	-0.1525	0.5656	0.1468
Prm2	-0.8404	1.0000	0.5751	0.1282	-0.6730	-0.1747
Prm3	-0.6844	0.5751	1.0000	0.1044	-0.8264	-0.1005
Prm4	-0.1525	0.1282	0.1044	1.0000	-0.0862	-0.9625
Prm5	0.5656	-0.6730	-0.8264	-0.0862	1.0000	0.1175
Prm6	0.1468	-0.1747	-0.1005	-0.9625	0.1175	1.0000

Analysis Of Parameter Estimates

Parameter			DF	Estimate	Standard Error	Wald 95% Confidence Limits		Chi-Square	Pr > ChiSq
Intercept			1	-2.4769	0.1543	-2.7794	-2.1745	257.68	<.0001
age	2		1	1.1006	0.1836	0.7407	1.4605	35.93	<.0001
car	2		1	-1.0022	0.2255	-1.4441	-0.5603	19.76	<.0001
car	3		1	-2.1282	1.0118	-4.1114	-0.1451	4.42	0.0354
age*car	2	2	1	0.4544	0.2728	-0.0803	0.9892	2.77	0.0958
age*car	2	3	1	0.4399	1.0513	-1.6206	2.5003	0.18	0.6757
Scale			0	1.0000	0.0000	1.0000	1.0000		

NOTE: The scale parameter was held fixed.

LR Statistics For Type 3 Analysis

Source	DF	Chi-Square	Pr > ChiSq

age	1	40.03	<.0001
car	2	23.43	<.0001
age*car	2	2.82	0.2441

12.3 We use SAS GENMOD to estimate the Poisson regression model with link

$$\ln \mu = \lambda_1 \mathrm{T}_1 + \lambda_2 \mathrm{T}_2$$

The deviance is D = 4.00 and the standardized deviance is 0.67. While the standardized deviance is somewhat smaller than one, the deviance is not small enough to suggest underdispersion ($\mathrm{P}(\chi^2(6) \le 4.00) = 0.32$).

The estimate of λ_2 is not significantly different from zero; the likelihood ratio test statistic is 0.81 with probability value 0.3685 (larger than 0.05). Alternatively, one can look at the confidence interval for λ_2; it covers zero.

The model without T_2 (that is, the Poisson regression with link $\ln \mu = \lambda_1 T_1$) is estimated next). The estimate of λ_1 is significant. A scatter plot of the observations against T_1, and the Poisson fit $\hat{\mu} = \exp(\hat{\lambda}_1 T_1)$ are shown below.

<u>Fitting results for the model with T_1 and T_2:</u>

```
                         The GENMOD Procedure

                          Model Information

                  Data Set                WORK.EXER12N3
                  Distribution                  Poisson
                  Link Function                     Log
                  Dependent Variable             nufail
                  Observations Used                   9

                        Parameter Information

                       Parameter       Effect

                       Prm1            Intercept
                       Prm2            time1
                       Prm3            time2

                 Criteria For Assessing Goodness Of Fit

      Criterion                 DF           Value        Value/DF

      Deviance                   6          4.0033          0.6672
      Scaled Deviance            6          4.0033          0.6672
      Pearson Chi-Square         6          3.9505          0.6584
      Scaled Pearson X2          6          3.9505          0.6584
      Log Likelihood                      362.7354

Algorithm converged.

                      Estimated Correlation Matrix
```

	Prm1	Prm2	Prm3
Prm1	1.0000	-0.7791	-0.2690
Prm2	-0.7791	1.0000	-0.3272
Prm3	-0.2690	-0.3272	1.0000

Analysis Of Parameter Estimates

Parameter	DF	Estimate	Standard Error	Wald 95% Confidence Limits		Chi-Square	Pr > ChiSq
Intercept	1	2.1752	0.2555	1.6745	2.6759	72.50	<.0001
time1	1	0.0070	0.0024	0.0023	0.0118	8.34	0.0039
time2	1	0.0025	0.0028	-0.0030	0.0081	0.81	0.3685
Scale	0	1.0000	0.0000	1.0000	1.0000		

NOTE: The scale parameter was held fixed.

<u>Fitting results for the model without T_2:</u>

The GENMOD Procedure

Model Information

Data Set	WORK.EXER12N3
Distribution	Poisson
Link Function	Log
Dependent Variable	nufail
Observations Used	9

Parameter Information

Parameter	Effect
Prm1	Intercept
Prm2	time1

Criteria For Assessing Goodness Of Fit

Criterion	DF	Value	Value/DF
Deviance	7	4.8078	0.6868
Scaled Deviance	7	4.8078	0.6868
Pearson Chi-Square	7	4.6345	0.6621
Scaled Pearson X2	7	4.6345	0.6621
Log Likelihood		362.3331	

Algorithm converged.

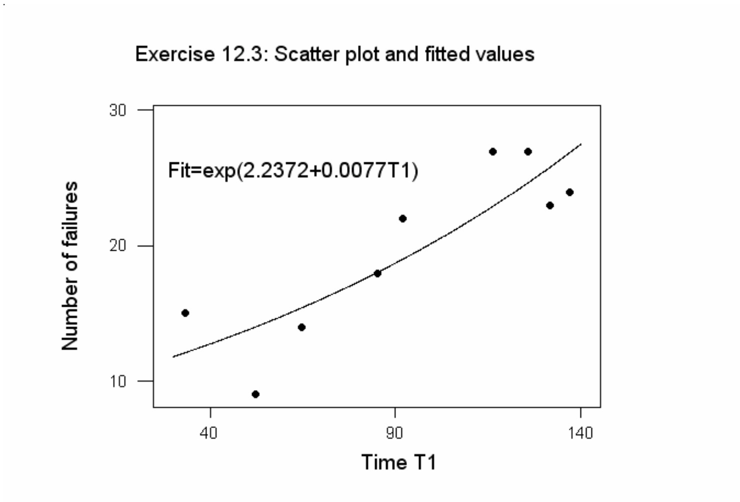
Estimated Correlation Matrix

	Prm1	Prm2
Prm1	1.0000	-0.9515
Prm2	-0.9515	1.0000

Analysis Of Parameter Estimates

Parameter	DF	Estimate	Standard Error	Wald 95% Confidence Limits		Chi-Square	Pr > ChiSq
Intercept	1	2.2372	0.2431	1.7608	2.7136	84.72	<.0001
time1	1	0.0077	0.0023	0.0033	0.0121	11.58	0.0007
Scale	0	1.0000	0.0000	1.0000	1.0000		

NOTE: The scale parameter was held fixed.



12.4

(a) Cancer incidence should be directly proportional to the size of the population. Hence it is reasonable to consider $\ln(\text{POP})$ as an offset. Age is a categorical variable. We use indicator variables for the eight age groups (X1 through X8) and consider the Poisson regression with link

$$\ln \mu = \beta_0 + \ln(\text{POP}) + \beta_2 \text{X2} + ... + \beta_8 \text{X8} + \beta_9 \text{Town}$$

The results of the model fit are shown below. Both age and town are significant; you can see this from the (partial; type 3) likelihood-ratio test statistics and their probability values at the end of the output. The estimate of the town effect is $\hat{\beta}_9 = 0.85$, with standard error 0.06. There is a significant location effect; women in Texas have a $100[\exp(0.85) - 1] = 134$ percent higher incidence of skin cancer. The deviance and the Pearson Chi-Square statistics are approximately one and indicate no problem with over/under-dispersion.

<u>Fitting results for the full model with an offset:</u>

```
                          The GENMOD Procedure

                           Model Information

                 Data Set                 WORK.EXER12N4
                 Distribution                   Poisson
                 Link Function                      Log
                 Dependent Variable             nucases
                 Offset Variable                  lnpop
                 Observations Used                   15

                        Class Level Information

   Class     Value                  Design Variables

   age       1          0      0      0      0      0      0      0
             2          1      0      0      0      0      0      0
```

3	0	1	0	0	0	0	0
4	0	0	1	0	0	0	0
5	0	0	0	1	0	0	0
6	0	0	0	0	1	0	0
7	0	0	0	0	0	1	0
8	0	0	0	0	0	0	1

Parameter Information

Parameter	Effect	age
Prm1	Intercept	
Prm2	town	
Prm3	age	2
Prm4	age	3
Prm5	age	4
Prm6	age	5
Prm7	age	6
Prm8	age	7
Prm9	age	8

Criteria For Assessing Goodness Of Fit

Criterion	DF	Value	Value/DF
Deviance	6	5.2089	0.8682
Scaled Deviance	6	5.2089	0.8682
Pearson Chi-Square	6	5.1482	0.8580
Scaled Pearson X2	6	5.1482	0.8580
Log Likelihood		6204.3156	

Estimated Correlation Matrix

	Prm1	Prm2	Prm3	Prm4	Prm5	Prm6	Prm7	Prm8	Prm9
Prm1	1.0000	-0.0944	-0.9521	-0.9788	-0.9868	-0.9885	-0.9900	-0.9819	-0.9730
Prm2	-0.0944	1.0000	-0.0031	-0.0047	-0.0037	-0.0024	0.0007	0.0927	0.0039
Prm3	-0.9521	-0.0031	1.0000	0.9410	0.9486	0.9501	0.9513	0.9349	0.9347
Prm4	-0.9788	-0.0047	0.9410	1.0000	0.9753	0.9769	0.9781	0.9610	0.9610
Prm5	-0.9868	-0.0037	0.9486	0.9753	1.0000	0.9847	0.9860	0.9689	0.9687
Prm6	-0.9885	-0.0024	0.9501	0.9769	0.9847	1.0000	0.9875	0.9706	0.9703
Prm7	-0.9900	0.0007	0.9513	0.9781	0.9860	0.9875	1.0000	0.9721	0.9715
Prm8	-0.9819	0.0927	0.9349	0.9610	0.9689	0.9706	0.9721	1.0000	0.9554
Prm9	-0.9730	0.0039	0.9347	0.9610	0.9687	0.9703	0.9715	0.9554	1.0000

Analysis Of Parameter Estimates

Parameter		DF	Estimate	Standard Error	Wald 95% Confidence Limits		Chi-Square	Pr > ChiSq
Intercept		1	-11.6921	0.4492	-12.5725	-10.8116	677.43	<.0001
town		1	0.8527	0.0596	0.7358	0.9696	204.54	<.0001
age	2	1	2.6290	0.4675	1.7128	3.5452	31.63	<.0001
age	3	1	3.8456	0.4547	2.9545	4.7367	71.54	<.0001
age	4	1	4.5938	0.4510	3.7098	5.4778	103.74	<.0001
age	5	1	5.0864	0.4503	4.2038	5.9690	127.59	<.0001
age	6	1	5.6457	0.4497	4.7642	6.5272	157.58	<.0001
age	7	1	6.2032	0.4575	5.3065	7.0999	183.83	<.0001
age	8	1	6.1757	0.4577	5.2785	7.0728	182.02	<.0001
Scale		0	1.0000	0.0000	1.0000	1.0000		

NOTE: The scale parameter was held fixed.

LR Statistics For Type 3 Analysis

Source	DF	Chi-Square	Pr > ChiSq
town	1	226.52	<.0001
age	7	2199.01	<.0001

(b) The estimation results for the more general model

$$\ln \mu = \beta_0 + \beta_1 \ln(\text{POP}) + \beta_2 \text{X2} + ... + \beta_8 \text{X8} + \beta_9 \text{Town}$$

are given below. It seems reasonable to suppose that the number of cancers is directly proportional to the population, and that one can expect that the coefficient β_1 is one. Let us test whether $\beta_1 = 1$. The estimate is $\hat{\beta}_1 = 1.96$, and the 95 percent Wald confidence interval is given by $1.96 \pm (1.96)(0.63)$, 1.96 ± 1.23, or $0.73 \leq \beta_1 \leq 3.18$. The interval is quite wide (because there are few observations). The interval includes one, which makes the off-set interpretation plausible.

<u>Fitting results for the full model without an offset:</u>

The GENMOD Procedure

Model Information

Data Set	WORK.EXER12N4
Distribution	Poisson
Link Function	Log
Dependent Variable	nucases
Observations Used	15

Class Level Information

Class	Value	Design Variables						
age	1	0	0	0	0	0	0	0
	2	1	0	0	0	0	0	0
	3	0	1	0	0	0	0	0
	4	0	0	1	0	0	0	0
	5	0	0	0	1	0	0	0
	6	0	0	0	0	1	0	0
	7	0	0	0	0	0	1	0
	8	0	0	0	0	0	0	1

Parameter Information

Parameter	Effect	age
Prm1	Intercept	
Prm2	lnpop	
Prm3	town	
Prm4	age	2
Prm5	age	3
Prm6	age	4
Prm7	age	5
Prm8	age	6
Prm9	age	7
Prm10	age	8

Criteria For Assessing Goodness Of Fit

Criterion	DF	Value	Value/DF
Deviance	5	2.8539	0.5708
Scaled Deviance	5	2.8539	0.5708
Pearson Chi-Square	5	2.8439	0.5688
Scaled Pearson X2	5	2.8439	0.5688
Log Likelihood		6205.4931	

The GENMOD Procedure

Algorithm converged.

Estimated Correlation Matrix

	Prm1	Prm2	Prm3	Prm4	Prm5	Prm6	Prm7	Prm8	Prm9	Prm10
Prm1	1.0000	-0.9982	0.7154	-0.3692	-0.5729	-0.6353	-0.7844	-0.8810	-0.9360	-0.9851
Prm2	-0.9982	1.0000	-0.7206	0.3160	0.5241	0.5888	0.7465	0.8516	0.9138	0.9736
Prm3	0.7154	-0.7206	1.0000	-0.2317	-0.3831	-0.4284	-0.5401	-0.6131	-0.6327	-0.7003
Prm4	-0.3692	0.3160	-0.2317	1.0000	0.9260	0.9135	0.8357	0.7422	0.6489	0.5100
Prm5	-0.5729	0.5241	-0.3831	0.9260	1.0000	0.9800	0.9448	0.8830	0.8112	0.6971
Prm6	-0.6353	0.5888	-0.4284	0.9135	0.9800	1.0000	0.9691	0.9192	0.8560	0.7520
Prm7	-0.7844	0.7465	-0.5401	0.8357	0.9448	0.9691	1.0000	0.9802	0.9444	0.8742
Prm8	-0.8810	0.8516	-0.6131	0.7422	0.8830	0.9192	0.9802	1.0000	0.9852	0.9453
Prm9	-0.9360	0.9138	-0.6327	0.6489	0.8112	0.8560	0.9444	0.9852	1.0000	0.9783
Prm10	-0.9851	0.9736	-0.7003	0.5100	0.6971	0.7520	0.8742	0.9453	0.9783	1.0000

Analysis Of Parameter Estimates

Parameter		DF	Estimate	Standard Error	Wald 95% Confidence Limits		Chi-Square	Pr > ChiSq
Intercept		1	-23.2489	7.5392	-38.0256	-8.4723	9.51	0.0020
lnpop		1	1.9613	0.6259	0.7345	3.1880	9.82	0.0017
town		1	0.7556	0.0862	0.5866	0.9245	76.81	<.0001
age	2	1	2.8684	0.4927	1.9027	3.8341	33.89	<.0001
age	3	1	4.2766	0.5339	3.2303	5.3230	64.17	<.0001
age	4	1	5.0990	0.5580	4.0053	6.1927	83.49	<.0001
age	5	1	5.8623	0.6768	4.5358	7.1888	75.02	<.0001
age	6	1	6.7681	0.8579	5.0866	8.4496	62.23	<.0001
age	7	1	7.7827	1.1265	5.5748	9.9906	47.73	<.0001
age	8	1	9.1783	2.0057	5.2473	13.1094	20.94	<.0001
Scale		0	1.0000	0.0000	1.0000	1.0000		

NOTE: The scale parameter was held fixed.

LR Statistics For Type 3 Analysis

Source	DF	Chi-Square	Pr > ChiSq
lnpop	1	9.77	0.0018
town	1	81.60	<.0001
age	7	988.50	<.0001

Additional model: We estimate a model that includes an interaction between town and age. We want to check whether the town effect depends on the age group. The results are given below. The likelihood-ratio test for the town by age interaction is insignificant. Note that such a test is possible in the saturated Poisson regression model, as the variance is the same as the mean; the scale parameter is kept fixed.

Fitting results for the model with interaction:

```
                         The GENMOD Procedure

                          Model Information

                 Data Set                WORK.EXER12N4
                 Distribution                  Poisson
                 Link Function                     Log
                 Dependent Variable            nucases
                 Offset Variable                 lnpop
                 Observations Used                  15

                       Class Level Information

     Class     Value                  Design Variables

     age       1          0      0      0      0      0      0      0
               2          1      0      0      0      0      0      0
               3          0      1      0      0      0      0      0
               4          0      0      1      0      0      0      0
               5          0      0      0      1      0      0      0
               6          0      0      0      0      1      0      0
               7          0      0      0      0      0      1      0
               8          0      0      0      0      0      0      1
     town      0          0
               1          1

                        Parameter Information

               Parameter       Effect       age    town

               Prm1            Intercept
               Prm2            town                1
               Prm3            age          2
               Prm4            age          3
               Prm5            age          4
               Prm6            age          5
               Prm7            age          6
               Prm8            age          7
               Prm9            age          8
               Prm10           age*town     2      1
               Prm11           age*town     3      1
               Prm12           age*town     4      1
               Prm13           age*town     5      1
               Prm14           age*town     6      1
               Prm15           age*town     7      1
               Prm16           age*town     8      1

                Criteria For Assessing Goodness Of Fit

       Criterion                 DF           Value        Value/DF

       Deviance                   0          0.0000            .
       Scaled Deviance            0          0.0000            .
       Pearson Chi-Square         0          0.0000            .
       Scaled Pearson X2          0          0.0000            .
       Log Likelihood                     6206.9201

 Algorithm converged.
```

Analysis Of Parameter Estimates

Parameter			DF	Estimate	Standard Error	Wald 95% Confidence Limits		Chi-Square	Pr > ChiSq
Intercept			1	-12.0592	1.0000	-14.0191	-10.0992	145.42	<.0001
town	1		1	1.3373	1.1180	-0.8540	3.5286	1.43	0.2316
age	2		1	3.1113	1.0308	1.0910	5.1316	9.11	0.0025
age	3		1	3.9860	1.0165	1.9937	5.9784	15.38	<.0001
age	4		1	4.8917	1.0070	2.9180	6.8655	23.60	<.0001
age	5		1	5.4975	1.0049	3.5280	7.4671	29.93	<.0001
age	6		1	6.0167	1.0038	4.0492	7.9842	35.92	<.0001
age	7		1	6.5703	1.0038	4.6029	8.5376	42.85	<.0001
age	8		1	6.7207	1.0124	4.7364	8.7050	44.07	<.0001
age*town	2	1	1	-0.6446	1.1571	-2.9124	1.6232	0.31	0.5774
age*town	3	1	1	-0.1917	1.1365	-2.4193	2.0359	0.03	0.8661
age*town	4	1	1	-0.3922	1.1263	-2.5998	1.8154	0.12	0.7277
age*town	5	1	1	-0.5455	1.1241	-2.7487	1.6578	0.24	0.6275
age*town	6	1	1	-0.4901	1.1229	-2.6910	1.7107	0.19	0.6625
age*town	7	1	0	0.0000	0.0000	0.0000	0.0000	.	.
age*town	8	1	1	-0.7581	1.1360	-2.9845	1.4683	0.45	0.5045
Scale			0	1.0000	0.0000	1.0000	1.0000		

NOTE: The scale parameter was held fixed.

LR Statistics For Type 3 Analysis

Source	DF	Chi-Square	Pr > ChiSq
town	1	1.78	0.1817
age	7	845.79	<.0001
age*town	7	5.21	0.6342

Another model: Finally, we introduce age as a continuous variable, and not as a factor as was done in the previous models. The output is shown below. Both age and town are significant. A graph of the number of cancer deaths against age (with the two towns indicated by different plotting symbols) and the Poisson model fit is given in the following graph. Every ten years the cancer rate (deaths per population) increases by a factor of exp(0.6133) = 1.85; that is, by 85 percent.

Fitting results for the model with age as continuous variable:

The GENMOD Procedure

Model Information

Data Set	WORK.EXER12N4
Distribution	Poisson
Link Function	Log
Dependent Variable	nucases
Offset Variable	lnpop
Observations Used	15

Criteria For Assessing Goodness Of Fit

Criterion	DF	Value	Value/DF
Deviance	12	184.8091	15.4008
Scaled Deviance	12	184.8091	15.4008
Pearson Chi-Square	12	141.4307	11.7859

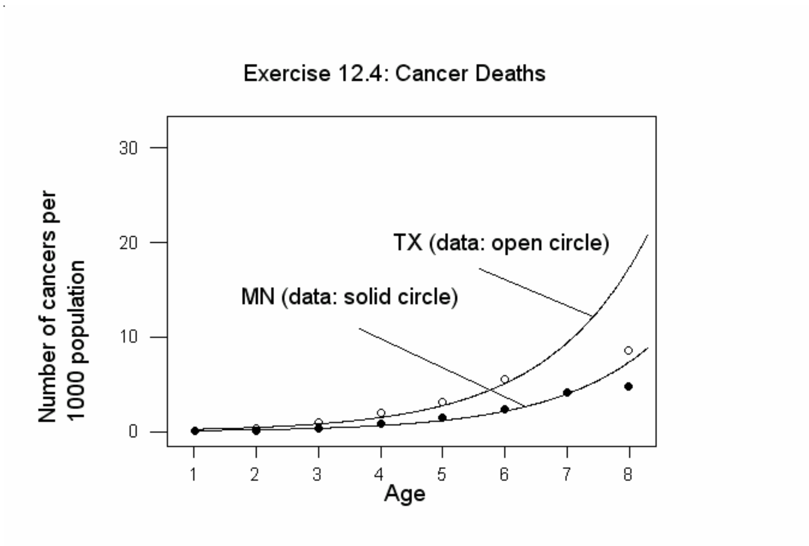
Scaled Pearson X2	12	141.4307	11.7859	
Log Likelihood		6114.5155		

Algorithm converged.

Analysis Of Parameter Estimates

Parameter	DF	Estimate	Standard Error	Wald 95% Confidence Limits		Chi-Square	Pr > ChiSq
Intercept	1	-9.8191	0.0902	-9.9959	-9.6423	11846.5	<.0001
town	1	0.8584	0.0545	0.7515	0.9652	247.95	<.0001
age	1	0.6133	0.0142	0.5855	0.6411	1871.42	<.0001
Scale	0	1.0000	0.0000	1.0000	1.0000		

NOTE: The scale parameter was held fixed.



12.5 We use SAS GENMOD to estimate the Poisson regression model with link

$$\ln \mu = \beta_0 + \beta_1 \ln(\text{Pop}) + \beta_2 \text{X2} + ... + \beta_9 \text{X9} + \beta_{10} \text{Z2} + \beta_{11} \text{Z3} + \beta_{12} \text{Z4}$$

The output shows that age and smoking are statistically significant factors. Lung cancer deaths increase monotonically with age. Lung cancer deaths also increase with smoking. The situation is worst for people who smoke cigarettes only (smoking = 4). The surprising fact that people who smoke cigarettes and pipe (or cigar) have lower incidences is probably explained by the number of cigarettes smoked (which is not recorded). People who smoke cigarettes only probably smoke more cigarettes than people who smoke both cigarettes and pipe (or cigar).

The deviance is D = 16.38 and the standardized deviance is 0.71. While the standardized deviance is somewhat smaller than one, the deviance is not small enough to suggest underdispersion ($P(\chi^2(23) \le 16.38) = 0.16$).

Let us test whether $\beta_1 = 1$. The estimate is $\hat{\beta}_1 = 1.0761$, and the 95 percent Wald confidence interval is given by $1.0761 \pm (1.96)(0.0340)$, 1.076 ± 0.067, or $1.01 \le \beta_1 \le 1.14$. The interval fails to cover one – however just barely (the lower limit is about one). While we would reject at the 0.05 significance level that $\beta_1 = 1$, the off-set interpretation is not entirely implausible.

Fitting results for the model without an offset:

```
                       The GENMOD Procedure

                        Model Information

                 Data Set              WORK.EXER12N5
                 Distribution                Poisson
                 Link Function                   Log
                 Dependent Variable          nudeath
                 Observations Used                36

                     Class Level Information

     Class       Value                      Design Variables

     age         1          0      0      0      0      0      0      0      0
                 2          1      0      0      0      0      0      0      0
                 3          0      1      0      0      0      0      0      0
                 4          0      0      1      0      0      0      0      0
                 5          0      0      0      1      0      0      0      0
                 6          0      0      0      0      1      0      0      0
                 7          0      0      0      0      0      1      0      0
                 8          0      0      0      0      0      0      1      0
                 9          0      0      0      0      0      0      0      1
     smoking     1          0      0      0
                 2          1      0      0
                 3          0      1      0
                 4          0      0      1

                      Parameter Information

               Parameter       Effect       age    smoking

               Prm1            Intercept
               Prm2            lnpop
               Prm3            age          2
               Prm4            age          3
               Prm5            age          4
               Prm6            age          5
               Prm7            age          6
               Prm8            age          7
               Prm9            age          8
               Prm10           age          9
               Prm11           smoking             2
               Prm12           smoking             3
               Prm13           smoking             4

                       The GENMOD Procedure

                 Criteria For Assessing Goodness Of Fit

            Criterion                 DF           Value        Value/DF

            Deviance                  23         16.3820          0.7123
```

Scaled Deviance	23	16.3820	0.7123
Pearson Chi-Square	23	16.3745	0.7119
Scaled Pearson X2	23	16.3745	0.7119
Log Likelihood		45620.8854	

Algorithm converged.

Analysis Of Parameter Estimates

Parameter		DF	Estimate	Standard Error	Wald 95% Confidence Limits		Chi-Square	Pr > ChiSq
Intercept		1	-4.2192	0.2505	-4.7103	-3.7282	283.61	<.0001
lnpop		1	1.0761	0.0340	1.0095	1.1427	1002.25	<.0001
age	2	1	0.5855	0.0812	0.4263	0.7447	51.97	<.0001
age	3	1	1.0304	0.0800	0.8736	1.1872	165.93	<.0001
age	4	1	1.3814	0.0653	1.2535	1.5093	447.97	<.0001
age	5	1	1.6401	0.0629	1.5169	1.7634	680.41	<.0001
age	6	1	2.0158	0.0633	1.8917	2.1398	1014.09	<.0001
age	7	1	2.3330	0.0701	2.1957	2.4704	1108.03	<.0001
age	8	1	2.6721	0.0848	2.5060	2.8383	993.31	<.0001
age	9	1	2.9916	0.0970	2.8015	3.1817	951.64	<.0001
smoking	2	1	0.0148	0.0494	-0.0820	0.1117	0.09	0.7643
smoking	3	1	0.1159	0.0598	-0.0012	0.2330	3.76	0.0524
smoking	4	1	0.3485	0.0503	0.2498	0.4471	47.91	<.0001
Scale		0	1.0000	0.0000	1.0000	1.0000		

NOTE: The scale parameter was held fixed.

LR Statistics For Type 3 Analysis

Source	DF	Chi-Square	Pr > ChiSq
lnpop	1	1244.55	<.0001
age	8	3254.75	<.0001
smoking	3	143.40	<.0001

The regression results treating ln(POP) as an offset are given next. The interpretation of the results is mostly unchanged.

Fitting results for the model with an offset:

The GENMOD Procedure

Model Information

Data Set	WORK.EXER12N5
Distribution	Poisson
Link Function	Log
Dependent Variable	nudeath
Offset Variable	lnpop
Observations Used	36

Class Level Information

Class	Value	Design Variables							
age	1	0	0	0	0	0	0	0	0
	2	1	0	0	0	0	0	0	0
	3	0	1	0	0	0	0	0	0

	4	0	0	1	0	0	0	0	0
	5	0	0	0	1	0	0	0	0
	6	0	0	0	0	1	0	0	0
	7	0	0	0	0	0	1	0	0
	8	0	0	0	0	0	0	1	0
	9	0	0	0	0	0	0	0	1
smoking	1	0	0	0					
	2	1	0	0					
	3	0	1	0					
	4	0	0	1					

Parameter Information

Parameter	Effect	age	smoking
Prm1	Intercept		
Prm2	age	2	
Prm3	age	3	
Prm4	age	4	
Prm5	age	5	
Prm6	age	6	
Prm7	age	7	
Prm8	age	8	
Prm9	age	9	
Prm10	smoking		2
Prm11	smoking		3
Prm12	smoking		4

The GENMOD Procedure

Criteria For Assessing Goodness Of Fit

Criterion	DF	Value	Value/DF
Deviance	24	21.4867	0.8953
Scaled Deviance	24	21.4867	0.8953
Pearson Chi-Square	24	20.6194	0.8591
Scaled Pearson X2	24	20.6194	0.8591
Log Likelihood		45618.3330	

Algorithm converged.

Analysis Of Parameter Estimates

Parameter		DF	Estimate	Standard Error	Wald 95% Confidence Limits		Chi-Square	Pr > ChiSq
Intercept		1	-3.6800	0.0682	-3.8138	-3.5463	2908.35	<.0001
age	2	1	0.5539	0.0800	0.3971	0.7107	47.95	<.0001
age	3	1	0.9804	0.0768	0.8298	1.1309	162.88	<.0001
age	4	1	1.3795	0.0653	1.2515	1.5074	446.80	<.0001
age	5	1	1.6542	0.0626	1.5316	1.7769	699.00	<.0001
age	6	1	1.9982	0.0628	1.8751	2.1212	1012.79	<.0001
age	7	1	2.2714	0.0644	2.1453	2.3975	1245.78	<.0001
age	8	1	2.5586	0.0678	2.4257	2.6914	1424.74	<.0001
age	9	1	2.8469	0.0724	2.7050	2.9889	1545.27	<.0001
smoking	2	1	0.0478	0.0470	-0.0443	0.1399	1.03	0.3090
smoking	3	1	0.2180	0.0387	0.1421	0.2938	31.73	<.0001
smoking	4	1	0.4170	0.0399	0.3387	0.4952	109.14	<.0001
Scale		0	1.0000	0.0000	1.0000	1.0000		

NOTE: The scale parameter was held fixed.

LR Statistics For Type 3 Analysis

Source	DF	Chi-Square	Pr > ChiSq
age	8	3889.22	<.0001
smoking	3	170.24	<.0001

Results of the Poisson regression with the factors age, smoking and the interaction between age and smoking is given below. The interaction between age and smoking turns out to be insignificant.

Fitting results for the model with interaction:

The GENMOD Procedure

Model Information

Data Set	WORK.EXER12N5
Distribution	Poisson
Link Function	Log
Dependent Variable	nudeath
Offset Variable	lnpop
Observations Used	36

Class Level Information

Class	Value	Design Variables							
age	1	0	0	0	0	0	0	0	0
	2	1	0	0	0	0	0	0	0
	3	0	1	0	0	0	0	0	0
	4	0	0	1	0	0	0	0	0
	5	0	0	0	1	0	0	0	0
	6	0	0	0	0	1	0	0	0
	7	0	0	0	0	0	1	0	0
	8	0	0	0	0	0	0	1	0
	9	0	0	0	0	0	0	0	1
smoking	1	0	0	0					
	2	1	0	0					
	3	0	1	0					
	4	0	0	1					

Criteria For Assessing Goodness Of Fit

Criterion	DF	Value	Value/DF
Deviance	0	0.0000	.
Scaled Deviance	0	0.0000	.
Pearson Chi-Square	0	0.0000	.
Scaled Pearson X2	0	0.0000	.
Log Likelihood		45629.0764	

Algorithm converged.

Analysis Of Parameter Estimates

Parameter		DF	Estimate	Standard Error	Wald 95% Confidence Limits		Chi-Square	Pr > ChiSq
Intercept		1	-3.5958	0.2357	-4.0578	-3.1338	232.73	<.0001
age	2	1	0.8035	0.3178	0.1806	1.4264	6.39	0.0115
age	3	1	1.0228	0.3289	0.3781	1.6674	9.67	0.0019
age	4	1	1.1542	0.2715	0.6220	1.6865	18.07	<.0001
age	5	1	1.3854	0.2532	0.8891	1.8816	29.94	<.0001
age	6	1	1.9325	0.2479	1.4467	2.4183	60.79	<.0001
age	7	1	2.2789	0.2473	1.7942	2.7635	84.94	<.0001
age	8	1	2.4944	0.2528	1.9990	2.9898	97.39	<.0001
age	9	1	2.7702	0.2528	2.2747	3.2656	120.11	<.0001
smoking	2	1	-0.6878	0.7454	-2.1487	0.7731	0.85	0.3561
smoking	3	1	0.1810	0.2495	-0.3080	0.6701	0.53	0.4681

smoking	4		1	0.2816	0.2522	-0.2128	0.7760	1.25	0.2642
age*smoking	2	2	1	0.2220	0.9225	-1.5861	2.0301	0.06	0.8098
age*smoking	2	3	1	-0.2752	0.3371	-0.9359	0.3855	0.67	0.4143
age*smoking	2	4	1	-0.2615	0.3409	-0.9296	0.4067	0.59	0.4431
age*smoking	3	2	1	-0.2255	0.9703	-2.1273	1.6762	0.05	0.8162
age*smoking	3	3	1	-0.0715	0.3465	-0.7507	0.6076	0.04	0.8364
age*smoking	3	4	1	-0.0010	0.3487	-0.6844	0.6825	0.00	0.9978
age*smoking	4	2	1	0.8480	0.7746	-0.6702	2.3663	1.20	0.2736
age*smoking	4	3	1	0.1651	0.2867	-0.3967	0.7270	0.33	0.5646
age*smoking	4	4	1	0.3096	0.2894	-0.2576	0.8768	1.14	0.2846
age*smoking	5	2	1	0.8851	0.7569	-0.5985	2.3687	1.37	0.2423
age*smoking	5	3	1	0.2300	0.2680	-0.2952	0.7552	0.74	0.3907
age*smoking	5	4	1	0.3452	0.2710	-0.1860	0.8764	1.62	0.2028
age*smoking	6	2	1	0.6489	0.7531	-0.8272	2.1251	0.74	0.3889
age*smoking	6	3	1	0.0221	0.2632	-0.4937	0.5379	0.01	0.9330
age*smoking	6	4	1	0.1250	0.2664	-0.3971	0.6470	0.22	0.6390
age*smoking	7	2	1	0.6471	0.7522	-0.8272	2.1215	0.74	0.3896
age*smoking	7	3	1	-0.0630	0.2636	-0.5797	0.4536	0.06	0.8110
age*smoking	7	4	1	0.0178	0.2674	-0.5063	0.5420	0.00	0.9469
age*smoking	8	2	1	0.7795	0.7537	-0.6977	2.2566	1.07	0.3010
age*smoking	8	3	1	-0.0292	0.2712	-0.5608	0.5024	0.01	0.9142
age*smoking	8	4	1	0.1081	0.2768	-0.4344	0.6507	0.15	0.6961
age*smoking	9	2	1	0.7608	0.7536	-0.7161	2.2378	1.02	0.3127
age*smoking	9	3	1	0.0428	0.2755	-0.4971	0.5827	0.02	0.8765
age*smoking	9	4	1	-0.0402	0.2964	-0.6211	0.5406	0.02	0.8920
Scale			0	1.0000	0.0000	1.0000	1.0000		

NOTE: The scale parameter was held fixed.

LR Statistics For Type 3 Analysis

Source	DF	Chi-Square	Pr > ChiSq
age	8	382.21	<.0001
smoking	3	3.80	0.2843
age*smoking	24	21.49	0.6099

Finally, we consider the Poisson regression that includes smoking as a factor (with the three indicators) and age as a continuous variable. The results are given below. We can test whether a class factor for age is needed or whether it is sufficient to include age as a continuous variable. The log-likelihood of the model that considers age as a factor (the full model) is 45,618.3330; the log-likelihood of the model that considers age as a continuous variable (the restricted model) is 45,591.2091. We compare the log-likelihood ratio statistic, 2(45,618.3330-45,591.2091) = 54.25, to a chi-square with 7 degrees of freedom (the nine intercepts in the unrestricted model, one for each age group, are tested against the linear formulation which includes two parameters, the intercept and the slope). The test statistic is large (probability value $P(\chi^2(7) \geq 54.25) < 0.001$ is small), indicating that it is not adequate to consider a linear component of size. Size must be treated as a class factor.

<u>Fitting results for the model with age as continuous variable:</u>

The GENMOD Procedure

Model Information

Data Set WORK.EXER12N5

Distribution	Poisson
Link Function	Log
Dependent Variable	nudeath
Offset Variable	lnpop
Observations Used	36

Class Level Information

Class	Value	Design Variables		
smoking	1	0	0	0
	2	1	0	0
	3	0	1	0
	4	0	0	1

Criteria For Assessing Goodness Of Fit

Criterion	DF	Value	Value/DF
Deviance	31	75.7345	2.4430
Scaled Deviance	31	75.7345	2.4430
Pearson Chi-Square	31	71.9749	2.3218
Scaled Pearson X2	31	71.9749	2.3218
Log Likelihood		45591.2091	

Algorithm converged.

Analysis Of Parameter Estimates

Parameter		DF	Estimate	Standard Error	Wald 95% Confidence Limits		Chi-Square	Pr > ChiSq
Intercept		1	-3.7389	0.0500	-3.8369	-3.6409	5589.70	<.0001
age		1	0.3330	0.0056	0.3220	0.3440	3547.25	<.0001
smoking	2	1	0.0329	0.0469	-0.0590	0.1248	0.49	0.4826
smoking	3	1	0.2364	0.0386	0.1607	0.3120	37.50	<.0001
smoking	4	1	0.4379	0.0398	0.3599	0.5160	121.07	<.0001
Scale		0	1.0000	0.0000	1.0000	1.0000		

NOTE: The scale parameter was held fixed.

LR Statistics For Type 3 Analysis

Source	DF	Chi-Square	Pr > ChiSq
age	1	3834.97	<.0001
smoking	3	196.00	<.0001

12.6 The output from estimating the Poisson regression with link

$$\ln \mu = \beta_0 + \beta_1 \text{DIST} + \beta_2 \text{INC} + \beta_3 \text{SIZE2} + \beta_4 \text{SIZE3} + \beta_5 \text{SIZE4} + \beta_6 \text{SIZE5}$$

is shown below. Here we treat SIZE as a class variable, specifying 4 indicators for the factor with five outcomes (1 through 5 people; size 1 is the baseline). Income does not affect the number of visits to the lake (probability value = 0.27) and is omitted in the next run. The deviance and the Pearson chi-square statistics are roughly the size of the critical 95th percentile (280.36).

Fitting results for the full model:

The GENMOD Procedure

Model Information

Data Set	WORK.EXER12N6
Distribution	Poisson
Link Function	Log
Dependent Variable	nuvisits
Observations Used	250

Class Level Information

Class	Value	Design Variables			
size	1	0	0	0	0
	2	1	0	0	0
	3	0	1	0	0
	4	0	0	1	0
	5	0	0	0	1

Parameter Information

Parameter	Effect	size
Prm1	Intercept	
Prm2	dist	
Prm3	inc	
Prm4	size	2
Prm5	size	3
Prm6	size	4
Prm7	size	5

Criteria For Assessing Goodness Of Fit

Criterion	DF	Value	Value/DF
Deviance	243	313.6999	1.2909
Scaled Deviance	243	313.6999	1.2909
Pearson Chi-Square	243	286.2022	1.1778
Scaled Pearson X2	243	286.2022	1.1778
Log Likelihood		11.3651	

Algorithm converged.

Estimated Correlation Matrix

	Prm1	Prm2	Prm3	Prm4	Prm5	Prm6	Prm7
Prm1	1.0000	-0.4081	-0.6605	-0.4409	-0.5760	-0.5506	-0.5740
Prm2	-0.4081	1.0000	-0.0275	0.0247	0.1431	0.0573	0.0438
Prm3	-0.6605	-0.0275	1.0000	-0.0391	0.0536	0.0374	0.0749
Prm4	-0.4409	0.0247	-0.0391	1.0000	0.5739	0.5990	0.6019
Prm5	-0.5760	0.1431	0.0536	0.5739	1.0000	0.6386	0.6437
Prm6	-0.5506	0.0573	0.0374	0.5990	0.6386	1.0000	0.6678
Prm7	-0.5740	0.0438	0.0749	0.6019	0.6437	0.6678	1.0000

Analysis Of Parameter Estimates

Parameter	DF	Estimate	Standard Error	Wald 95% Confidence Limits		Chi-Square	Pr > ChiSq
Intercept	1	1.6578	0.1907	1.2840	2.0315	75.57	<.0001
dist	1	-0.0215	0.0016	-0.0245	-0.0184	190.19	<.0001

inc		1	0.0203	0.0184	-0.0158	0.0563	1.22	0.2700
size	2	1	-0.0249	0.1595	-0.3375	0.2877	0.02	0.8758
size	3	1	0.1032	0.1521	-0.1949	0.4014	0.46	0.4973
size	4	1	0.3344	0.1454	0.0495	0.6194	5.29	0.0214
size	5	1	0.4731	0.1442	0.1904	0.7558	10.76	0.0010
Scale		0	1.0000	0.0000	1.0000	1.0000		

NOTE: The scale parameter was held fixed.

LR Statistics For Type 3 Analysis

Source	DF	Chi-Square	Pr > ChiSq
dist	1	213.97	<.0001
inc	1	1.22	0.2699
size	4	21.19	0.0003

The output of the simplified Poisson regression with link

$$\ln \mu = \beta_0 + \beta_1 \text{DIST} + \beta_3 \text{SIZE2} + \beta_4 \text{SIZE3} + \beta_5 \text{SIZE4} + \beta_6 \text{SIZE5}$$

is shown below.

<u>Fitting results for the restricted model without income:</u>

The GENMOD Procedure

Model Information

Data Set	WORK.EXER12N6
Distribution	Poisson
Link Function	Log
Dependent Variable	nuvisits
Observations Used	250

Class Level Information

Class	Value	Design Variables			
size	1	0	0	0	0
	2	1	0	0	0
	3	0	1	0	0
	4	0	0	1	0
	5	0	0	0	1

Criteria For Assessing Goodness Of Fit

Criterion	DF	Value	Value/DF
Deviance	244	314.9173	1.2906
Scaled Deviance	244	314.9173	1.2906
Pearson Chi-Square	244	284.7341	1.1669
Scaled Pearson X2	244	284.7341	1.1669
Log Likelihood		10.7564	

Algorithm converged.

Analysis Of Parameter Estimates

Parameter	DF	Estimate	Standard Error	Wald 95% Confidence Limits		Chi-Square	Pr > ChiSq
Intercept	1	1.7957	0.1431	1.5152	2.0762	157.44	<.0001
dist	1	-0.0214	0.0016	-0.0245	-0.0184	189.88	<.0001

size	2	1	-0.0184	0.1594	-0.3308	0.2939	0.01	0.9079
size	3	1	0.0941	0.1519	-0.2035	0.3917	0.38	0.5355
size	4	1	0.3283	0.1453	0.0436	0.6130	5.11	0.0238
size	5	1	0.4610	0.1439	0.1790	0.7429	10.27	0.0014
Scale		0	1.0000	0.0000	1.0000	1.0000		

NOTE: The scale parameter was held fixed.

Additional model: Treating SIZE as a continuous variable and not as a factor leads to the Poisson link

$$\ln \mu = \beta_0 + \beta_1 \text{DIST} + \beta_2 \text{INC} + \beta_3 \text{SIZE}.$$

The estimation results show that income can be omitted (output not shown). Omitting income leads to the results shown below. Both distance and family size are statistically significant. A change in distance by 10 miles reduces the mean number of visits by a factor of exp(-0.0212(10)) = 0.81, or 19 percent. A change in the family size by one unit increases the mean number of visits by a factor exp(0.1358) = 1.145, or 14.5 percent.

We can test whether a class factor for size is needed or whether it is sufficient to treat size as a continuous variable. The log-likelihood of the model that considers size as a factor (the full model) is 10.7564; the log-likelihood of the model that considers size as a continuous variable (the restricted model) is 9.8849. We compare the log-likelihood ratio statistic, 2(10.7564-9.8849) =1.74, to a chi-square with 3 degrees of freedom (the five intercepts in the unrestricted model, one for each of the five size groups, are tested against the linear formulation which includes two parameters, the intercept and the slope). The test statistic is small (probability value $P(\chi^2(3) \geq 1.74) = 1 - 0.37 = 0.63$ is large), indicating that it is sufficient to consider a linear component of size. A scatter plot of the number of visits against distance and fitted values from the Poisson regression against distance is also shown.

<u>Fitting results for the model with size as continuous variable:</u>

The GENMOD Procedure

Model Information

Data Set	WORK.EXER12N6
Distribution	Poisson
Link Function	Log
Dependent Variable	nuvisits
Observations Used	250

Criteria For Assessing Goodness Of Fit

Criterion	DF	Value	Value/DF
Deviance	247	316.6602	1.2820
Scaled Deviance	247	316.6602	1.2820
Pearson Chi-Square	247	287.1920	1.1627
Scaled Pearson X2	247	287.1920	1.1627
Log Likelihood		9.8849	

Algorithm converged.

Analysis Of Parameter Estimates

Parameter	DF	Estimate	Standard Error	Wald 95% Confidence Limits		Chi-Square	Pr > ChiSq
Intercept	1	1.5453	0.1367	1.2774	1.8133	127.75	<.0001
dist	1	-0.0212	0.0015	-0.0242	-0.0182	191.10	<.0001
size	1	0.1358	0.0317	0.0736	0.1980	18.32	<.0001
Scale	0	1.0000	0.0000	1.0000	1.0000		

NOTE: The scale parameter was held fixed.