Appendix 1

Life table for Switzerland; 1978-1983

(source : Office fédéral de la statistique, Berne, 1985)

1 Age x	2 Probability of death $\hat{q}_x$	3 Survival probability $\hat{p}_x$	4 Death rate $\hat{\lambda}_x$	5 Survivor function ℓ_x	6 Number of deaths d_x	7 Expectation of life $\overset{\circ}{e}_x$	8 Age x
0	0.009487	0.990513		100000	949	72.40	0
1	0.000887	0.999113	0.003823	99051	88	72.09	1
2	0.000675	0.999325	0.000760	98963	66	71.16	2
3	0.000551	0.999449	0.000602	98897	55	70.20	3
4	0.000460	0.999540	0.000501	98842	45	69.24	4
5	0.000403	0.999597	0.000423	98797	40	68.27	5
6	0.000379	0.999621	0.000393	98757	38	67.30	6
7	0.000355	0.999645	0.000371	98719	35	66.33	7
8	0.000332	0.999668	0.000338	98684	32	65.35	8
9	0.000313	0.999687	0.000318	98652	31	64.37	9
10	0.000300	0.999700	0.000309	98621	30	63.39	10
11	0.000296	0.999704	0.000298	98591	29	62.41	11
12	0.000307	0.999693	0.000295	98562	30	61.43	12
13	0.000340	0.999660	0.000321	98532	34	60.45	13
14	0.000405	0.999595	0.000371	98498	40	59.47	14
15	0.000513	0.999487	0.000448	98458	50	58.49	15
16	0.000682	0.999318	0.000582	98408	67	57.52	16
17	0.000931	0.999069	0.000801	98341	92	56.56	17
18	0.001207	0.998793	0.001067	98249	118	55.61	18
19	0.001469	0.998531	0.001346	98131	145	54.68	19
20	0.001676	0.998324	0.001590	97986	164	53.76	20
21	0.001787	0.998213	0.001752	97822	175	52.85	21
22	0.001760	0.998240	0.001786	97647	171	51.94	22
23	0.001686	0.998314	0.001728	97476	165	51.03	23
24	0.001600	0.998400	0.001651	97311	156	50.12	24
25	0.001532	0.998468	0.001561	97155	148	49.20	25
26	0.001474	0.998526	0.001498	97007	143	48.27	26
27	0.001407	0.998593	0.001448	96864	137	47.34	27
28	0.001340	0.998660	0.001374	96727	129	46.41	28
29	0.001286	0.998714	0.001304	96598	124	45.47	29

1 Age x	2 Probability of death $\hat{q}_x$	3 Survival probability $\hat{p}_x$	4 Death rate $\hat{\lambda}_x$	5 Survivor function ℓ_x	6 Number of deaths d_x	7 Expectation of life $\mathring{e}_x$	8 Age x
30	0.001255	0.998745	0.001272	96474	122	44.53	30
31	0.001248	0.998752	0.001253	96352	120	43.58	31
32	0.001257	0.998743	0.001249	96232	121	42.64	32
33	0.001282	0.998718	0.001267	96111	123	41.69	33
34	0.001323	0.998677	0.001300	95988	127	40.74	34
35	0.001382	0.998618	0.001348	95861	132	39.80	35
36	0.001451	0.998549	0.001413	95729	139	38.85	36
37	0.001530	0.998478	0.001495	95590	147	37.91	37
38	0.001630	0.998370	0.001578	95443	155	36.97	38
39	0.001757	0.998243	0.001689	95288	168	36.02	39
40	0.001923	0.998077	0.001841	95120	183	35.09	40
41	0.002127	0.997873	0.002014	94937	201	34.15	41
42	0.002366	0.997634	0.002243	94736	225	33.23	42
43	0.002635	0.997365	0.002505	94511	249	32.30	43
44	0.002931	0.997069	0.002780	94262	276	31.39	44
45	0.003251	0.996749	0.003088	93986	305	30.48	45
46	0.003572	0.996428	0.003416	93681	335	29.58	46
47	0.003897	0.996103	0.003742	93346	364	28.68	47
48	0.004258	0.995742	0.004079	92982	396	27.79	48
49	0.004690	0.995310	0.004469	92586	434	26.91	49
50	0.005225	0.994775	0.004947	92152	481	26.03	50
51	0.005872	0.994128	0.005549	91671	539	25.17	51
52	0.006610	0.993390	0.006249	91132	602	24.31	52
53	0.007424	0.992576	0.007026	90530	672	23.47	53
54	0.008301	0.991699	0.007884	89858	746	22.64	54
55	0.009229	0.990771	0.008803	89112	823	21.83	55
56	0.010172	0.989828	0.009748	88289	898	21.03	56
57	0.011129	0.988861	0.010700	87391	973	20.24	57
58	0.012182	0.987818	0.011711	86418	1053	19.46	58
59	0.013356	0.986644	0.012826	85365	1140	18.69	59
60	0.014712	0.985288	0.014102	84225	1239	17.94	60
61	0.016236	0.983764	0.015566	82986	1347	17.20	61
62	0.017894	0.982106	0.017186	81639	1461	16.47	62
63	0.019706	0.980294	0.018953	80178	1580	15.77	63
64	0.021694	0.978306	0.020884	78598	1705	15.07	64
65	0.023879	0.976121	0.023028	76893	1837	14.40	65
66	0.026199	0.973801	0.025340	75056	1966	13.74	66
67	0.028641	0.971359	0.027767	73090	2093	13.09	67
68	0.031296	0.968704	0.030380	70997	2222	12.46	68
69	0.034256	0.965744	0.033267	68775	2356	11.85	69

1 Age x	2 Probability of death $\hat{q}_x$	3 Survival probability $\hat{p}_x$	4 Death rate $\hat{\lambda}_x$	5 Survivor function ℓ_x	6 Number of deaths d_x	7 Expectation of life $\mathring{e}_x$	8 Age x
70	0.037614	0.962386	0.036533	66419	2498	11.25	70
71	0.041271	0.958729	0.040190	63921	2638	10.67	71
72	0.045166	0.954834	0.044126	61283	2768	10.11	72
73	0.049446	0.950554	0.048389	58515	2894	9.56	73
74	0.054258	0.945742	0.053147	55621	3018	9.04	74
75	0.059752	0.940248	0.058576	52603	3143	8.53	75
76	0.065881	0.934119	0.064761	49460	3258	8.04	76
77	0.072548	0.927452	0.071618	46202	3352	7.57	77
78	0.079819	0.920181	0.079123	42850	3420	7.12	78
79	0.087762	0.912238	0.087383	39430	3461	6.69	79
80	0.096444	0.903556	0.096495	35969	3469	6.29	80
81	0.105845	0.894155	0.106497	32500	3440	5.91	81
82	0.115919	0.884081	0.117722	29060	3368	5.55	82
83	0.126306	0.873694	0.128832	25692	3245	5.21	83
84	0.137156	0.862844	0.140845	22447	3079	4.89	84
85	0.148622	0.851378	0.153833	19368	2879	4.59	85
86	0.160853	0.839147	0.167876	16489	2652	4.31	86
87	0.174001	0.825999	0.183060	13837	2408	4.03	87
88	0.187986	0.812014	0.199478	11429	2148	3.78	88
89	0.202840	0.797160	0.217228	9281	1883	3.54	89
90	0.218595	0.781405	0.236421	7398	1617	3.31	90
91	0.235280	0.764720	0.257173	5781	1360	3.10	91
92	0.252919	0.747081	0.279610	4421	1118	2.90	92
93	0.271533	0.728467	0.303869	3303	897	2.71	93
94	0.291138	0.708862	0.330099	2406	701	2.53	94
95	0.311742	0.688258	0.358459	1705	531	2.37	95
96	0.333347	0.666653	0.389123	1174	391	2.21	96
97	0.355943	0.644057	0.422278	783	279	2.06	97
98	0.379515	0.620485	0.458126	504	191	1.93	98
99	0.404031	0.595969	0.496885	313	127	1.80	99
100	0.429449	0.570551	0.538793	186	80	1.68	100
101	0.455714	0.544286	0.584104	106	48	1.58	101
102	0.482754	0.517246	0.633096	58	28	1.47	102
103	0.510481	0.489519	0.686067	30	15	1.37	103
104	0.538790	0.461210	0.743341	15	8	1.23	104
105	0.567559	0.432441	0.805267	7	4	1.07	105
106	0.596649	0.403351	0.872222	3	2	0.83	106
107	0.625905	0.374095	0.944617	1	1	0.50	107

Appendix 2

Using GLIM in descriptive epidemiology

The software GLIM[1] was specifically designed for fitting generalized linear models which are commonly used in the analysis of epidemiological data. It is therefore one of the most useful tools for carrying out epidemiological calculations

We should first recall the concept of the linear model. Suppose that Y is a normal variate with mean μ and variance σ^2 and that μ is linearly related to several covariates represented by the vector $\mathbf{z}$:

$$\mu = \boldsymbol{\beta}\mathbf{z} = \beta_1 z_1 + \ldots + \beta_p z_p$$

or

$$Y = \boldsymbol{\beta}\mathbf{z} + \varepsilon$$

where $\varepsilon \rightsquigarrow N(0, \sigma^2)$ is usually called the *error.* Suppose further that Y has been observed for several values of $\mathbf{z}$. The *response variable,* called also the *dependent variable* Y can therefore be represented by a vector of dimension n, the number of observations. Denoting Y_i as the ith observation corresponding to the value $\mathbf{z}_i$ of $\mathbf{z}$, the maximum likelihood method enables $\boldsymbol{\beta}$ to be estimated by minimizing the expression

$$L(\boldsymbol{\beta}) = \sum_{i=1}^{n} (Y_i - \boldsymbol{\beta}\mathbf{z}_i)^2$$

which is the negative of the log-likelihood.

GLIM provides estimates of the coordinates of $\boldsymbol{\beta}$, the variance-covariance of these estimates as well as fitted values ($\hat{Y}_i$) and residuals ($Y_i - \hat{Y}_i$).

GLIM is an interactive programme which can be run on either a personal computer or a mainframe and which is usually activated by simply typing GLIM on the keyboard. In order to introduce the reader to its use, the estimation of the parameters of the regression equation which was fitted to the data presented in table 3.3 (Chapter 3 page) is reproduced and commented upon below. Comments are framed and printed in italics; instructions given to the programme are printed in bold type, while output of the programme is printed in smaller character using the current typeface. As a rule an instruction is introduced by a '$' and remains activated until another $ character is input. With these conventions the dialogue between the computer and the user may be as follows :

GLIM

GLIM 3.77 update 2 (copyright)1985 Royal Statistical Society, London

[1] Generalised Linear Interactive Modelling, NAG Ltd, Wilkinson House, Jordan Hill Road, Oxford OX2 8DR, UK

After this welcome message the user is invited with a? to input a directive *(a $ followed by a word). Directives are either a statement or a request for an action to be carried out.Their name can be abbreviated provided that it is unambiguous.*

? **$unit 11**

? **$data z y**

These directives state that there are 11 observations (n = 11) and that there are two values per unit in the input data (z and y)

$DAT ? **$read**

The above directive requests data to be read from the keyboard. They should be input as described in the directive ***data****. The computer therefore expects 2 × 11 numbers as a series of z y pairs*

$REA ? **0 0.65 0.1 1.21 0.2 1.36 0.3 1.77 0.4 2.55 0.5 2.63 0.6 1.21**

$REA ? **0.7 3.22 0.8 2.50 0.9 4.23 1.0 2.34**

? **$look z y $**

This directive requests the output of the values of z and y. It is used here to check that the data have been input correctly.

	Z	Y
1	0.0000	0.6500
2	0.1000	1.2100
3	0.2000	1.3600
4	0.3000	1.7700
5	0.4000	2.5500
6	0.5000	2.6300
7	0.6000	1.2100
8	0.7000	3.2200
9	0.8000	2.5000
10	0.9000	4.2300
11	1.0000	2.3400

? **$yvar y $err n fit**

The directive ***yvar*** *enables the* dependant variable *to be specified, and* ***err*** *state that the* error ε *is normally distributed ('n' for normal). This complete the specification of the model. The estimation starts with the directive* ***fit*** *and initially produces the deviance and its d.f.*

deviance = 10.806

d.f. = 10

Since the argument of ***fit*** *is empty, the request is for the adjustment of a constant mean (***Y** $= \beta \times$ **1** $+ \varepsilon$ *where* **1** *is a vector with eleven coordinates equal to 1). In this case the estimate* $\hat{\beta}$ *of* β *is therefore* $\overline{Y} = \sum_{i=1}^{n} Y_i$ *and the value of the negative of the maximum log-likelihood* $L(\hat{\beta}) = \sum_{i=1}^{n} (Y_i - \overline{Y})^2$, *which is the minimum deviation from the observed data when fitted with this class of models, is called the* **deviance**; *when divided by* σ^2, *the deviance is distributed as* χ^2 *with* **d. f.** *degrees of freedom.*

? **$disp e**

In order to display the estimates (e) the directive **displ**ay *must be used. This latter directive automatically produces standard errors of estimates and the value of* $L(\hat{\beta})$ /df, *named the* ***scale parameter*** *which is here an estimate of* σ^2.

	estimate	s.e.	parameter
1	2.152	0.3134	1

scale parameter taken as 1.081

$DIS ? **$fit z $disp e**

deviance = 4.8900

d.f. = 9

	estimate	s.e	parameter
1	0.9923	0.4158	1
2	2.319	0.7028	Z

scale parameter taken as 0.5433

The directive ***fit z*** *requests the estimation of the linear model* $\mu = \beta_0 + \beta_1 z$. *Note that the constant* **1** *is always included in a model except if explicitly excluded (***fit z-1***). The estimates of the parameters of the regression equation are therefore* $\hat{\beta}_0 = 0.9923$, $\hat{\beta}_1 = 2.319$ and $\hat{\sigma}^2 = 4.89/9 = 0.5433$.

$DIS ? **r**

The letter **r** *is an argument of the directive* ***display***. *When typed while display is still activated, it requests that the fitted values* $(\hat{Y}_i)$ *and the standardized residuals* $(Y_i - \hat{Y}_i)/SE(Y_i)$ *be output*

unit	observed	fitted	residual
1	0.6500	0.9923	-0.342
2	1.2100	1.2242	-0.014
3	1.3600	1.4561	-0.096
4	1.7700	1.6880	0.082
5	2.5500	1.9199	0.630
6	2.6300	2.1518	0.478
7	1.2100	2.3837	-1.174
8	3.2200	2.6156	0.604
9	2.5000	2.8475	-0.348
10	4.2300	3.0795	1.151
11	2.3400	3.3114	-0.971

$DIS ? **$stop**

The generalized linear model differs from the above simple normal model in two respects, (i) the model now aims to describe a function of the mean and not the mean itself; (ii) the error is no longer distributed as a normal variate but belongs to a class of random variables which enable reliable estimation of linear models to be achieved. As pointed out often in this book, descriptive epidemiology collects data which are often distributed according to the Poisson distribution; in this context it is the logarithm of the mean which is modelled and the name *'Poisson regression'* is now commonly used to designate genralized linear modelling using Poisson distributed error and logarithmic transformation of the mean. We shall illustrate the principle of this method and its implementation in GLIM with the data of table 2.8, adjusting the model of equation 2.33 (see page...).

Suppose that the data are stored in a file named MHP.DAT and organized as shown below, where each line corresponds to a computer record :

8 8 36 54 53 96 115 145

10 6 7 18 17 25 35 37

22801 27291 26762 25899 19853 17431 15024 11961

13506 12480 11012 9887 7010 6845 6066 4492

A possible way of fitting the multiplicative model (2.33) to the above data is given below. Comments briefly introduce the directives.

GLIM

GLIM 3.77 update 2 (copyright)1985 Royal Statistical Society, London

Eight observations from each of two cancer registries.

? **$unit 16**

Read the number of cases (k) in the file MHP.DAT *which will be connected to the reading unit 1 after answering the file name request; then read the number of person-years (m). Note that* ***dinput*** *is used instead of* ***read*** *when reading from a file.*

? **$data k $dinput 1**

File name ? **mhp.dat**

$DIN ? **$data m $dinp 1 $**

Create the variables AGE *and* REG *(for registry) using the function* ***%gl*** *: this function creates a vector with values given by the first argument (here 8 and 2 because 1≤ age ≤ 8 and 1 ≤ reg ≤ 2). The second argument gives the number of repetitions of each value. Note that, if not specified otherwise, the dimension of a vector equals the number of units. The character ':' enables the activated directive to be repeated with other arguments (here the* ***calculate*** *directive).*

? **$cal age=%gl(8,1) :reg=%gl(2,8) $loo reg age k m** $

	REG	AGE	K	M
1	1.000	1.000	8.000	22801.
2	1.000	2.000	8.000	27291.
3	1.000	3.000	36.000	26762.
4	1.000	4.000	54.000	25899.
5	1.000	5.000	53.000	19853.
6	1.000	6.000	96.000	17431.
7	1.000	7.000	115.000	15024.
8	1.000	8.000	145.000	11961.
9	2.000	1.000	10.000	13506.
10	2.000	2.000	6.000	12480.
11	2.000	3.000	7.000	11012.
12	2.000	4.000	18.000	9887.
13	2.000	5.000	17.000	7010.
14	2.000	6.000	25.000	6845.
15	2.000	7.000	35.000	6066.
16	2.000	8.000	37.000	4492.

State, using ***yvar****, that the response variable is the number of cases (k), state, using* ***err****, that the error distribution is the Poisson distribution and, using* ***offset,*** *that the origin of the response variable scale is shifted by log(m) (i.e., the mean* μ *is such that* $\log(\mu) = \text{zero} + \beta z$).

? **$yvar k $err p $cal zero=%log(m) $offset zero**

State that AGE *and* REG *are categorical variables (factors). This directive requests the computer to create dummy variables for each level of the factors but one (i.e., 7 for* AGE *and 1 for* REG).

? **$factor age 8 :reg 2 $**

The successive fits enable the contribution of each factor to be assessed. Remember that the change in deviance is distributed as a χ^2 *with df equal to the corresponding change in degrees of freedom.*

? **$fit :+ age :+ reg $**

scaled deviance = 676.59 at cycle 4

d.f. = 15

```
scaled deviance = 18.142 (change = -658.4) at cycle 3
     d.f. = 8       (change = -7)
scaled deviance = 9.3920 (change = -8.750) at cycle 4
     d.f. = 7       (change = -1)
```

? **$disp e r $**

	estimate	s.e.	parameter
1	-7.519	0.2374	1
2	-0.3572	0.3564	AGE(2)
3	0.8109	0.2808	AGE(3)
4	1.377	0.2636	AGE(4)
5	1.631	0.2644	AGE(5)
6	2.285	0.2527	AGE(6)
7	2.642	0.2495	AGE(7)
8	3.080	0.2472	AGE(8)
9	-0.2651	0.09168	REG(2)

scale parameter taken as 1.000

unit	observed	fitted	residual
1	8	12.376	-1.244
2	8	10.364	-0.734
3	36	32.683	0.580
4	54	55.690	-0.227
5	53	55.080	-0.280
6	96	92.987	0.312
7	115	114.526	0.044
8	145	141.293	0.312
9	10	5.624	1.845
10	6	3.636	1.240
11	7	10.317	-1.033
12	18	16.310	0.419
13	17	14.920	0.539
14	25	28.013	-0.569
15	35	35.474	-0.080
16	37	40.707	-0.581

The programme provides several statistics, values of which can be requested through the look directive. For example to get the classical goodness of fit χ^2 type :

? **$loo %x2 $**

9.892

Note that most statistics can also be calculated directly using system-built vectors storing the main results of the fit. For example the above χ^2 is obtained through ***%fv*** *which stores the fitted values.*

? **$cal (k-%fv)**2/%fv $**

1.547
0.5393
0.3366
0.05131
0.07855
0.09761
0.001958
0.09728
3.405
1.537
1.066
0.1752
0.2900
0.3240
0.006321
0.3376

Note that the above values are the squares of the standardized residuals, $\frac{(Y_i - \hat{Y}_i)}{\sqrt{Y_i}}$ listed previously. The residuals could also be stored in a vector :

? $cal r2=(k-%fv)2/%fv $**

The directive **table** is a powerful tabulation programme. It is used here in its simplest form to get the total of **r2** coordinates, that is the value of the goodness of fit χ^2.

? **$tab the r2 t $**

9.892

? **$stop**

Before going to the next example of Poisson regression, we should remember that the estimates of the coordinates of β are the logarithms of the estimated relative rates; for example the relative rate of age-group 8 (70-74 years) compared with age-group 1 (35-39 years) is exp(3.08)=21.76. The incidence rate of this latter age-group is estimated as exp(–7.519)/5=10.85/100 000 (the estimated rate is divided by five because we input the populations instead of the person-years; see table 2.8); similarly the relative rate of Geneva (REG(2)) compared with Zaragoza is exp(–0.2651) = 0.767.

The method of Poisson regression is now applied to the data of table 2.13 (page), stored as previously in a file named HOMINCG.DAT. Only the data corresponding to age greater than 20 were used, since no case was observed before that age.

GLIM

GLIM 3.77 update 2 (copyright) 1985 Royal Statistical Society, London

Thirteen age-groups and six cantons of Côte d'Or make 78 units of observation.

? **$unit 78**

? **$data k $dinput 1**

File name? **homincg.dat**

$DIN? **$data m $dinput 1**

Create the variable AGE *and* PLACE *(for canton). The first two arguments of the directive* ***look*** *select the output interval for the vectors looked at, 1 to 12 in the present example. Note that GLIM 3.77 retains only four meaningful letters to identify a variable (plac for place)*

$DIN? **$cal age=%gl(13,6) :place=%gl(6,1) $loo 1 12 place age k m $**

	PLAC	AGE	K	M
1	1.000	1.000	0.000	52794.0
2	2.000	1.000	0.000	10073.0
3	3.000	1.000	0.000	8402.0
4	4.000	1.000	0.000	19034.0
5	5.000	1.000	1.000	9539.0
6	6.000	1.000	0.000	1186.0
7	1.000	2.000	0.000	54321.0
8	2.000	2.000	0.000	10499.0
9	3.000	2.000	0.000	7984.0
10	4.000	2.000	1.000	19009.0
11	5.000	2.000	0.000	7936.0
12	6.000	2.000	0.000	1345.0

Specify the model and the factors to be used in the fit; then fit the multiplicative model.

? **$yvar k $err p $cal zero=%log(m) $offset zero**

? **$factor age 13 plac 6**

$FAC? **$fit age+plac $disp e $**

scaled deviance = 68.198 at cycle 8

d.f. = 60

	estimate	s.e	parameter
1	-11.37	1.001	1
2	-0.006026	1.414	AGE(2)
3	-6.130	14.39	AGE(3)
4	1.806	1.118	AGE(4)
5	1.515	1.155	AGE(5)
6	2.730	1.049	AGE(6)
7	3.141	1.035	AGE(7)
8	3.901	1.021	AGE(8)
9	3.873	1.029	AGE(9)

10	4.491	1.014	AGE(10)
11	5.427	1.007	AGE(11)
12	5.466	1.010	AGE(12)
13	5.516	1.013	AGE(13)
14	-0.3498	0.2127	PLAC(2)
15	-0.5899	0.2198	PLAC(3)
16	-0.2367	0.1560	PLAC(4)
17	-0.3564	0.1902	PLAC(5)
18	-0.8039	0.4566	PLAC(6)

scale parameter taken as 1.000

Note that the incidence rate estimate in age-group 3 is almost zero and has a very large standard error. Actually no case has been observed in this age-group and the incidence rate estimate should be zero. The next step is to assess the significance of the factor PLAC; *to this end a model containing only the factor age is fitted to the data and the corresponding increase in deviance evaluated*

?

? **$fit –plac $**

scaled deviance = 80.781 (change = +12.58) at cycle 8

d.f. = 65 (change = +5)

This calculation confirms that the incidence differs in the various cantons of Côte d'Or. It is then possible to test whether this difference is mainly between the town of Dijon and the other cantons : a dummy variable is created which takes on the value 0 for Dijon and 1 for the other cantons; the best way to do this is to use the logical functions which are available in GLIM.

? **$cal other=(plac > 1) $fit age+other $disp e $**

scaled deviance = 71.663 at cycle 8

d.f. = 64

estimate s.e. parameter

	estimate	s.e	parameter
1	-11.36	1.001	1
2	-0.005996	1.414	AGE(2)
3	-6.097	14.18	AGE(3)
4	1.809	1.118	AGE(4)
5	1.515	1.155	AGE(5)
6	2.729	1.049	AGE(6)
7	3.136	1.035	AGE(7)
8	3.898	1.021	AGE(8)
9	3.867	1.029	AGE(9)
10	4.481	1.014	AGE(10)
11	5.417	1.007	AGE(11)
12	5.456	1.010	AGE(12)
13	5.508	1.013	AGE(13)
14	0.3700	-0.1214	OTHE

scale parameter taken as 1.000

The deviance is not increased significantly (71.66 – 68.20=3.46 *for* 4 *degrees of freedom*); *this observation leads us to accept the homogeneity of incidence in the cantons other than Dijon. The relative rate for these regions compared with Dijon is estimated as* exp(–0.37)=0.69; *a confidence interval may be obtained as* exp(–0.37±1.96*0.1214).
The relationship between age and incidence rate can be modelled with a polynomial in order to describe the data with a more parsimonious model. The variable age is first centred, then the polynomial degree to be used is roughly evaluated.

$cal x=age-6$

? **$cal x2=x*x :x3=x2*x :x4=x3*x $fit x+othe :+x :+x2 :+x3 :+x4 $**

scaled deviance = 102.96 at cycle 4
d.f. = 75
scaled deviance = 102.96 (change = 0.00) at cycle 4
d.f. = 75 (change = 0)
scaled deviance = 91.70 (change = –11.261) at cycle 5
d.f. = 74 (change = –1)
scaled deviance = 88.98 (change = –2.73) at cycle 5
d.f. = 73 (change = –1)
scaled deviance = 87.58 (change = –1.40) at cycle 5
d.f. = 72 (change = –1)

A third degree polynomial provides an acceptable model.....

? **$fit –x4 $disp e $**

scaled deviance = 88.98 (change = +1.40) at cycle 5
d.f. = 73 (change = +1)

	estimate	s.e.	parameter
1	-8.916	0.1818	1
2	0.6368	0.05715	X
3	-0.3677	0.1213	OTHE
4	0.001819	0.01794	X2
5	-0.004087	0.002379	X3

scale parameter taken as 1.000

.....which provides practically the same estimate of the relative rate as that obtained when age was modelled as a factor.

? **$stop**

We shall consider as a last example the data from table 3.15 giving the trends in mortality from lung cancer among young adults in France Scotland and the USA. These data were stored in the computer as a file (TREND.DAT) with 18 records

each containing two numbers, the number of cases (k) and the person-years in thousands (m). The records are sorted by country (USA, Scotland, France) and by time of death (1955 to 1984 by 6 groups of five years). The calculations which have been described on page... and in table 3.16 are reproduced below in detail.

GLIM

GLIM 3.77 update 2 (copyright)1985 Royal Statistical Society, London

Fitting a model for the USA.

? **$unit 6**

$data k m $dinput 1 $

File name? **trend.dat**

Create the variable time period (t).

? **$cal t=%gl(6,1) $loo t k m $**

	T	K	M
1	1.000	3762.	27599.
2	2.000	4900.	29249.
3	3.000	6147.	29859.
4	4.000	6318.	28416.
5	5.000	5638.	27590.
6	6.000	5413.	30569

Specify a model for a Poisson regression

? **$yvar k $err p $cal zero=%log(m) $offset zero $**

? **$cal t2=t*t $fit t+t2 $disp e $**

scaled deviance = 14.094 at cycle 3

d.f. = 3

	estimate	s.e.	parameter
1	-2.393	0.02757	1
2	0.4284	0.01696	T
3	-0.05299	0.002310	T2

scale parameter taken as 1.000

Wald's test based on the standard error of the T2 coefficient (–0.05299/0.002310=–22.9) *shows that the quadratic term is strongly significant. The same evaluation can be made using the likelihood ratio test :*

? **$fit –t2 $disp e $**

scaled deviance = 553.06 (change = +539.0) at cycle 3

d.f. = 4 (change = +1)

	estimate	s.e.	parameter
1	–1.857	0.01322	1
2	0.04836	0.003269	T

scale parameter taken as 1.000

We shall now fit the same model expression using a different error distribution in considering that the logarithm of the incidence rate is normally distributed with a mean equal to the proposed expression and a variance proportional to the observed number of cases in each unit.

? **$cal y=%log(k/m) $yvar y $err n $cal w=k $weight w**

– model changed

*Do not forget to set the origin back to zero. After having done so (**offset**), fit the quadratic model with the method of weighted least sqares.*

? **$offset $fit t+t2 $disp e $**

deviance = 14.042

d.f. = 3

	estimate	s.e.	parameter
1	-2.393	0.05916	1
2	0.4283	0.03639	T
3	-0.05299	0.004965	T2

scale parameter taken as 4.681

Fit the linear model by the same method.

? **$fit –t2 $**

deviance = 547.19 (change = +533.1)

d.f. = 4 (change = +1)

? **$disp e $**

	estimate	s.e.	parameter
1	–1.848	0.1622	1
2	0.04825	0.04053	T

scale parameter taken as 136.8

Note the value of the standard error of the T coefficient, obtained when fitting this model by least squares and compare it with the same coefficient in the linear model fitted by Poisson regression.

Fitting a model for Scotland.

? **$data k m $dinput 1 $**

? **$loo k m $**

	K	M
1	242.0	811.0
2	222.0	803.2
3	247.0	798.8
4	195.0	747.4
5	138.0	717.6
6	116.0	724.

The method of least squares will be applied first, since it is the model which is activated.

? **$cal y=%log(k/m) $**

– change to data affects model

Do not forget to change the values stored in w before fitting....

? **$cal w=k $fit t+t2 $**

– change to data affects model

deviance = 4.0117

d.f. = 3

? **$disp e $**

	estimate	s.e.	parameter
1	–1.359	0.1395	1
2	0.1637	0.09516	T
3	–0.04125	0.01391	T2

scale parameter taken as 1.337

Note that the scale parameter is close to one and that the coefficient of the quadratic term is strongly significant, as confirmed by the calculation reported below, which is based on the Poisson distribution :

? **$err p $yvar k $cal zero=%log(m) $offset zero $weight**

– model changed

$WEI ?

This last directive eliminates the weighting, which is irrelevant in the Poisson regression.

? **$fit t+t2 $**

scaled deviance = 4.0565 at cycle 3

d.f. = 3

? **$disp e $**

	estimate	s.e.	parameter
1	–1.359	0.1223	1
2	0.1622	0.08373	T
3	–0.04101	0.01224	T2

scale parameter taken as 1.000

? **$fit –t2 $**

scaled deviance = 15.495 (change = +11.44) at cycle 3
d.f. = 4 (change = +1)

? **$disp e $**

	estimate	s.e.	parameter
1	–1.012	0.06175	1
2	–0.1124	0.01754	T

scale parameter taken as 1.000

Fitting a model for France.

? **$data k m $dinput 1**

$DIN ? **$loo k m $**

	K	M
1	479.0	5878.
2	612.0	6586.
3	968.0	8333.
4	1265.0	8507.
5	1308.0	8032.
6	1349.0	7484.

– change to data affects model

? **$cal zero=%log(m) $fit t $disp e $**

scaled deviance = 14.875 at cycle 3
d.f. = 4

	estimate	s.e.	parameter
1	–2.641	0.03567	1
2	0.1627	0.008182	T

scale parameter taken as 1.000

? **$fit +t2 $**

scaled deviance = 6.4909 (change = –8.384) at cycle 3
d.f. = 3 (change = –1)

? **$disp e $**

	estimate	s.e.	parameter
1	–2.823	0.07334	1
2	0.2818	0.04219	T
3	–0.01589	0.005513	T2

scale parameter taken as 1.000

Now apply the least squares method :

? **$cal y=%log(k/m) $err n $offsset $cal w=k $weight w**

– – model changed

? **$yvar y $fit t $disp e $**

– – model changed

deviance = 15.015

d.f. = 4

	estimate	s.e.	parameter
1	–2.640	0.07020	1
2	0.1627	0.01614	T

scale parameter taken as 3.754

? **$fit +t2 $**

deviance = 6.5056 (change = –8.509)

d.f. = 3 (change = –1)

? **$disp e $**

	estimate	s.e.	parameter
1	–2.822	0.1063	1
2	0.2816	0.06125	T
3	–0.01591	0.008032	T2

scale parameter taken as 2.169

Note that the reduction in deviance is identical for the two error models; However, the standard error of the coefficient of the quadratic term is greater in this second situation where the lack of fit is taken into consideration in the estimation of σ^2.

? **$stop**

This brief description of the capabilities of GLIM for carrying out calculation in descriptive epidemiology may be supplemented by references [36] and [37] of Chapter 2. A new release of this software is now available and details can be found in :

Francis B J, Green M and Payne C P (eds) *The GLIM System : Release 4 Manual*, Oxford University Press, Oxford.

Subject index