Poisson regression

Contents

1. Introduction 1

2. Poisson-log model 2

1. Introduction

In this example we examine the relationship between smoking and coronary heart disease. Data are from a study by Sir Richard Doll, and we base our analysis on a table of the number of deaths from coronary heart disease for smokers and non-smokers from a 1951 study. Respondents were British medical doctors. We use the table below as point of departure, as originally published by Dobson and Barnett (2008, Table 9.1).

	Smokers		Non-smokers	
Age group	Deaths	Person-years	Deaths	Person-years
35-44	32	52407	2	18790
45-54	104	43238	12	10673
55-64	206	28612	28	5710
65-74	186	12663	28	2585
75-84	102	5317	31	1462

These data are given in the LISREL file **corheart2.lsf**. Data and syntax files can be found in the **MVABOOK\Chapter3** folder.

corheart2.LSF

	SMOKE	AGECAT	DEATHS	PYEARS	DTHRATE	AGESQ	SMKAGE	LNPYEARS
1	1.00	1.00	32.00	52407.00	0.00	1.00	1.00	10.87
2	1.00	2.00	104.00	43248.00	0.00	4.00	2.00	10.67
3	1.00	3.00	206.00	28612.00	0.01	9.00	3.00	10.26
4	1.00	4.00	186.00	12633.00	0.01	16.00	4.00	9.44
5	1.00	5.00	102.00	5317.00	0.02	25.00	5.00	8.58
6	0.00	1.00	2.00	18790.00	0.00	1.00	0.00	9.84
7	0.00	2.00	12.00	10673.00	0.00	4.00	0.00	9.28
8	0.00	3.00	28.00	5710.00	0.00	9.00	0.00	8.65
9	0.00	4.00	28.00	2585.00	0.01	16.00	0.00	7.86
10	0.00	5.00	31.00	1462.00	0.02	25.00	0.00	7.29

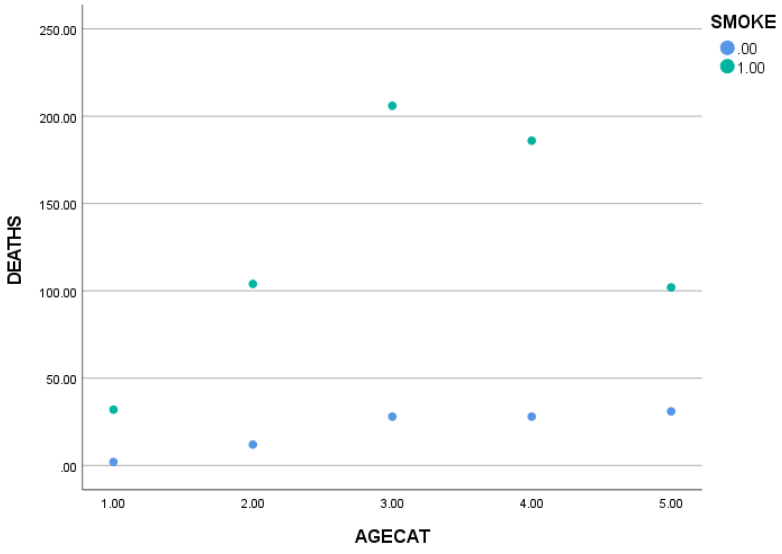
The variables are:

- SMOKE: An indicator variable that assumes the value 1 for smokers, and 0 for non-smokers
- AGECAT: An ordinal variable representing the age groups in the original table, assuming values between 1 (age group 35-44) and 5 (age group 75-84).
- DEATHS: The number of deaths recorded for the cell formed by the cross-tabulation of SMOKE and AGECAT.
- PYEARS: The number of person-years of observation
- DTHRATE: The ratio between the number of DEATHS and PYEARS
- AGESQ: The squared value of the variable AGECAT
- SMKAGE: An interaction term between SMOKE and AGECAT, calculated as the product of these two variables
- LNPYEARS: The natural logarithm of PYEARS.

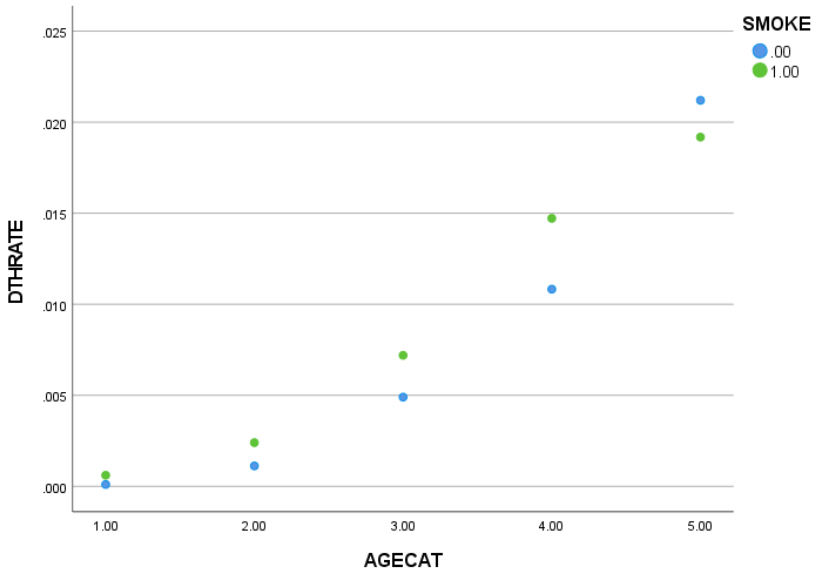
2. Poisson-log model

The outcome variable of interest here is the number of deaths. A Poisson distribution is assumed, as an appropriate distribution for a rare event (in this case dying due to smoking induced coronary heart disease).

A scatterplot of the data by categories of SMOKE is shown below.



In order to take the number of person-years into account, we remake this graph using the DTHRATE as outcome instead of the number of deaths.

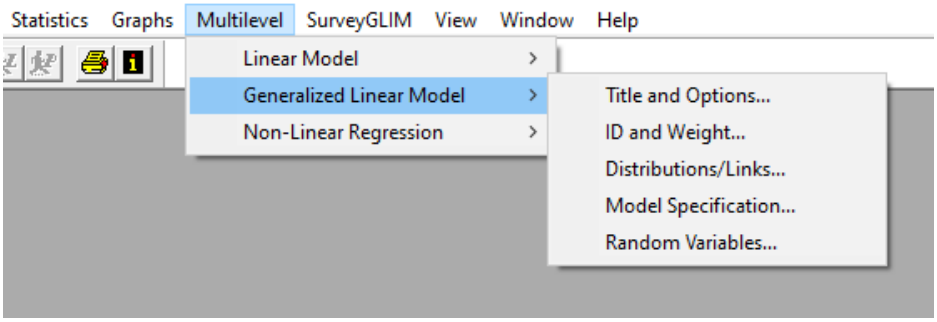


We note that there seems to be a different pattern for smokers than for non-smokers. There seems to be some interaction between the age and smoking status, and the trend in the second plot indicates that the relationship may not be linear.

We thus opt to use the model

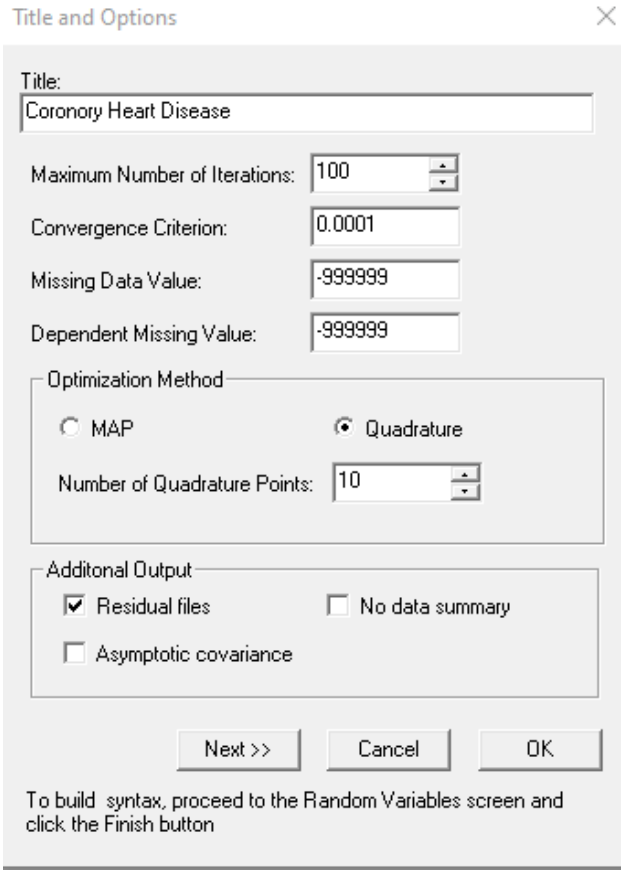
$$\ln(\text{DEATHS}) = \ln(\text{PYEARS}) + \alpha + \gamma_1(\text{SMOKE}) + \gamma_2(\text{AGECAT}) + \gamma_3(\text{AGESQ}) + \gamma_4(\text{SMKAGE})$$

and we fit it to the data using the GLIM (Generalized LInear Model) module in LISREL to do so. This module is accessed via the **Multilevel** option on the main toolbar.



When the options on this menu is compared to those for the **Linear Model** option, we note the additional **Distributions/Links** option. This is because a generalized linear model not only includes the response variable and a linear part consisting of the explanatory variable(s), but also a link function which transforms the mean of the response variable to linear form.

On the **Titles and Options** dialog box, we enter a title and check to request residual files in the **Additional Output** section.



Proceed to the **Distributions and Links** tab by using the **Next** button. Here we select Poisson as the **distribution type** and leave the **link function** field at its default (Log) value for a Poisson model.

On the **Dependent and Independent Variables** tab DEATHS is selected as outcome, while SMOKE and AGECAT are added as predictors. We also add the squared value of age (AGESQ) and the interaction term between SMOKE and AGECAT as predictors. The **Offset variable** field is set to LNPYEARS. The variable LNPYEARS is thus used as covariate whose coefficient is fixed equal to 1.

Click next to move to the final tab and then **Finish** to generate the syntax file shown below.

corheart2.PRL

```
MGlimOptions Converge=0.0001 MaxIter=100 MissingCode=-999999
             Method=Quad NQUADPTS=10 Output=Residuals ;
Title=Coronory Heart Disease;
SY='C:\gerhard vbe\MVABOOK\CHAPTER1\corheart2.LSF';
DEPENDENT_MISS=-999999;
Distribution=POI;
Link=LOG;
Intercept=Yes;
Scale=None;
DepVar=DEATHS;
CoVars=SMOKE AGECAT AGESQ SMKAGE;
Offset=LNPYEARS;
```

The first part of the output file indicates that there is no hierarchical structure defined (an option available for this kind of analysis). This is followed by basic descriptive statistics for all variables.

```
 ------------------------------------------------------
    WARNING: NO HIERARCHICAL STRUCTURE DEFINED:
    CHECK THE ID3= OR ID2=  PARAGRAPHS
  FOR THIS ANALYSIS A SIMPLE RANDOM SAMPLE IS ASSUMED
 ------------------------------------------------------

o=======================o
| Coronary Heart Disease |
|                        |
o=======================o
```

Model and Data Descriptions

Sampling Distribution	= Poisson
Link Function	= Log
Number of Level-2 Units	1
Number of Level-1 Units	10
Number of Level-1 Units per Level-2 Unit	= 10

Variable	Minimum	Maximum	Mean	Standard Deviation
DEATHS	2.0000	206.0000	73.1000	73.4218
intcept	1.0000	1.0000	1.0000	0.0000
SMOKE	0.0000	1.0000	0.5000	0.5270
AGECAT	1.0000	5.0000	3.0000	1.4907
AGESQ	1.0000	25.0000	11.0000	9.1165
SMKAGE	0.0000	5.0000	1.5000	1.9003

This is followed by goodness-of-fit statistics that indicate that the model describes the data well.

Goodness of fit statistics

Statistic	Value	DF	Ratio
Likelihood Ratio Chi-square	3.2585	5	0.6517
Pearson Chi-square	3.1900	5	0.6380
Log Likelihood	2726.8318		
Akaike Information Criterion	-5443.6635		
Schwarz Criterion	-5442.1506		

The estimated regression weights are given next.

Estimated regression weights

Parameter	Estimate	Standard Error	z Value	P Value
intcept	-3.6027	0.4770	-7.5524	0.0000
SMOKE	2.8786	0.3961	7.2678	0.0000
AGECAT	2.4486	0.2130	11.4965	0.0000
AGESQ	-0.2933	0.0274	-10.7133	0.0000
SMKAGE	-0.3426	0.1035	-3.3105	0.0009

Event Rate Ratio and 95% Event Rate Confidence Intervals

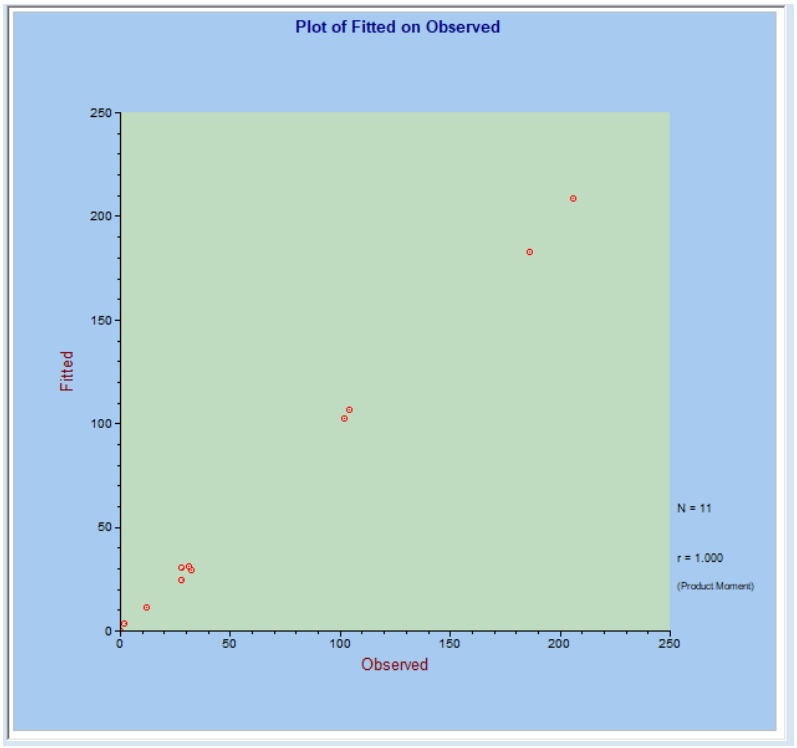
Parameter	Estimate	Event Rate	Bounds Lower	Bounds Upper
intcept	-3.6027	0.0272	0.0107	0.0694
SMOKE	2.8786	17.7893	8.1849	38.6640
AGECAT	2.4486	11.5723	7.6229	17.5678
AGESQ	-0.2933	0.7458	0.7068	0.7869
SMKAGE	-0.3426	0.7099	0.5795	0.8696

All the regression weights are statistically significant. The highest event rate is associated with the variable SMOKE. The risk of death from coronary heart disease is higher for smokers than for non-smokers when age is taken into account, but it also seems from the results that this relationship attenuates with increased age.

The residuals requested for this analysis are written to a new LSF file **corhart_res.lsf**.

corheart_RES.LSF

	Observed	Fitted	Raw	Pearson	Deviance	Likelihd	SPearson	SDevianc
1	32.00	29.56	2.44	0.45	0.44	0.71	0.71	0.70
2	104.00	106.82	-2.82	-0.27	-0.27	-0.37	-0.37	-0.37
3	206.00	208.36	-2.36	-0.16	-0.16	-0.25	-0.25	-0.25
4	186.00	182.59	3.41	0.25	0.25	0.34	0.34	0.34
5	102.00	102.67	-0.67	-0.07	-0.07	-0.15	-0.15	-0.15
6	2.00	3.41	-1.41	-0.76	-0.83	-0.96	-0.91	-0.98
7	12.00	11.54	0.46	0.14	0.13	0.17	0.17	0.17
8	28.00	24.75	3.25	0.65	0.64	0.82	0.83	0.81
9	28.00	30.24	-2.24	-0.41	-0.41	-0.50	-0.50	-0.51
10	31.00	31.06	-0.06	-0.01	-0.01	-0.02	-0.02	-0.02
11	0.00	0.00	0.00	1.57	1.65	2.71	2.63	2.74



A plot of the fitted and observed show a straight line, indicating a good fit. The last line of the data file contains residual sum of squares, which indicate a good fit.