Applications of Generalized Linear Models

I. Introduction

In lecture 5 we have introduced generalized linear models (GLMs). In this lecture we will focus on some common applications of GLMs for different forms and scales of the response and explanatory variables. In section II we consider methods for analysing binary response data. We will focus on the most common method to analyse binary response data, namely *logistic regression*. Logistic regression is used to model relationships between the response variable and several explanatory variables, which may be discrete or continuous. Section III discusses generalizations of logistic regression for dichotomous responses to model nominal or ordinal responses with more than two categories. The *nominal logistic regression* model will be introduced for situations where there is no natural order among the response categories and the *ordinal logistic regression model* will be introduced for situations where the response categories are ordinal. Section IV focuses on the analysis of count data. Here we are interested in the number of times an event occurs. We will distinguish between two different situations. In the first situation, the events relate to varying amounts of 'exposure', which need to be taken into account when modelling the rate of events. *Poisson regression* is used in this case. The other explanatory variables (in addition to 'exposure') may be continuous or categorical. The counts may for example be the number of traffic accidents, which need to be analysed in relation to some exposure variable such as the number of registered motor vehicles. In the second situation, 'exposure' is constant (and therefore not relevant to the model) and the explanatory variables are usually categorical. If there are only a few explanatory variables the data are summarized in a cross-classified table. The response is the frequency or count in each cell of the table. The variables used to define the table are all treated as explanatory variables. The study design may mean that there are certain constraints on the cell frequencies (for example, the totals of the frequencies in each row of the table may be equal) and this need to be taken into account in the modelling. *Log-linear models* are appropriate for this situation. We will consider an example relating to each of the above situations. Finally, we will give a short overview of further applications of generalized linear models.

II. Binary data and logistic regression

Introduction

In this section we consider the situation where the response, Y, can only take one of two possible values. In practice, these responses may be alive or dead, or present or absent. This situation also arises frequently in medical trials, where at the end of the trial period, the patient has either recovered or has not. It is convenient to denote the two levels by 0 and 1 and to refer to the categories as a "failure" or a "success". The main objective is to investigate the relationship between the response probability $\pi = \pi(\underline{x})$ and the explanatory variables $x_1, x_2, \ldots, x_n$. A binary random variable can be defined as

$$Z = \begin{cases} 1 \text{ if the outcome is a success} \\ 0 \text{ if the outcome is a failure} \end{cases}$$

with probabilities $P(Z=1) = \pi$ and $P(Z=0) = 1-\pi$. Typically, there are a number of explanatory variables or covariates $x_1, x_2, \ldots, x_n$ associated with each observation in the study. For example, in a designed experiment the covariates may consist of indicator variables associated with blocking or related to treatment factors, together with other quantitative variables relating to the response. In an observational study the covariates may mainly consist of measured variables relating to the probability of a positive response.

It is important to distinguish between grouped and ungrouped data when we are dealing with binary data. Suppose that we have *N* independent random variables, $Y_1, Y_2, \ldots, Y_N$, corresponding to the numbers of successes in *N* different subgroups. It is assumed that each random variable is distributed $Y_i \sim binomial(n_i, \pi_i)$. Thus, we are dealing with *N* independent binomial distributions. In other words, of the $n_{total} = n_1 + n_2 + \cdots + n_N$ observations in the study, n_i share the covariate values $x_{i1}, x_{i2}, \ldots, x_{ip}$. These observations are said to form a *covariate class*. This situation is illustrated below:

	Subgroups			
	1	2	…	N
Successes	Y_1	Y_2	…	Y_N
Failures	$n_1 - Y_1$	$n_2 - Y_2$	…	$n_N - Y_N$
Totals	n_1	n_2	…	n_N

When binary data are grouped by covariate class, the responses have the form $y_1/n_1,\ldots,y_N/n_N$, where $0 \le y_i \le n_i$ is the number of successes out of the n_i subjects in the *i*th covariate class. Ungrouped data, or data listed by individual subjects, can be considered as a special case for which $n_1 = \ldots = n_N$.

Our goal is to describe the proportion of success, $P_i = Y_i/n_i$, in each subgroup in terms of factor levels and other explanatory variables. From the fact that $Y_i \sim binomial(n_i, \pi_i)$, it follows that $E(Y_i) = n_i\pi_i$ and $E(P_i) = \pi_i$. So within the generalized linear model set-up we can model the probabilities π_i as

$$g(\pi_i) = \beta_{i0} + \beta_{i1}x_{i1} + \ldots + \beta_{ip}x_{ip}$$

where $x_1,\ldots,x_p$ are the observed explanatory variables, $\beta_{i1},\ldots,\beta_{ip}$ are the parameters and g is a link function. The explanatory variables can be dummy variables for factors and measured values for covariates. A simple method to model the probability of 'success' would be the linear model

$$\pi = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p .$$

The disadvantage of this approach is that π is a probability and is therefore restricted to the interval $[0,1]$, while the fitted values are not restricted to this interval. The next subsection will introduce the general logistic regression model that uses link function, *g*, that is different from the identity link and which overcomes the above problem.

Logistic regression model

The general linear logistic regression model is defined as

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \text{logit}(\pi_i) = \beta_{i0} + \beta_{i1}x_{i1} + \cdots + \beta_{ip}x_{ip}$$

where $x_{i1},\ldots,x_{ip}$ are continuous measurements corresponding covariates and/or dummy variables corresponding to factor levels and $\beta_{i1},\ldots,\beta_{ip}$ are the parameters. This model is very widely used for analysing data involving binary or binomial responses and several explanatory variables. Estimates for the parameters and response probabilities are typically obtained by the method of maximum likelihood. These estimates will be computed and returned by your software package. Note that the general linear logistic regression model can be rewritten in terms of the probability of a positive response

$$\pi_i = \frac{\exp\left(\beta_{i0} + \beta_{i1}x_{i1} + \ldots + \beta_{ip}x_{ip}\right)}{1 + \exp\left(\beta_{i0} + \beta_{i1}x_{i1} + \ldots + \beta_{ip}x_{ip}\right)}.$$

Goodness of fit

To evaluate the goodness of fit we can calculate the deviance defined in lecture 5. There are no nuisance parameters (like σ^2 for the Normal distribution), so goodness of fit can be assessed and the hypotheses can be tested directly using the approximation

$$D \sim \chi^2(N - p)$$

under the hypothesis that the model is correct, where p is the number of parameters estimated and N the number of covariate classes. It must be noted that the above approximation can be poor if the number of observations in the respective covariate classes are small.

Another goodness of fit statistic that is used is the Pearson chi-square statistic that can also be tested using the approximation

$$X^2 \sim \chi^2(N - p).$$

X^2 is asymptotically equivalent to the deviance. The choice between D and X^2 depends on the adequacy of the approximation to the $\chi^2(N - p)$ distribution.

Model Diagnostics

There are two main forms of residuals used to evaluate the model adequacy: the Pearson residual and the deviance residual. There are also standardized versions of each of these types of residuals. We defined these residuals in lecture 5 and also discussed the various residual plots that may be used to evaluate the model, such as plots of the residuals against the explanatory variables. Sometimes there may be very few distinct values in each group or covariate class that may lead to uninformative plots. In cases like these we may have to rely on the deviance, Pearson chi-square or other statistics.

Over-dispersion

Over-dispersion is another important issue that must be given consideration when assessing the adequacy of models for binary or binomial data. Over-dispersion refers to the situation where the observations Y_i have a variance larger than that expected under the binomial model, i.e. $n_i\pi_i(1-\pi_i)$. This is phenomenon is quite common in practice. It can occur for various regions, such as an inadequate model specification (e.g., relevant explanatory variables have been omitted or the link function is incorrect). One approach to deal with over-dispersion in this case is to include an extra parameter ϕ in the model so that $\text{var}(Y_i) = n_i\pi_i(1-\pi_i)\phi$. The implementation of this depends on the statistical software you use. Another reason for over-dispersion may be that the Y_i

are correlated. In this case we must use modelling approaches designed for correlated data. A more detailed discussion of over-dispersion can be found in Nelder and McCullagh (1989).

Example

Weisberg (1985) reports data on an experiment carried out on cows[1]. The effect of small electrical currents on farm animals is of interest, with the eventual goal of understanding the effects of high-voltage powerlines on livestock. This experiment was carried out with 7 cows, and 6 shock intensities 0, 1, 2, 3, 4 and 5 milliamps. It is reported that shocks in the order of 15 milliamps are painful for any human. Each cow was given 30 shocks, five at each intensity level, in random order. The entire experiment was then repeated, so each unfortunate cow received a total of 60 shocks! For each shock, the response, mouth movement, was either present or absent. The data in Table 1 give the total number of responses, out of the 70 resulting trials for each shock level. We ignore differences between cows and blocks in this experiment.

We use SPSS 11.0 to fit the following logistic regression model to the data:

$$\text{logit}(\pi_i) = \log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 x_i$$

Binary logistic regressions models can be fitted using either the Logistic Regression procedure or the Multinomial Logistic Regression procedure. For example, the fitted values and residuals are only available in the Logistic regression procedure, while the results from the deviance and Pearson goodness of fit tests are only available in the Multinomial Logistic Regression procedure. Each procedure has options not available in the other. The result summaries from the fit of the above model in SPSS are presented below Table 1.

[1] The data were taken from Weisberg, S. (1985). *Applied linear regression (2nd ed.)*. Toronto: Wiley.

Table 1: Cow shock data

Current (Mill amperes)	Number of Trials	Number of Responses	Proportion of Responses
0	70	0	0.000
1	70	9	0.129
2	70	21	0.300
3	70	47	0.671
4	70	60	0.857
5	70	63	0.900

Likelihood Ratio Test

Model	-2 Log Likelihood	Chi-Square	df	Sig.
Intercept Only	271.227			
Final	30.093	241.134	1	.000

Goodness-of-Fit

	Chi-Square	df	Sig.
Pearson	7.583	4	.108
Deviance	9.353	4	.053

Parameter Estimates

	B	S.E.	Wald	df	Sig.	Exp(B)
b1	1.246	.112	123.918	1	.000	3.476
Constant	-3.301	.324	103.928	1	.000	.037

We see that the estimated parameters for the above model are $\beta_0 = -3.301$ and $\beta_1 = 1.246$. The Likelihood ratio test compares the above model to the reduced model with just a constant term β_0. There is clear evidence from this test that the response rate increases with intensity, since the test shows that the model including the β_1 term provides a much better fit. This is confirmed the Wald test.

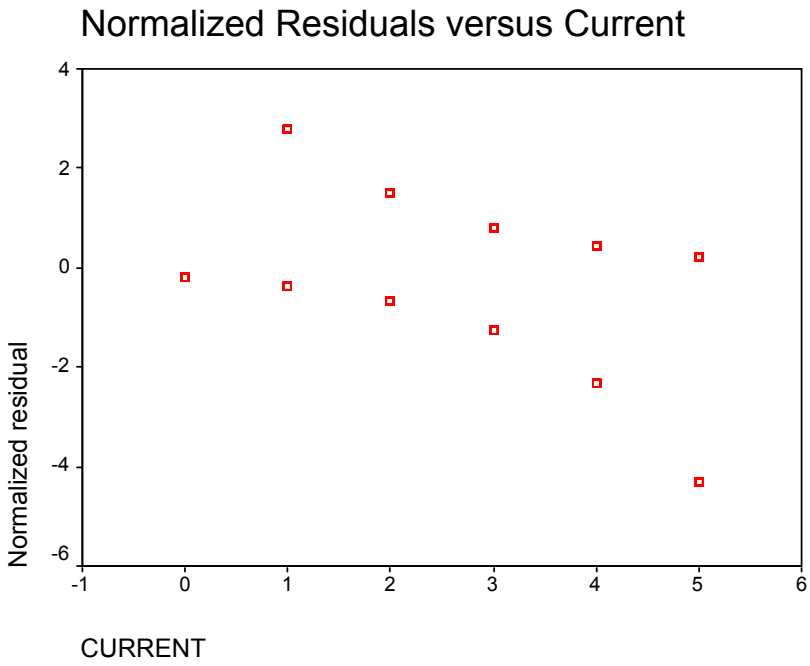

Figure 1: Standardized residuals versus the dose for the cow shock example.

The plot of the standardized residuals against the current shows some systematic pattern, which suggests that we may need alternative terms in the model. The two curve like patterns represent the residuals of a positive (top curve) and negative (bottom curve) response, respectively. It seems as though the residuals for a negative response decrease as the current increases and the other way around for the residuals relating to a positive response. This lack of fit is also reflected by the relatively small p-value relating to the deviance (0.053). However, we will not pursue this issue any further here.

III. Polytomous data

Introduction

In the previous section we have focused on response variables with two categories. In this section the response variable takes on a fixed set of possible values. Such a response is called polytomous. There are two broad approaches to model this type of data using generalized linear models. The first approach is a *generalization of the logistic regression model* of the previous section for polytomous responses that are either nominal or ordinal. The second option is to model the counts for the covariates as the response, by assuming that they are events from a Poisson process. This approach is called *log-linear modelling*. We will discuss the latter case in the next section.

In the *nominal or ordinal logistic regression* approach one observed polytomous variable

is regarded as the response, while the other variables are regarded as explanatory variables. In the case of log-linear models there is no single variable that clearly stands out as a response, thus all the variables are treated alike. The choice between these two modelling approaches may also depend on how we intend to interpret and present the results. For example, log-linear models are good for exploring complex interactions in the data, while parameter estimates are not always easy to interpret.

Polytomous responses can be classified according to different scales that require different kinds of models. We distinguish between the following major types:

1. *Nominal scales* in which the categories are regarded as exchangeable and totally devoid of structure

2. *Ordinal scales* in which the categories are ordered. The 'distance' or 'spacing' between the categories is not of importance.

3. *Interval scales* in which the categories are ordered and numerical labels or scores are attached to the various categories. Differences between the categories are therefore interpreted as measures of separation between categories.

The distinction between nominal and ordinal scales is usually, but not always clear. For instance responses relating to perception of food quality – excellent, good, bad and appalling – are clearly ordinal.

Nominal logistic regression

We will first focus on the situation where there is no natural ordering of the response categories. The first step is to arbitrarily choose one category as the reference category. For convenience we will choose the first category for this purpose. Now for the nominal logistic regression model we define *logits* for the non-reference categories as

$$\text{logit}(\pi_j) = \log\left(\frac{\pi_j}{\pi_1}\right) = \beta_{0j} + \beta_{1j}x_{1j} + \cdots + \beta_{pj}x_{pj}, \quad \text{for } j = 2,\ldots,J.$$

This represents $J-1$ logit equations that are used simultaneously to obtain estimates of the parameters $\underline{\beta}_j$, say $\underline{b}_j$. The parameters estimates can then be used to obtain estimates for the category probabilities $\hat{\pi}_1,\ldots,\hat{\pi}_J$, subject to the constraint that $\hat{\pi}_1 + \hat{\pi}_2 + \ldots + \hat{\pi}_J = 1$. This is done by rewriting the above model, so that the π_i s are the responses. Once we have estimated probabilities, we can calculate fitted values for the model by multiplying each of the estimated probabilities by the total number of observations, n.

Model diagnostics

The summary statistics to evaluate the goodness of fit for the above model is analogous to those for the logistic regression model for binary responses. We can evaluate

goodness of fit using the chi-square statistic $\left(X^2\right)$, deviance (D)or likelihood ratio chi-square statistic as before. If the model fits well, then both X^2 and D have asymptotic χ^2 distributions on $N-p$ degrees of freedom, where p is the number of parameters. We can also use the residuals as before to assess the adequacy of the model.

Odds ratios and model interpretation

Often it is not easy to directly interpret the model parameters $\beta_0, \beta_1, \ldots, \beta_p$. Odds ratios often provide a much easier interpretation. In order to explain the use of odds ratios, we will consider a simple model involving a response variable with J categories and a binary explanatory variable x. The explanatory variable x denotes the presence $(x=1)$ or absence $(x=0)$ of some 'exposure' factor. The odds ratio for 'exposure' for response j $(j=2,\ldots,J)$ relative to the reference category $j=1$ is the defined as

$$\theta_j = \frac{\pi_{j,present}}{\pi_{j,absent}} \Big/ \frac{\pi_{1,present}}{\pi_{1,absent}}$$

where π_{jp} and π_{ja} denote the probabilities of response category j $(j=1,\ldots,J)$ for the cases where the exposure is present or absent, respectively. The model for our example can be specified as

$$\log\left(\frac{\pi_j}{\pi_1}\right) = \beta_{0j} + \beta_{1j}x, \quad j=2,\ldots,J.$$

The logarithm of the odds ratio for the above model can be written as

$$\log\left(\theta_j\right) = \log\left(\frac{\pi_{jp}}{\pi_{1p}}\right) - \log\left(\frac{\pi_{ja}}{\pi_{1a}}\right) = \beta_{1j}.$$

It is clear that we estimate the odds ratio by $\hat{\theta}_j = \exp\left(b_{1j}\right)$, where b_{1j} is the estimate of parameter β_{1j}. The odds ratio can equal any non-negative number. When $1<\theta_j<\infty$, the presence of the exposure factor is more likely for response category j than for the baseline response category (category 1 in this case). For instance, when $\theta_j=4$, the odds for the presence of the exposure factor are four times higher for response category j than it is for response category 1. It is important to note that this does not mean that the probability for the presence of the exposure factor is four times higher for response category j than it is for response category 1. When $0<\theta_j<1$, the first response is less likely in response category j than it is in response category 1.

Example

We consider an example in which subjects were asked to rate how important various features were to them when they were buying a car[2]. Table 2 shows the ratings for air conditioning and power steering, according to sex and age of the subject. In this example the response, importance of air conditioning and power steering, is rated on an ordinal scale but for the purpose of this example we ignore order and the 3-point scale is treated as nominal. The category 'Very Important' is chosen as the reference or baseline category.

Sex	Age	No or little importance	Important	Very Important	Total
Female	18-23	26	12	7	45
	24-40	9	21	15	45
	>40	5	14	41	60
Male	18-23	40	17	8	65
	24-40	17	15	12	44
	>40	8	15	18	41
Total		105	94	101	300

Table 2: Importance of air conditioning and power steering in cars.

We fit a nominal logistic regression model to the data with age group and sex as covariates, that is

$$\log\left(\frac{\pi_j}{\pi_{\text{very important}}}\right) = \beta_{0j} + \beta_{1j}x_1 + \beta_{2j}x_2, \quad j = \text{not important, important}$$

where

$$x_1 = \begin{cases} 1 & \text{for men} \\ 0 & \text{for women} \end{cases} \qquad \text{and} \qquad x_2 = \begin{cases} 0 & \text{age 18 - 23} \\ 1 & \text{age 24 - 40} \\ 2 & \text{age > 40} \end{cases}$$

The deviance and Pearson chi-square statistics suggest that the model fits the data well. Below we also include the parameter estimates and predicted frequencies for the fitted model. The parameter estimates for the model are: $\beta_{0,\text{not important}} = 1.092$, $\beta_{0,\text{important}} = 0.590$, $\beta_{1,\text{not important}} = 0.813$, $\beta_{1,\text{important}} = 0.424$, $\beta_{2,\text{not important}} = -1.521$ and $\beta_{2,\text{important}} = -0.691$. The predicted proportions seem to match the observed frequencies

[2] The data were taken from Dobson, A. J. (1990). *An introduction to generalized linear models (2nd ed.)*. New York: Chapman & Hall.

reasonably well in general. We can use the parameter estimates to calculate odds ratios. For example, say we want to compare how sex influences an 'important' response, given a certain age. At a particular age, the odds ratio for a response of 'important' between men and women is

$$\theta_{\text{important}} = \exp\left(\beta_{1,\text{imp}}\right) = 1.528.$$

Thus, given a certain age, men are 1.528 times more likely to have a response of 'important' than women of the same age.

Goodness-of-Fit

	Chi-Square	df	Sig.
Pearson	5.445	6	.488
Deviance	5.337	6	.501

Parameter Estimates

RESPONSE		B	Std. Error	Wald	df	Sig.	Exp(B)	95% Confidence Interval for Exp(B)	
								Lower Bound	Upper Bound
import	Intercept	.590	.314	3.540	1	.060			
	SEX	.424	.299	2.006	1	.157	1.528	.850	2.748
	AGE	-.691	.191	13.108	1	.000	.501	.345	.728
ni	Intercept	1.092	.309	12.535	1	.000			
	SEX	.813	.321	6.411	1	.011	2.255	1.202	4.231
	AGE	-1.521	.211	51.769	1	.000	.218	.144	.331

Observed and Predicted Frequencies

AGE	SEX	RESPONSE	Frequency			Percentage	
			Observed	Predicted	Pearson Residual	Observed	Predicted
.00	.00	import	12	14.036	-.655	26.7%	31.2%
		ni	26	23.186	.839	57.8%	51.5%
		very	7	7.778	-.307	15.6%	17.3%
	1.00	import	17	17.106	-.030	26.2%	26.3%
		ni	40	41.692	-.438	61.5%	64.1%
		very	8	6.203	.759	12.3%	9.5%
1.00	.00	import	21	15.924	1.583	46.7%	35.4%
		ni	9	11.466	-.844	20.0%	25.5%
		very	15	17.611	-.797	33.3%	39.1%
	1.00	import	15	15.793	-.249	34.1%	35.9%
		ni	17	16.778	.069	38.6%	38.1%
		very	12	11.429	.196	27.3%	26.0%
2.00	.00	import	14	17.040	-.870	23.3%	28.4%
		ni	5	5.348	-.158	8.3%	8.9%
		very	41	37.612	.904	68.3%	62.7%
	1.00	import	15	14.102	.295	36.6%	34.4%
		ni	8	6.530	.627	19.5%	15.9%
		very	18	20.368	-.740	43.9%	49.7%

The percentages are based on total observed frequencies in each subpopulation.

Ordinal logistic regression

In the previous subsection we considered the situation where there was no ordering in the response categories. When the response categories have a natural ordering, model specification should take that into account so that the extra information is utilized in

the model. This ordering can be incorporated directly in the way we specify the logits. Ordinal responses like this are common in areas such as market research, opinion polls and psychiatry. Sometimes the response z may be some continuous variable, which is difficult to measure, so that its range is divided into J ordinal categories with associated probabilities $\pi_1, \pi_2, \ldots, \pi_J$. There are a few commonly used models for this situation. We will consider of these models.

Cumulative logit model

Before we introduce the model we first need to define the cumulative odds for the j th response category. This is given by

$$\frac{\pi_1 + \pi_2 + \cdots + \pi_j}{\pi_{j+1} + \cdots + \pi_J}$$

where J is the total number of response categories. The cumulative logit model is

$$\log \frac{\pi_1 + \pi_2 + \cdots + \pi_j}{\pi_{j+1} + \cdots + \pi_J} = \beta_{0j} + \beta_{1j} x_{1j} + \cdots \beta_{pj} x_{pj} .$$

Proportional odds model

In the above model all the parameters depend on the category j. The proportional odds model is based on the assumption that the effects of the covariates $x_1, \ldots, x_p$ are the same for all categories, on the logarithmic scale. Thus, in this model only the intercept term β_{0j} depends on the category j so that the model is

$$\log \frac{\pi_1 + \pi_2 + \cdots + \pi_j}{\pi_{j+1} + \cdots + \pi_J} = \beta_{0j} + \beta_1 x_1 + \cdots + \beta_p x_p .$$

The appropriateness of the model can be tested separately for each variable by comparing it to a cumulative odds model with the relevant parameter not depending on j. The proportional odds model is usually the default form of ordinal logistic regression provided by statistical software packages.

Residuals and goodness of fit statistics are analogous to those for nominal logistic regression. The choice of model for ordinal data depends mainly on the practical problem being investigated. Two further models that are commonly used for this kind data are the *adjacent categories logit model* and the *continuation ratio logit model*; however, we will only mention these models here as options to look out for in statistical software packages.

Example

The response variable in the previous example is ordinal, although we treated it as nominal. Here we return to the car preference data example to fit an ordinal

regression model to the data. The default model that is fitted by the SPSS 11.0 ordinal regression procedure is the proportional odds model. We fit the proportional odds model

$$\log\left(\frac{\pi_{\text{ni}}}{\pi_{\text{imp}}+\pi_{\text{very imp}}}\right)=\beta_{01}+\beta_1 x_1+\beta_2 x_2$$

$$\log\left(\frac{\pi_{\text{ni}}+\pi_{\text{imp}}}{\pi_{\text{very imp}}}\right)=\beta_{02}+\beta_1 x_1+\beta_2 x_2$$

with x_1 and x_2 defined as in the previous example. The results for this model and the nominal logistic model are similar for this example, but we would prefer the ordinal model because it is simpler (it has 2 less parameters) and takes the ordering of the response categories into account. The results of the model fit are presented below.

Goodness-of-Fit

	Chi-Square	df	Sig.
Pearson	4.575	8	.802
Deviance	4.549	8	.805

Link function: Logit.

6

Parameter Estimates

		Estimate	Std. Error	Wald	df	Sig.	95% Confidence Interval	
							Lower Bound	Upper Bound
Threshold	[RESPONSE = 0]	3.310E-02	.217	.023	1	.879	-.393	.459
	[RESPONSE = 1]	1.644	.242	46.263	1	.000	1.171	2.118
Location	SEX	-.577	.226	6.512	1	.011	-1.020	-.134
	AGE	1.116	.145	59.131	1	.000	.832	1.401

Link function: Logit.

Cell Information

Frequency

SEX	AGE		RESPONSE		
			0	1	2
.00	.00	Observed	26	12	7
		Expected	22.872	14.844	7.284
		Pearson Residual	.933	-.902	-.115
	1.00	Observed	9	21	15
		Expected	11.379	16.926	16.694
		Pearson Residual	-.816	1.254	-.523
	2.00	Observed	5	14	41
		Expected	5.986	15.433	38.580
		Pearson Residual	-.425	-.423	.652
1.00	.00	Observed	40	17	8
		Expected	42.118	16.523	6.360
		Pearson Residual	-.550	.136	.685
	1.00	Observed	17	15	12
		Expected	16.546	16.506	10.948
		Pearson Residual	.141	-.469	.367
	2.00	Observed	8	15	18
		Expected	6.758	13.625	20.617
		Pearson Residual	.523	.456	-.818

Link function: Logit.

It is also possible to obtain estimated probabilities for the various covariate patterns.

We did not include those here. The model seems to describe the data well, as suggested by the goodness of fit criteria above. The parameter estimates for our model are also presented above, with $\beta_{01} = 0.031$, $\beta_{02} = 1.644$, $\beta_1 = -0.577$ and $\beta_2 = 1.116$.

IV. Count data

Introduction

This section focuses on the analysis of count data. Here we are interested in the number of times an event occurs. We will distinguish between two different situations, as we have mentioned earlier. In the first situation, the events relate to varying amounts of 'exposure', which need to be taken into account when modelling the rate of events. *Poisson regression* is used in this case. The other explanatory variables (in addition to 'exposure') may be continuous or categorical. The counts may for example be the number of traffic accidents, which need to be analysed in relation to some exposure variable such as the number of registered motor vehicles. In the second situation, 'exposure' is constant (and therefore not relevant to the model) and the explanatory variables are usually categorical. If there are only a few explanatory variables the data are summarized in a cross-classified table. The response is the frequency or count in each cell of the table. The variables used to define the table are all treated as explanatory variables. The study design may mean that there are certain constraints on the cell frequencies (for example, the totals of the frequencies in each row of the table may be equal) and this need to be taken into account in the modelling. *Log-linear models* are appropriate for this situation. We will consider an example relating to each of the above situations.

Poisson Regression

Let $Y_1, Y_2, \ldots, Y_N$ be independent random variables with Y_i denoting the number of events observed from exposure n_i for the *i*th covariate class. The expected value of Y_i can be written as

$$E(Y_i) = \mu_i = n_i \theta_i .$$

For example, suppose Y_i is the number of insurance claims for a particular make and model of car. This will depend on the number of cars of this type that are insured, n_i, and other variables that affect θ_i, such as the age of the cars and the location where they are used. The subscript *i* is used to denote the different combinations of make, model, age, location and so on. The dependence of θ_i on the explanatory variables is usually modelled by

$$\theta_i = e^{\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p} .$$

This can be modelled within the generalized linear model set-up as

$$E(Y_i) = \mu_i = \eta_i e^{\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p}; \quad Y_i \sim \text{Poisson}(\mu_i).$$

The natural link function is the logarithmic function

$$\log \mu_i = \log n_i + \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p .$$

The inclusion of the term $\log n_i$ differs from the usual specification of the linear component. This term is called the offset and is a known constant that is incorporated into the estimation of the parameters $\beta_0, \ldots, \beta_p$.

The parameter estimates are often interpreted on the exponential scale e^{β} in terms of ratios of rates. For a binary explanatory variable denoted by an indicator variable, $x_j = 0$ if the factor is absent or $x_j = 1$ if it is present, the *rate ratio*, RR, for presence versus absence is defined as

$$RR = \frac{E(Y_i \mid present)}{E(Y_i \mid absent)} = e^{\beta_j} .$$

Similarly, for a continuous explanatory variable x_k, a one-unit increase will result in a multiplicative effect of e^{β_k} on the rate μ. Hypotheses about the parameters can be tested using the Wald score or likelihood ratio statistic. Confidence intervals can also be constructed similarly. Alternatively, hypothesis testing can be performed by comparing the goodness of fit of appropriately defined nested models.

Goodness of fit

Pearson and deviance residuals can also be calculated for Poisson regression models as before to evaluate the fit. The deviance and X^2 statistics can be used directly as measures of goodness of fit, as they can be calculated from the data and the fitted model (there are no nuisance parameters such as σ^2 for the Normal distribution). Two other summary statistics that are often provided in software packages for Poisson regression models are the likelihood ratio chi-square statistic and the pseudo-R^2. These statistics are based on comparisons between the maximum value of the log-likelihood function for a minimal model with no covariates and the maximum value of the log-likelihood function for a model with p parameters.

Example

The data in Table 3 are from a study conducted in 1951, in which all British doctors were sent a brief questionnaire about their smoking status[3]. Since then information of

[3] The data were taken from Dobson, A. J. (1990). *An introduction to generalized linear models (2nd ed.)*. New York: Chapman & Hall.

their deaths has been collected. Table 3 shows the numbers of deaths from coronary heart disease among male doctors 10 years after the survey. It also shows the total number of person-years of observation at the time of the analysis. This is the 'exposure' that was mentioned above.

Table 4: Smoke data

Age Group	Smokers		Non-smokers	
	Deaths	Person-years	Deaths	Person-years
35-44	32	52407	2	18790
45-54	104	43248	12	10673
55-64	206	28612	28	5710
65-74	186	12663	28	2585
75-84	102	5317	31	1462

The questions of interest were:

1. Is the death rate higher for smokers than for non-smokers?
2. If so, by how much?
3. Is the differential effect related to age?

As illustration we will consider one Poisson regression model that describes the data well (there are more than one). The model we will consider is

$$\log(deaths_i) = \log(population_i) + \beta_1 + \beta_2 smoke_i + \beta_3 agecat_i + \beta_4 agesq_i + \beta_5 smkage_i$$

The subscript i denotes the ith subgroup defined by age group and smoking status ($i = 1,\ldots 5$ for ages 35-44,…, 75-84 for smokers and $i = 6,\ldots 10$ for the corresponding age groups for non-smokers). The term $deaths_i$ denotes the expected number of deaths and $population_i$ denotes the number of doctors at risk in group i. $Smoke_i$ is an indicator variable for smoking status and that takes a value of one for smokers and zero for non-smokers; $agecat_i$ takes values 1,…, 5 for age groups 35-44,…, 75-84; $agesq_i$ is the square

root of $agecat_i$ which is included to take account of a non-linear death rate of increase with age; $smkage_i$ is an extra age term that is equal to $agecat_i$ for smokers and zero for non-smokers, thus describing different rates of increase in the number of deaths for smokers and non-smokers. The model output is given below.

Table 4: Poisson regression output

Term	agecat	agesq	smoke	smkage
$\hat{\beta}$	2.376	-0.198	1.441	-0.308
s.e. ($\hat{\beta}$)	0.208	0.027	0.372	0.097
Wald statistics	11.43	-7.22	3.87	-3.17
p-value	<0.001	<0.001	<0.001	0.002
Rate Ratio	10.77	0.82	4.22	0.74
95% conf. Int.	(7.2,16.2)	(0.78,0.87)	(2.04,8.76)	(0.61,0.89)

The Wald statistics to test $\beta_j = 0$ all have very small p-values and the confidence intervals for e^{β_j} do not contain unity showing that all the terms are needed in the model. The estimates show that the risk of coronary deaths was, on average, about four times higher for smokers than non-smokers (based on the rate ratio for *smoke*), after the effect of age is taken into account. The goodness of fit measures $X^2 = 1.550$ and $D = 1.635$ show that the fit for the model is good, when compared to the appropriate χ^2 distribution.

Log-linear Models

In the case of log-linear models there is no single variable that clearly stands out as a response, thus all the variables are treated alike. Typical examples involve counts in a Poisson or Poisson-like process. These counts are treated as independent Poisson variables. If there are only a few variables the data are summarized in a cross-classified table, called a *contingency table*. We define a contingency table as follows: divide each variable into categories (if necessary) and count the number of cases falling into each of the possible combinations of categories. It is these counts that we wish to model. So if Y_i is the random variable representing the count in the i th cell of the

contingency table, then $Y_i \sim \text{Poisson}(\mu_i)$, where $E(Y_i) = \mu_i$ and $\text{var}(Y_i) = \mu_i$.

It follows that the *log-link* is the natural choice for the link function in a GLM set-up. This follows because log of the mean of each cell is equal to the linear predictor, i.e.

$$\log(\mu_i) = \eta_i = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p.$$

The term log-linear model is used to describe this relationship between the means of the cell counts and the linear predictor through the log-link. For contingency tables the main questions almost always relate to associations between variables. Therefore, in log-linear models, the terms of primary interest are the interactions involving two or more variables. We see that there is no offset term as was the case for Poisson regression. In this case 'exposure' is constant and therefore not relevant to the model. This follows from certain constraints on the cell frequencies (for example, the totals of the frequencies in each row of the table may be equal).

Goodness of fit

The adequacy of fit of a log-linear model can be assessed through the D and X^2 goodness of fit statistics as before. More insight into model adequacy can often be obtained by examining the Pearson or deviance residuals. Hypothesis tests can be conducted by comparing the difference in goodness of fit statistics between a general model corresponding to an alternative hypothesis and a nested, simpler model corresponding to a null hypothesis. We will illustrate some of these methods in the example below.

Example

We consider an example involving a $2 \times 2 \times 2$ contingency table that contains data from a study of the effects of racial characteristics on whether individuals convicted of homicide receive the death penalty[4]. The data are presented in Table 5.

Table 5: Death penalty data

[4] The data were taken from Agresti, A. (1990). *Categorical data analysis*. New York: Wiley.

Defendant's Race	Victim's Race	Death Penalty		Percentage Yes
		Yes	No	
White	White	26	12	12.6
	Black	9	21	0.0
Black	White	40	17	17.5
	Black	17	15	5.8

If we fit all models from the simplest model, containing just the main effects, to the most complex model containing all interactions between the three variables, it follows from goodness of fit test (in this case the likelihood ratio and Pearson chi-square test) that only two models fit adequately at a 0.05 level of significance. The model containing interaction terms between the victim's race and the penalty and between the victim's race and the defendant's race is the simplest model that provides a good fit. According to this model, the death penalty verdict is independent of the defendant's race, if we are given the victim's race. This model can be specified as

$$\log(\mu_{ijk}) = \alpha + \lambda_i^D + \lambda_j^V + \lambda_k^P + \lambda_{ij}^{DV} + \lambda_{jk}^{VP}, \quad i = 1,2;\ j = 1,2;\ k = 1,2$$

The results for the SPSS fit of the above model are given below.

```
Goodness-of-fit Statistics

                    Chi-Square        DF        Sig.

Likelihood Ratio        1.8819         2       .3903
         Pearson        1.4313         2       .4889

_
```

```
- - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -
-
                          GENERAL LOGLINEAR ANALYSIS
- - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -
-

Parameter Estimates

                                                        Asymptotic 95% CI
Parameter    Estimate         SE     Z-value      Lower       Upper

        1      4.5797      .1011       45.31       4.38        4.78
        2     -2.8717      .4196       -6.84      -3.69       -2.05
        3      .0000       .             .          .           .
        4     -2.4375      .3476       -7.01      -3.12       -1.76
        5      .0000       .             .          .           .
        6      -.5876      .1639       -3.59       -.91        -.27
        7      .0000       .             .          .           .
        8     1.0579       .4635        2.28        .15        1.97
        9      .0000       .             .          .           .
       10      .0000       .             .          .           .
       11      .0000       .             .          .           .
       12     3.3116       .3786        8.75       2.57        4.05
       13      .0000       .             .          .           .
       14      .0000       .             .          .           .
       15      .0000       .             .          .           .
```

The Pearson chi-square and Likelihood ratio goodness of fit statistics both show that the model provides a good fit to the data. The parameter estimates for our model follow from the above as:

$\alpha = 4.5797$, $\lambda_1^D = -2.8717$, $\lambda_1^V = -2.4375$, $\lambda_1^P = -0.5876$, $\lambda_{11}^{DV} = 1.0579$, $\lambda_{11}^{VP} = 3.3116$.

The parameters are computed under the constraints that $\lambda_2^D = 0$, $\lambda_2^V = 0$ and $\lambda_{12}^{DV} = \lambda_{21}^{DV} = \lambda_{12}^{VP} = \lambda_{21}^{VP} = 0$. Now the model can be interpreted by computing odds ratios. We will not go into this analysis, but interested readers are referred to Agresti (1990).

V. Survival Analysis and Repeated Measures

There are two important types of data that we have not discussed, which can also be modelled with GLMs (and other methods). We will not discuss these types of data in detail here. We will merely mention these as types of data that require their own approaches to modelling. The first of these is *survival data*. This data is the time from some well-defined starting point until some event, called 'failure' occurs. These times are non-negative and typically have skewed distributions with long tails. The analysis of survival data is the topic of numerous books and most statistical software packages have procedures to implement appropriate statistical methods for this data. The exponential distribution plays a central role in the analysis of survival data.

In all the models that we have considered so far we have assumed that the outcomes

$Y_i, i = 1,\ldots,n$ are independent. There are two common situations where this assumption is implausible. In one situation the outcomes are repeated measures over time of the same subjects. This is an example of *longitudinal data*. Longitudinal data for a group of subjects are likely to exhibit correlation between successive measurements. The other situation in which data are likely to be correlated is where they are measurements on related subjects; for example, the weights of samples of women aged 40 years selected from specific locations in different countries. This kind of data is sometimes referred to as *clustered data*. The term *repeated measures* is used to describe both longitudinal and clustered data. Methods such as *repeated measures ANOVA* and *generalized estimating equations* are examples of methods to deal with these kinds of data.

References

Agresti, A. (1990). *Categorical data analysis*. New York: Wiley.

Dobson, A. (1990). *An introduction to Generalized Linear Models (2nd ed.)*. Boca Raton, FL: Chapman and Hall/CRC.

McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models (2nd ed.)*. London: Chapman and Hall.

Weisberg, S. (1985). *Applied linear regression*. New York: Wiley