Logistic and Poisson Regression
An Introduction to Generalized Linear Models
Uppsala University
March, 2001

Jonathan Yuen

Department of Ecology and Crop Production Science
Swedish University of Agricultural Sciences

e-mail: Jonathan.Yuen@evp.slu.se
telephone (018) 672369

February 21, 2001
## Contents

1 An Introduction to General Linear Models (GLM)  
   1.1 Components of a GLM .......................... 5  
   1.2 A Simple GLM .................................. 6  
   1.3 A numeric example of a simple GLM ............. 7  

2 Unconditional Logistic Regression  
   2.1 A Numeric Example .............................. 12  
   2.2 Another Numeric Example of Logistic Regression 15  

3 Design Matrices and Aliasing  
   3.1 Studying Design Matrices with GLMMOD ............. 18  

4 Logistic Regression with Several Independent Variables  
   4.1 Another Numeric Example .......................... 22  

5 Deviance and Goodness of Fit  

6 A Detailed Analysis with Analysis of Deviance  

7 Poisson Regression  
   7.1 Introduction ..................................... 31  
   7.2 A Numeric Example of Poisson Regression .......... 32  
   7.3 Another Example of Poisson Regression ............. 34  

8 A Working Example  

9 Goodness of Fit and Over-dispersion  

References  

5
6
7
11
12
15
17
21
25
27
31
32
34
38
41
44
List of Tables

1 Structure of Some Common General Linear Models . . . . . . . . 6
2 Data for a Case-Control Study Arranged in a 2 x 2 table . . . . . 13
3 Data for a Simple Case-Control Study . . . . . . . . . . . . . . 14
4 Stratified Data for a Case-Control Study . . . . . . . . . . . . . 21
5 A Likelihood Ratio Test for Exposure . . . . . . . . . . . . . . 23
6 Analysis of Deviance from Esophageal Cancer Study . . . . . . 29
7 Analysis of Deviance from Non-melanoma Skin Cancer Analysis . 37

List of Figures

1 Proportion ’P’ as a function of logit(p) . . . . . . . . . . . . . . 12
2 Likelihood Surface . . . . . . . . . . . . . . . . . . . . . . . . 13
This is a collection of material I have written for different things. I have taken this text from web pages, seminars, lecture notes, and manuscripts published or somewhere in the publication process. For those of you that are unfamiliar with human disease epidemiology, I apologize for the use of examples that are based on case-control studies, which are analyzed with logistic regression. This is due to the fact that this material comes from an era when the students taking the courses had very mixed backgrounds, and this type of material was relevant for many of them at that time. An advantage of these studies is that the sampling proportion of diseased and non-diseased units need not be identical.

Jonathan Yuen
Uppsala February 2001
1 An Introduction to General Linear Models (GLM)

General Linear Models (McCullagh & Nelder, 1989), here abbreviated as GLM (not to be confused the procedure GLM in SAS which we will use briefly) is a concept that unifies many different types of statistical models. These models include:

- t tests
- Analysis of Variance
- Multiple Regression
- Analysis of Covariance
- Logistic Regression
- Poisson Regression
- Analysis of Dilution Assays
- Probit Analysis

1.1 Components of a GLM

A GLM has several components. These are

**Random Components** This concerns the dependent variable, and we allow for a discrepancy between observed and ‘true’ (undoubtedly unknown) values. Traditionally the observed values of the dependent variable are denoted by $y$.

**Systematic Component** The independent variables. Covariates (usually denoted by $x_j$) and their unknown parameters (usually denoted by $\beta_j$).

The product of each covariate and its parameter are summed. Assuming we have $p$ covariates:

$$x_1\beta_1 + x_2\beta_2 \cdots x_p\beta_p$$

Mathematically, it is often written like this

$$\sum_{j=1}^{p} x_j\beta_j$$
Within the context of a GLM, this is often referred to as the linear predictor (LP), and is referred to with the Greek letter $\eta$, pronounced 'eta'. Thus,

$$\eta = \sum_{j=1}^{p} x_j \beta_j$$

but I will generally refer to it as the linear predictor.

Link A link between the systematic component and the dependent variable. This can be (in simple cases) an identity function ($=$), or some other mathematical function. This is often referred to as $g$.

Table 1 lists some of the common types of GLM’s

<table>
<thead>
<tr>
<th>Analysis</th>
<th>Random Part</th>
<th>Systematic Part</th>
<th>Link</th>
</tr>
</thead>
<tbody>
<tr>
<td>$t$ tests</td>
<td>Normal</td>
<td>Categorical</td>
<td>Identity</td>
</tr>
<tr>
<td>ANOVA</td>
<td>Normal</td>
<td>Categorical</td>
<td>Identity</td>
</tr>
<tr>
<td>Multiple Reg</td>
<td>Normal</td>
<td>Continuous</td>
<td>Identity</td>
</tr>
<tr>
<td>Analysis of Cov</td>
<td>Normal</td>
<td>Cat. &amp; Cont.</td>
<td>Identity</td>
</tr>
<tr>
<td>Logistic Reg</td>
<td>Binomial</td>
<td>Cat &amp; Cont</td>
<td>Logit</td>
</tr>
<tr>
<td>Poisson Reg</td>
<td>Poisson</td>
<td>Cat &amp; Cont</td>
<td>Log</td>
</tr>
<tr>
<td>Dilution Assays</td>
<td>Binomial</td>
<td>Cat &amp; Cont</td>
<td>CLL</td>
</tr>
<tr>
<td>Probit Analysis</td>
<td>Binomial</td>
<td>Cat &amp; Cont</td>
<td>Probit</td>
</tr>
</tbody>
</table>

Table 1: Structure of Some Common General Linear Models

1.2 A Simple GLM

A conventional $t$-test can be formulated as a GLM. We observe a series of values from one of two groups. We assume that these are the result of a series of random variable and a true mean for each of the two groups. It is assumed that these random variables are independent from each other, and that together they have a normal distribution with mean zero.

In the systematic component, we can assign observations to one of two groups. In this case, we use dummy variables (which can take the value of 0 or 1) to denote assign group membership to each observation. In this case, the random and systematic parts of the GLM are coupled together with the identity function.
1.3 A numeric example of a simple GLM

Here we will use the SAS procedure GLM to fit some simple GLM’s. This procedure uses an identity link function with normally distributed errors and uses ‘least-squares’ to estimate the parameters. The SAS program is present in the file ‘dagis.sas’, and will be shown here in Courier text:

Like this

while my comments appear in this font.

All SAS program lines end with a semicolon, and SAS ignores lines that begin with an asterisk. Thus, the following line does nothing but supply us with information.

* simple linear regression stuff;

The next line gives us some printing options. I’ll eliminate this kind of stuff in the future listings, though they will be present in the disk files.

options linesize=75 pagesize = 64;

SAS works with data sets, which have to be given names. Here we create a data set called dagis, and read 4 variables directly into it with the cards statement. The four variables are sex (a character variable), wt, height, and ones. The variable ones takes the value of 1 for all observations. Data is arranged in a rectangular matrix, with observations corresponding to the rows, and different variables in each column.

data dagis;
   input sex $ wt height ones;
   cards;
   M 17 110 1
   M 15 105 1
   M 12 100 1
   F 15 104 1
   F 16 106 1
   F 14 102 1
   run;

The next stuff is for graphics on my computer (which uses unix), and I use proc gplot to produce some graphs. You can eliminate much of this if you are using Windows:
The next two statements fit a linear model or GLM. In the first, the dependent variable is height, and the independent variable is wt. This is reversed in the second model. This pattern (the dependent variable followed by the = sign, and then the independent variables) is one we will see in proc GENMOD. Note that both height and weight are continuous variables.

```
proc glm;
  model height=wt;
run;
```

We can also fit a GLM with wt as the dependent variable and height as the LP.

```
proc glm;
  model wt = height;
run;
```

The next model uses the CLASS statement to create dummy variables for sex, because this is a categorical variable. These dummy variables are then used as the independent variables in the model. If we want to see the actual parameter estimates for the dummy variables, we need to give the solution option after the '/' in the model statement.

```
proc glm;
  class sex;
  model wt = sex /solution;
  means sex;
run;
```

Had we wanted to create the dummy variables ourselves (instead of letting SAS do it with the CLASS statement, we would have needed some SAS statements in the DATA statement. Notice that the solution option is not needed in the model statement.
data dagis;
  input sex $ wt height ones;
  if sex = 'M' then sexd1 = 1;
  else sexd1 = 0;
  if sex = 'F' then sexd2 = 1;
  else sexd2 = 0;
cards;
  M 17 110 1
  M 15 105 1
  M 12 100 1
  F 15 104 1
  F 16 106 1
  F 14 102 1
run;

proc glm;
  model wt = sexd1 sexd2 ;
  means sex;
run;

While it is important to understand the relationship of categorical variables to the dummy variables (and there is a section in the future dedicated to just this topic) the CLASS statement is a valuable tool in the GENMOD procedure. While making two dummy variables, as we did here, is relatively simple, it can rapidly become complex with several variables, several different categories in each, and interaction terms. The CLASS statement is one reason we are doing these exercises with GENMOD not with PROC LOGISTIC.

GLM fits (by default) an intercept term in all models. This can be replaced by our variable 'ones', and the intercept term provided by SAS is removed with the 'noint' option.

proc glm;
  class sex;
  model wt = ones sex /noint solution;
  means sex;
run;

The information provided by the dummy variables together with the intercept term is overlapping (aliased). SAS eliminates one of these variables. In the two previous examples, the aliased variable removed was the last one (i.e. the dummy variable that was equal to one when sex was 'M'). This is only one way (of many
possible ways) to deal with the aliasing. If we remove the intercept to avoid aliasing, the model fit is the same, but the parameters are different. In the first two examples with the intercept term, the intercept parameter represents the weight of the boys, and the parameter estimate for the girls is the difference between the weight of the girls and the boys.

```plaintext
proc glm;
  class sex;
  model wt = sex /noint solution;
  means sex;
run;
```

Without the intercept term, the regression parameters for the dummy variables created for sex represent the average weight of the boys and the average weight of the girls.

Models can also combine both continuous variables and the dummy variables created by the CLASS statement. Interpretation of the parameters from these models is taken up later.

```plaintext
proc glm;
  class sex;
  model wt = height sex /solution;
run;
```

In this model, we assume an effect of height on weight, and an effect of sex on weight.

```plaintext
proc glm;
  class sex;
  model wt = height sex sex*height /solution;
run;
```

In the previous model, we can also examine an interaction between sex and height in addition to the effect of height and sex.

```plaintext
proc glm;
  class sex;
  model wt = sex sex*height /noint solution;
run;
```

The final model is exactly the same as the one before (an effect of sex, an effect of height, and an interaction between them), but we calculate the parameters differently, so that we obtain the intercept and slope of the two lines predicting weight as a function of height, one for the boys and one for the girls.
2 Unconditional Logistic Regression

Unconditional logistic regression (often referred to as logistic regression) is also done with a GLM, but using a different link function and different errors. In this case, the outcome consists of the number of successes that resulted from a given number of trials. For example, we may flip a coin 25 times and note the number of times it shows heads. Assume that it comes up heads 14 times. The number of trials in this case is 25, and the number of successes 14. Since the outcome is a proportion, we can use the binomial error distribution in our GLM. The systematic component can be a mixture of categorical or continuous variables. They must have a linear relationship to each other, but that is true of all linear predictors. In logistic regression, we use the logit (logistic transformation) as the link function.

We define

$$logit(y) = \ln \frac{y}{1-y}$$

Plant pathologists (and infectious disease epidemiologists) will recognize the logistic transformation. For a compound interest disease, a plot of the logit transformed disease proportion over time approximates a straight line, the slope of which is the parameter $\gamma$, the apparent infection rate (Vanderplank, 1963).

In logistic regression, we therefore relate the logit of the proportion to the linear predictor. The discrepancies between the observed proportion of events and the true proportion of events is accounted for by allowing the predicted proportion to have a binomial error distribution.

Figure 1 gives a graphical representation of the logistic transformation. Equipped with the logit, you can easily calculate the proportion. Note that the logit of zero or one is not defined.

$$y = \ln \frac{p}{1-p}$$

$$p = \frac{e^y}{1 + e^y} = \frac{1}{1 + \frac{1}{e^y}}$$

In practice, we can no longer use least-squares (the technique used in PROC GLM and PROC ANOVA in SAS) to estimate the logistic models. Most modern techniques rely on a numerical solution, where the initial estimates are continually refined until they can be no longer improved. The examples presented here are based on PROC GENMOD in SAS, which is much like the original program GLIM originally written to estimate these general linear models.

PROC GENMOD uses a technique called Newton-Raphson to maximize the likelihood of the regression parameters (the $\beta$’s in the LP), given the data that was observed. This likelihood maximization is akin to climbing a hill, where a hiker
can estimate the position of the top of the hill given the slope and how curved the surface is. This technique also calculates the ‘curviness’ of the likelihood surface, and this information is used to calculate the standard error of the estimates.

A likelihood curve might look figure 2, which is actually from the following numeric example.

2.1 A Numeric Example

Logistic regression is often used to analyze case/control studies. In these studies, we often have all the diseased individuals (the cases), but have sampled only a proportion of the non-diseased individuals. The data can be arranged in a 2 x 2 table based on case or control status and exposed or unexposed status. I’ve done the table in Swedish, so *ja* means yes, *nej* means no, *fall* is a case, and *kontroll* is a control person.

Traditionally, the odds ratio is used to estimate the risk of being a case for exposed individuals compared to the unexposed. This would be calculated as

\[
\frac{A}{B} \div \frac{C}{D}
\]

![Figure 1: Proportion ‘P’ as a function of logit(p)](image)
Figure 2: Likelihood Surface

<table>
<thead>
<tr>
<th>CC status</th>
<th>Exposure</th>
<th>Ja</th>
<th>Nej</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fall</td>
<td>A</td>
<td>B</td>
<td></td>
<td>A+B</td>
</tr>
<tr>
<td>Kontroll</td>
<td>C</td>
<td>D</td>
<td></td>
<td>C+D</td>
</tr>
<tr>
<td>Total</td>
<td>A+C</td>
<td>B+D</td>
<td></td>
<td>A+B+C+D</td>
</tr>
</tbody>
</table>

Table 2: Data for a Case-Control Study Arranged in a 2 x 2 table
or

\[
\frac{AD}{BC}
\]

using the information presented in table 2. A detailed description of the rationale behind case/control studies is outside the scope of this manuscript, but in the calculation of the odds ratio, the unequal sampling of cases and controls appears in both the numerator and the denominator, and is thus canceled out. For further information on specifically these types of studies, an introductory textbook in human disease epidemiology such as Ahlbom’s text (1993), or Hosmer and Lemeshow (1989) would be a good starting point.

Our logistic regression example is based on the data presented in table 3 (Ahlbom, 1993).

<table>
<thead>
<tr>
<th>CC status</th>
<th>Exposure</th>
<th>Ja</th>
<th>Nej</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fall</td>
<td>100</td>
<td>50</td>
<td></td>
<td>150</td>
</tr>
<tr>
<td>Kontroll</td>
<td>100</td>
<td>100</td>
<td></td>
<td>200</td>
</tr>
<tr>
<td>Total</td>
<td>200</td>
<td>150</td>
<td></td>
<td>350</td>
</tr>
</tbody>
</table>

Table 3: Data for a Simple Case-Control Study

These data would have an odds ratio of

\[
\frac{100 \times 100}{100 \times 50} = 2.0
\]

A simple example of such a data set might be one where each line consists of information on a single person. If we have data such that we have information about a single person on each line, then the data file would have 350 lines. The first 100 lines (representing cell A) might look like this:

1 Ja

This would be followed by 100 lines (cell C) like this:

0 Ja

Cell B would be represented by 50 lines like this:

1 Nej
and cell D would be 100 lines like this:

0 Nej

The data file could be read in like this, assuming the data file is called 'demo.dat'.

data cc2;
infile 'demo.dat';
input cc expose $;
atrisk = 1;
run;

This would read in the two variables and create a third, called atrisk, which is always equal to one. We use this as the denominator in the regression. We can perform the regression by invoking GENMOD.

proc genmod;
   class expose;
   model cc/atrisk = expose / link = logit error=binomial;
run;

In GENMOD we must specify a link function and an error distribution, which are logit and binomial for logistic regression. The dependent variable consists of two parts for logistic regression. The first is the outcome variable (cc in this case), and the other is the number of trials. Since each person represents a trial in this data set, this is the variable atrisk, which is always equal to one. The class statement is used with the variable expose to create dummy variables. This is the reason for using Swedish – the word ja sorts before the word nej.

2.2 Another Numeric Example of Logistic Regression

In this example from a book for ecologists (Crawley, 1993), the data are grouped, so that we have a number of trials, and a number for the outcome on each line. In addition, we need the independent variable. This was an experiment where approximately 40 insects were placed in petri dishes, and exposed to varying levels of a chemical. After a fixed period of time, the number of insects killed by the chemical were counted.

Here we read the file 'dishes.dat' and have the variables dose (a continuous variable), the number of dead insects (the dependent variable) and the variable initial (the number of insects placed in each dish). In addition we create a second dependent variable by calculating the natural logarithm of dose.

The data file dishes.dat is an ordinary text file that looks like this:
The SAS code to read in the data looked like this:

```sas
data bugs;
  infile 'dishes.dat';
  input dose dead initial;
  ldose = log(dose);
run;
```

We then run GENMOD. The dependent variable has two parts, the outcome (dead) and the denominator (the number of trials, in this case the variable initial). This is followed by the dependent variable (dose), and then a slash (/). We then tell GENMOD which link function we are using (logit link) and what error distribution we are using. Binomial errors can be abbreviated with the letter b.

```sas
proc genmod;
  model dead/initial = dose / link=logit error=b;
run;
```

This second model is similar to the first, except that the independent variable is the natural logarithm of dose.

```sas
proc genmod;
  model dead/initial = ldose / link=logit error=b;
run;
```
3 Design Matrices and Aliasing

Aliasing is the term that is used to describe redundant variables. If the information that one independent variable contains is also found either in another independent variable, or a combination of some other independent variables, it is said to be aliased. Aliasing can be either internal or external. Internal aliasing often results from the formation of dummy variables.

External aliasing can be a result of insufficient information. Two variables may contain slightly different information, but we may lack the ability to distinguish the difference with the data we have on hand. If aliasing is not complete, data are often said to be confounded.

Assume that we have two different categorical variables. Just for the sake of discussion, let’s call them smoking and sex. We have three possibilities for the smoking variable: former, current, and never. We have only two possibilities for sex, male and female.

This leads to 6 possible combinations of data, namely:

\[
\begin{array}{ll}
\text{Sex} & \text{smoke} \\
M & \text{former} \\
M & \text{current} \\
M & \text{never} \\
F & \text{former} \\
F & \text{current} \\
F & \text{never} \\
\end{array}
\]

If we then calculate the all the dummy variables that correspond to these data, we obtain the following matrix (I’ve included the original variables at the beginning):

\[
\begin{array}{cccccccc}
\text{sex} & \text{smk} & \text{int} & M & F & \text{smf} & \text{smc} & \text{smn} \\
M & for & 1 & 1 & 0 & 1 & 0 & 0 \\
M & cur & 1 & 1 & 0 & 0 & 1 & 0 \\
M & nev & 1 & 1 & 0 & 0 & 0 & 1 \\
F & for & 1 & 0 & 1 & 1 & 0 & 0 \\
F & cur & 1 & 0 & 1 & 0 & 1 & 0 \\
F & nev & 1 & 0 & 1 & 0 & 0 & 1 \\
\end{array}
\]

One of dummy variables for sex is aliased with the intercept term, as is one of the smoking dummy variables. We need to remove one of each of dummy variables for both sex and smoking. In practice, SAS does this when it creates dummy variables via the CLASS statement. The sort order for the different values that can be taken determines which aliases dummy variable is removed. Thus, for the example above, the dummy variable smn would normally be left out, since the value never comes after for and cur in the normal sort order.
3.1 Studying Design Matrices with GLMMOD

The SAS procedure GLMMOD can be used to generate dummy variables, and other model effects, such as the combinations of dummy variables for interaction terms. Although the procedure is intended to produce the design matrix for SAS procedures that lack the CLASS statement, it can also be used to see what design matrix would result from certain model statements. For the examples presented in lecture, the SAS code looked like this:

```sas
data dagis;
  input sex $ wt height ones;
  cards;
  M 17 110 1
  M 15 105 1
  M 12 100 1
  F 15 104 1
  F 16 106 1
  F 14 102 1
run;
```

We then use Proc Glmmod to create the design matrix with dummy variables for sex.

```sas
proc glmmod;
  class sex;
  model wt = sex;
run;
```

We can also create the design matrix for more complicated models. Here we can look at the design matrix for a model with effects for sex, height, and a sex-height interaction.

```sas
proc glmmod;
  class sex;
  model wt = height sex sex*height;
run;
```

The next model is really the same model, but with parameters for the intercept and the slope of the line indicating the relationship between height and weight for each sex.
proc glmmod;
   class sex;
   model wt = sex sex*height /noint;
run;

The next example I discussed was produced with SAS code that looked like this:

data fake;
   input sex $ smk $ fake;
   cards;
   M F 100
   M C 100
   M N 100
   F F 100
   F C 100
   F N 100
run;

We can examine the main effects of sex and smoking status:

proc glmmod;
   class sex smk;
   model fake = sex smk;
run;

We can also examine the main effects plus an interaction term.

proc glmmod;
   class sex smk;
   model fake = sex smk sex*smk;
run;

We can also view the interaction model as six dummy variables describing the different combinations of sex and smoking and status.

proc glmmod;
   class sex smk;
   model fake = sex * smk ;
run;

Or we can view the interaction model as the effect of sex, and then the effects of smoking nested with sex.
proc glmmmod;
  class sex smk;
  model fake = sex smk(sex);
run;
4 Logistic Regression with Several Independent Variables

In practice, we have several independent variables. The simplest multivariate models might only have one additional variable, perhaps strata. Here is a simple example with two strata (Ahlbom, 1993). In this case we input the data with two additional variables. The first is a frequency variable indicating the number of lines with this pattern. The other one is strata. The original data appears in tabular form in table 4.

<table>
<thead>
<tr>
<th>CC status</th>
<th>Exposure</th>
<th>Ja</th>
<th>Nej</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fall</td>
<td>30</td>
<td>5</td>
<td>35</td>
<td></td>
</tr>
<tr>
<td>Kontroll</td>
<td>30</td>
<td>10</td>
<td>40</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>60</td>
<td>15</td>
<td>75</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>CC status</th>
<th>Exposure</th>
<th>Ja</th>
<th>Nej</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fall</td>
<td>30</td>
<td>225</td>
<td>255</td>
<td></td>
</tr>
<tr>
<td>Kontroll</td>
<td>10</td>
<td>90</td>
<td>100</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>40</td>
<td>315</td>
<td>355</td>
<td></td>
</tr>
</tbody>
</table>

Table 4: Stratified Data for a Case-Control Study

A SAS program to create these data might look like this:

* a simple example of logistic regression;
* for controlling for strata;
* see Ahlboms book Biostatistics for Epidemiologists p 114;

*options linesize=65 pagesize = 24;
options ls = 75 ps=60;

data st1;
input case freq strata exposed $ ones;
cards;
  1 30 1 ja 1
  0 30 1 ja 1
  1  5 1 no 1
Thus the first line indicate that there are a total of 30 observations with exposed cases in strata 1.

To perform an analysis with these data using GENMOD, one must first use the 'FREQ' statement to indicate that the variable freq represents the frequency of that observation. We then run two models, one with only strata, and another with both strata and exposed in the class statement (to create dummy variables for both) and in the model statement.

```plaintext
proc genmod;
  freq freq;
  class strata;
  model case/ones = strata/
    link = logit error =b;
run;

proc genmod;
  freq freq;
  class strata exposed;
  model case/ones = strata exposed/
    link = logit error =b;
run;
```

We can perform a likelihood ratio test (table 5) to determine the importance of the variable ‘exposed’ We do this by calculating the difference in -2 times the log-likelihood between the two models, and then comparing this amount to a $\chi^2$ distributed variable, with the number of degrees of freedom equal to the difference in degrees of freedom between the two models.

The difference between the two models is then checked against a $\chi^2$ distribution.

### 4.1 Another Numeric Example

Here we look at the esophageal cancer data from Breslow and Day (1980), considering only alcohol consumption in two groups. We fit two models, one with
agegrp and one with both agegrp and alcohol. The variable 'algrp' is created with two levels. Because 'REF' sorts after 'HI', the dummy variable for REF will be left out of the model.

* esophagus data from breslow and day vol 1 page 137;
* analysis with genmod;

data eso;
  input agegrp alcohol case number;
  atrisk = 1;
  if alcohol = 0 then Algrp = 'REF';
    else algrp = 'HI';
cards;
  1  1  1  1
  1  0  1  0
  1  1  0  9
  1  0  0 106
  2  1  1  4
  2  0  1  5
  2  1  0  26
  2  0  0 164
  3  1  1  25
  3  0  1  21
  3  1  0  29
  3  0  0 138
  4  1  1  42
  4  0  1  34
  4  1  0  27
  4  0  0 139
  5  1  1  19
  5  0  1  36
  5  1  0  18
  5  0  0  88
proc genmod;
  freq number;
  class agegrp;
  model case/atrisk = agegrp/
                 link=logit
                 error = binomial;
run;

proc genmod;
  freq number;
  class agegrp algrp;
  model case/atrisk = agegrp algrp /
                  link=logit
                  error = binomial;
run;
5 Deviance and Goodness of Fit

A model attempts to replace the observed $y$’s with fitted values. In general, these fitted values do not agree exactly with the observed values. It is quite easy to see that small discrepancies are permissible, whereas large ones are not.

We can think of model construction as a process where more and more variables are added to the LP. The simplest model has only one parameter, representing the mean value. The most complicated model has one parameter for each observation. The simplest model (the mean only) generally gives little insight into the processes behind the data. The most complicated model (often referred to as a saturated model) is also of little use, since it merely replaces the observations themselves with an equal number of regression parameters. Since the goal of our statistics is to condense the information, this full model is, in practice, of little use.

For a given set of data, there will be larger discrepancies for the simpler model, and smaller discrepancies for the more complicated models.

There are a number of different measures of goodness of fit but the ones most commonly used in GLM’s are deviance and Pearson $\chi^2$.

The deviance is formed from the logarithm of the ratio of two likelihoods. One of these is the likelihood of the current model. The other is the likelihood of the saturated model. Deviance ($D$) is calculated as follows:

$$D = -2 \log \left( \frac{\text{current}}{\text{saturated}} \right)$$

Where current and saturated are the likelihoods of the current and saturated models, respectively.

Note that we can manipulate this formula so that

$$D = -2 \log (\text{current}) + 2 \log (\text{saturated})$$

In practice, if we have single individuals in each strata, so that the number ’at risk’ in each observation is equal to 1, the likelihood of a saturated model is 1, and the deviance reported for the model equals 2 times the log likelihood of the current model. If we have grouped data, the likelihood of the saturated model is no longer 1. In this case, then the ’deviance’ reported by SAS is no longer equal to -2 times the log-likelihood.

These represent the two ways in which data for logistic regression can be entered. In practice, if we confine the likelihood ratio statistics that we calculate to the differences between two models, it matters little which way we enter the data. Where it makes a difference is in the interpretation of the residual deviance figure from a single model. If we have data observations consisting of many trials with a given number of outcomes as the dependent variable (ie the denominator in the model statement is not one), then the residual deviance is a reasonable
indicator of goodness of fit. If this is not the case, and the data consist of single trials, then the goodness of fit must be determined by some other means. One of these is the Hosmer-Lemeshow test, which is covered in their book (Hosmer & Lemeshow, 1989).

The other measure of goodness of fit is the Pearson \( \chi^2 \) statistic which takes the form

\[
\chi^2 = \sum \frac{(y - \hat{y})^2}{V(\hat{y})}
\]

where \( V(\hat{y}) \) is the variance function for the distribution. For the normal link, this is the residual sums of squares, but for the binomial or Poisson error, this is the original Pearson \( \chi^2 \) statistic.

Either of these can also be divided by the scale parameter in the case of over-dispersed data.
6 A Detailed Analysis with Analysis of Deviance

The esophageal cancer in Breslow and Day (Breslow & Day, 1980) actually has both alcohol and tobacco consumption in 4 different levels in addition to age. A detailed analysis would consider the effects of alcohol and tobacco separately. This might be followed by a model with both variables. The effects of age might be considered a nuisance variable, so it is included in all models. In the file 'eso.dat', I have coded, in addition to the pattern of the independent variables age-group, smoking, and alcohol, two number corresponding to the number of cases and controls with that pattern. The SAS code to read that data set and create a data set with one individual per line looked like this:

* esophagus data;
* 975 records;

* two levels of alcohol consumption;

data eso;
   infile 'eso.dat';
   input agegrp alcg tobg casen contn;
run;

data eso2 ;
   set eso;
   keep agegrp alcg tobg atrisk case;
   atrisk = 1;
   do i= 1 to casen;
      case = 1;
      output;
   end;
   do i= 1 to contn;
      case = 0;
      output;
   end;
run;

data eso3;
   set eso2;
   a4 = alcg;
   if a4 = 1 then a4 = 9;
   if tobg = 1 then tobg = 9;
if alcg = 1 then c_alc = 20;
if alcg = 2 then c_alc = 60;
if alcg = 3 then c_alc = 100;
if alcg = 4 then c_alc = 150;
run;

The variable alcg is the original grouped variable, with values from 1 (lowest level of consumption) to 4 (highest level). The variable a4 is the original grouped variable, with the lowest level recoded as 9 to make it the reference group. A variable c_alc is also created to represent the midpoint of the different values.

We then proceed to run the different models. First, a null model with only age.

```
proc genmod;
  class agegrp;
  model case/atrisk = agegrp / 
    noint link=logit error=b;
run;
```

Then, two models with age and tobacco or age and alcohol.

```
proc genmod;
  class tobg agegrp;
  model case/atrisk = agegrp tobg / 
    noint link=logit error=b;
run;
```

```
proc genmod;
  class a4 agegrp;
  model case/atrisk = agegrp a4 / 
    noint link=logit error=b;
run;
```

We could also fit a model to examine the interaction of alcohol and tobacco.

```
proc genmod;
  class a4 tobg agegrp;
  model case/atrisk = agegrp a4 tobg a4*tobg / 
    noint link=logit error=b;
run;
```

The next model examines the effect of alcg (without the class statement). This is equivalent to examining the 'trend' effect of alcohol.
proc genmod;
  class tobg agegrp;
  model case/atrisk = agegrp alcg tobg /
      noint link=logit error=b;
run;

The last model looks at the midpoints of alcohol level within each group.

proc genmod;
  class tobg agegrp;
  model case/atrisk = agegrp tobg c_alc /
      noint link=logit error=b;
run;

The resulting data can be summarized in an analysis of deviance table (table 6). Here deviance represents the 'residual deviance' for complete models, and the change in deviance when comparing models. The latter is compared to a $\chi^2$ distribution.

<table>
<thead>
<tr>
<th>Model</th>
<th>Description</th>
<th>Deviance</th>
<th>df</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>Agegrp Only</td>
<td>868.4</td>
<td>969</td>
</tr>
<tr>
<td>B</td>
<td>A + Alcohol (4 grps)</td>
<td>727.4</td>
<td>966</td>
</tr>
<tr>
<td></td>
<td>Change from A</td>
<td>141.0</td>
<td>3</td>
</tr>
<tr>
<td>C</td>
<td>A + Tobacco</td>
<td>831.8</td>
<td>966</td>
</tr>
<tr>
<td></td>
<td>Change from A</td>
<td>36.6</td>
<td>3</td>
</tr>
<tr>
<td>D</td>
<td>A + Alcohol + Tobacco</td>
<td>703.9</td>
<td>963</td>
</tr>
<tr>
<td></td>
<td>Change from B</td>
<td>23.5</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>Change from C</td>
<td>127.9</td>
<td>3</td>
</tr>
<tr>
<td>E</td>
<td>D + AlcxTob Interaction</td>
<td>698.4</td>
<td>954</td>
</tr>
<tr>
<td></td>
<td>Change from D</td>
<td>5.4</td>
<td>9</td>
</tr>
<tr>
<td>F</td>
<td>C + alcg (trend)</td>
<td>709.8</td>
<td>965</td>
</tr>
<tr>
<td></td>
<td>Change from C</td>
<td>122.0</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Change from D</td>
<td>-5.9</td>
<td>2</td>
</tr>
<tr>
<td>G</td>
<td>C + c_alc</td>
<td>709.0</td>
<td>965</td>
</tr>
<tr>
<td></td>
<td>Change from C</td>
<td>123.0</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Change from D</td>
<td>-5.1</td>
<td>1</td>
</tr>
</tbody>
</table>

Table 6: Analysis of Deviance from Esophageal Cancer Study
The effects of alcohol and tobacco are confounded. This is evident from the change in deviance that results from the addition of these variables. The effect of alcohol is less if tobacco is in the model, and the effect of tobacco is less if alcohol is in the model. There is no evidence of a statistical interaction between tobacco and alcohol.
7 Poisson Regression

7.1 Introduction

Poisson regression is very similar to logistic regression. It can be viewed as the limiting case as the proportion $p$ becomes smaller but the denominator becomes larger. True Poisson processes are rare, but Poisson models can still be fit to a variety of data.

For the random component, one assumes Poisson errors. Often 'counting' type processes have a Poisson distribution.

For the systematic component, one can use a mixture of categorical or continuous variables. They must have a linear relationship to each other (as discussed earlier), but that is true of all linear predictors.

The normal link function is the natural logarithm.

Thus, in Poisson regression, we relate the logarithm of the observed counts to the linear predictor. The predicted counts are permitted to have a Poisson distributed error distribution to allow for discrepancies between the observed counts and those predicted by the model.

One difference between the Poisson regression models and the other GLM’s is that there can be independent variables in the LP which have a known $\beta$. These are often referred to as the offset in a Poisson regression.

Common offsets would be correction factors for the following:

- The length of time over which observations are made.
- The number of persons who were at risk for the outcome.
- A combination of the above (ie the number of persons and the time over which they were observed. This quantity is often referred to as person-years.
- The number of cases expected, calculated from the person-years and the background rates, usually stratified by age and calendar year.
- The total area from which individuals were gathered (quadrats of varying size, for example).

The offset constitutes the denominator of a rate, but has to be log-transformed before it is included in the LP. We can see that if

$$\log \frac{\text{predicted}}{\text{denom}} = LP$$

then
\[ \log(\text{predicted}) - \log(\text{denom}) = LP \]

so

\[ \log(\text{predicted}) = \log(\text{denom}) + LP \]

Thus, we would want to include the logarithm of the offset in the LP, but to fix the value of $\beta$ at 1.0.

### 7.2 A Numeric Example of Poisson Regression

The first example is taken from McCullagh and Nelder (McCullagh & Nelder, 1989). The data on ship type, construction period, and operation period, followed by the number of months in operation and the number of ‘incidents’ constitute the data. Not all combinations are possible, so these observations are deleted. A variable consisting of the natural logarithm of the number of months in operation is created to use as the offset. Thus, we are examining incidents per month.

* ship damage data;
* McCullagh and Nelder, p 205;

```sas
data ships;
  infile 'ship.dat';
  input type $ con $ op $ months incidnt;
  if months = 0 then delete;
  logmo = log(months); * remember this is natural log-
  arithmetic in SAS;
run;
```

The procedure GENMOD is also used for Poisson regression, but a different link function (link=log) and error distribution (error=p) are used. In addition the offset is specified. Two additional options are given, which produce the Type1 (sequential) and Type3 (that variable last) changes in deviance.

```sas
proc genmod;
  class type con op;
  model incidnt = type con op /
    link = log
    error = p
    offset = logmo
    type1 type3;
  make 'ParmEst' out = est;
run;
```
The above procedure also uses the 'make' statement to create a new SAS data set that contains the parameter estimates. The 'ParmEst' indicates which data are to be exported, and out= specifies the new data set name. In this case the data are called 'est'.

These data can be further manipulated in SAS. Here we take create a new data set and start by copying the data set est. We then calculate the rate ratio by exponentiating the parameter estimate. 95% confidence intervals are also calculated making large sample assumptions.

```sas
data expo;
  set est;
  rr = exp(estimate);
  ll = exp(estimate - 1.96 * stderr);
  ul = exp(estimate + 1.96 * stderr);
run;

proc print;
  run;
```

We can also duplicate table 6.4 in McCullagh and Nelder by just looking at all possible combinations of type and con, without including the main effects first.

```sas
proc genmod data = ships;
  class type con;
  model incidnt = type*con /
    link = log
    error = p
    offset = logmo
    noint;
  make 'ParmEst' out = est;
run;

data expo;
  set est;
  rate = exp(estimate) * 1000;
  ll = exp(estimate - 1.96 * stderr);
  ul = exp(estimate + 1.96 * stderr);
  keep rate parm level1 level2 ll ul;
run;

proc print;
  run;
```

33
Another Example of Poisson Regression

The second example is data initially published by Scotto (Scotto et al., 1974), and used by Frome and Chekoway (Frome & Chekoway, 1985) in their paper on Poisson regression. It consists of the number of non-melanoma skin cancers from eight different age groups from two cities, Minneapolis and Dallas. The average population in these age groups is also available for both cities and logarithm is used as the offset in the Poisson regression. The value t is the same as that created by Frome and Chekoway, and represents a linear transform of the midpoint for each age group. This is also log-transformed.

First we read in the data:

```plaintext
* poisson regression example 1;
* skin cancer data initially published by Scotto et al (1974);
* and subsequently analyzed by Frome and Chekoway (1985);

data cancer;
  input city $ agegrp cases pop;
  agecont = (agegrp * 10) + 10;
  t = (agegrp*20 - 15)/35;
  logt = log(t);
  logpop = log(pop);
  rate = cases/pop;
  cards;
  minnes 1 1 172675
  minnes 2 16 123065
  minnes 3 30 96216
  minnes 4 71 92051
  minnes 5 102 72159
  minnes 6 130 54722
  minnes 7 133 32185
  minnes 8 40 8328
  dallas 1 4 181343
  dallas 2 38 146207
  dallas 3 119 121374
  dallas 4 221 111353
  dallas 5 259 83004
  dallas 6 310 55932
  dallas 7 226 29007
  dallas 8 65 7358
run;
```
We then fit the different models. First a null model with just an intercept term.

```
proc genmod;
  model cases = /
        link = log error=p offset = logpop;
run;
```

Then a model that fits only the effect of log(age) as a continuous variable.

```
proc genmod;
  model cases = logt /
        link = log error=p offset = logpop;
run;
```

We can also fit a model that only looks at the effect of age as a categorical variable.

```
proc genmod;
  class agegrp;
  model cases = agegrp /
        link = log
        error = p
        offset = logpop;
run;
```

We do the same for city, and fit a model examining the only this effect.

```
proc genmod;
  class city;
  model cases = city /
        link = log
        error = p
        offset = logpop;
run;
```

We can also look a multivariate models. The following is a model with the effects of city and logarithm(age) as a continuous variable.

```
proc genmod;
  class city;
  model cases = city logt /
        link = log
        error = p
        offset = logpop;
run;
```
We can also fit a model with both city and age as categorical variables.

```plaintext
proc genmod;
  class city agegrp;
  model cases = city agegrp /
    link = log
    error = p
    offset = logpop;
run;
```

We can also examining a model with an interaction between city and age.

```plaintext
proc genmod;
  class city agegrp;
  model cases = city*agegrp /
    link = log
    error = p
    offset = logpop ;
run;
```

The last model is also one with an interaction between city and age, where both are categorical variables. By leaving the intercept out of the model, the parameter estimates are the logarithm of the crude rate for that particular age group and city.

```plaintext
proc genmod;
  class city agegrp;
  model cases = city*agegrp /
    link = log
    error = p
    offset = logpop noint ;
run;
```

These results can be summarized in an analysis of deviance table (Table 7). Small differences between these results and those presented by Frome and Chekoway may be due to different methods used to estimate the regression parameters.
<table>
<thead>
<tr>
<th>Model</th>
<th>Description</th>
<th>Deviance</th>
<th>df</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>Intercept Only</td>
<td>2739</td>
<td>15</td>
</tr>
<tr>
<td>B</td>
<td>Intercept and log(t)</td>
<td>274.3</td>
<td>14</td>
</tr>
<tr>
<td></td>
<td>Change from A</td>
<td>2518.7</td>
<td>1</td>
</tr>
<tr>
<td>C</td>
<td>Intercept and Agegrp</td>
<td>267.7</td>
<td>8</td>
</tr>
<tr>
<td></td>
<td>Change from A</td>
<td>2525.3</td>
<td>7</td>
</tr>
<tr>
<td>D</td>
<td>Intercept and City</td>
<td>2572.3</td>
<td>14</td>
</tr>
<tr>
<td></td>
<td>Change from A</td>
<td>220.7</td>
<td>1</td>
</tr>
<tr>
<td>E</td>
<td>Intercept, City, and log(t)</td>
<td>16.6</td>
<td>13</td>
</tr>
<tr>
<td></td>
<td>Change from B</td>
<td>257.7</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Change from D</td>
<td>2555.5</td>
<td>1</td>
</tr>
<tr>
<td>F</td>
<td>Intercept, City, and Agegrp</td>
<td>7.9</td>
<td>7</td>
</tr>
<tr>
<td></td>
<td>Change from B</td>
<td>259.7</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Change from D</td>
<td>2564.4</td>
<td>1</td>
</tr>
<tr>
<td>G</td>
<td>Intercept + City*Agegrp</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

(Full Model)

Table 7: Analysis of Deviance from Non-melanoma Skin Cancer Analysis
8 A Working Example

This is an analysis published by me (Yuen et al., 1993). The data consisted of all breast cancer cases diagnosed in Sweden from 1958 to 1989, stratified by decennia of and month of birth. The number of individuals born during that particular time period (ie all those born in January during a particular decennia) were used as the denominator in the rates, and the natural logarithm was calculated to use as the 'offset'. Decennia was treated as a categorical variable, whereas month was used as both a categorical and a continuous variable. When used as a continuous variable (NMO) it was also sine and cosine transformed to allow for cyclical (yearly) patterns. Although the data were originally analyzed with GLIM, it is also easily done with GENMOD in SAS. The primary difference is to recode January so that it is the reference month in the GENMOD analyses.

The data were read in with the following program.

```
data simple;
  infile 'sbreast.dat';
  input dec month case pop;
  lpop = log(pop);
  nmo = month; *numeric month;
  smo = sin(month*3.1416/6.0);
  cmo = cos(month*3.1416/6.0);
  if month = 1 then month = 99;
  * Jan was reference month in article;
run;
```

In this context, a null model would have to at least include decennia to allow for the different effects of age and time.

```
proc genmod;
  title 'Model A Table II';
  class dec;
  model case = dec /
    link = log
    error = p
```
A model with month as a categorical variable was fit next. The parameters appear in Table III in the original article. Notice that the noint option was used in SAS to allow us to interpret the parameter estimates for decennia as the crude risk for individuals born in January during those decennia to develop breast cancer as of 1989??.

```plaintext
proc genmod;
  title 'Model B Table II, Table III';
  class dec month;
  model case = dec month /
      link = log
      error = p
      offset = lpop
      noint;
run;
```

We fit month as a continuous variable (specific to each decennia) to investigate the trend that the risk is higher for those born in the months later in the year in the oldest decennia, since more of these individuals have survived to 1958. In the middle decennia, this is not an important factor, but in the younger cohorts, an increased risk for those born earlier in the year corrects for the fact that these individuals are older.

```plaintext
proc genmod;
  title 'Model C';
  class dec;
  model case = dec dec*nmo / 
      link = log
      error = p 
      offset = lpop;
```

We can also look for a general fluctuation throughout the year by fitting a common smo and cmo effect.

```plaintext
proc genmod;
  title 'Model D';
  class dec;
  model case = dec dec*nmo smo cmo /
      link = log
      error = p
      offset = lpop;
run;
```
The full model investigates the possibility of an interaction between the cyclic components and decennia. It addresses the question *Does the cyclic effect vary over time?* and is fit with the last model.

```plaintext
proc genmod;
  title 'Model E, Table IV';
  class dec;
  model case = dec dec*month dec*smo dec*cmo /
                   link = log
                   error = p
                   offset = lpop
                   noint;
run;
```
9 Goodness of Fit and Over-dispersion

There are several areas that deserve attention in assessing goodness of fit. If we are doing logistic regression, and we have our data arranged so that there are many ’trials’ per observation in the data set, we can then assess the overall goodness of fit of the data by examining the residual deviance and the Pearson $\chi^2$ values, which should have $\chi^2$ distribution (with the appropriate degrees of freedom).

If all the available independent variables have been included in the model, and the deviance or Pearson $\chi^2$ is still too large, it may be that the data is over-dispersed. In other words the variance of the $Y$ is greater than binomial or Poisson distribution, for logistic and Poisson regression respectively. Identifying reasons for over-dispersion is dependent on subject matter. One possible reason is insufficient information. In human epidemiological studies, it is rare that all information on all possible risk factors can be collected, so that some over-dispersion can be expected.

Other things such as clustering or biased sampling schemes can lead to both over and underestimation of numbers of individuals.

In practice, we account for the over-dispersion by assuming that the parameters we estimate are not changed, but that the over-dispersion has increased the variance more than is predicted by either binomial or Poisson distributions. A method for doing this was described by (Breslow, 1984).

The easiest estimates of this additional variance is via the dscale or pscale option in the GENMOD model statement. With this option, the model is fit as normal, but the variance estimates (the curviness of the likelihood surface), are adjusted.

The scale parameter, which has previously been 1.0, is now different, representing the additional variation. The square root of the deviance or Pearson $\chi^2$ divided by the number of degrees of freedom, is used as the scale parameter.

The likelihoods calculated in this manner are not true likelihoods, but quasi-likelihoods.

Systematic variation can be examined in a manner similar to linear linear models, via residual plots. In this cases we would use either the deviance or Pearson $\chi^2$ residuals, not the raw residuals. This is easily arranged via SAS.

In this example, we attempt to correct for over-dispersion in the ship data using the DSscale and PSscale options of SAS.

First we read in the data.

```sas
* ship damage data;
* McCullagh and Nelder, p 205;

options linesize=80 ps=60;
```
data ships;
  infile 'ship.dat';
  input type $ con $ op $ months incidnt;
  if months = 0 then delete;
  logmo = log(months);
run;

Then we fit a model in the normal fashion.

proc genmod;
  Title 'Ship Accident Data';
  class type con op;
  model incidnt = type con op /
      link = log
      error = p
      offset = logmo
  ;
run;

In this model, the residual deviance and the Pearson $\chi^2$ are 38.69 and 42.27, respectively, with 25 df. This indicates a possibility of a poor fit due to overdispersed data. We can correct our estimates with this by using the DSCALE or PSCALE option in SAS.

proc genmod;
  Title 'Ship Accident Data with DSCALE';
  class type con op;
  model incidnt = type con op /
      link = log
      error = p
      offset = logmo
dscale;
run;

proc genmod;
  Title 'Ship Accident Data with PSCALE';
  class type con op;
  model incidnt = type con op /
      link = log

42
error = p
offset = logmo
PSCALE;
run;

Note that this 'decreases' the $\chi^2$ values for the individual parameters, and increases the estimates of the standard error, which will also increase the range of the confidence intervals. For example, the logarithm of the rate ratio of incidents for type A ships compared to type E ships is -0.3256. The standard error is 0.2359 in the regular analysis, and increases to 0.2935 and 0.3067 with the dscale and pscale options, respectively.
References


