# Generalized Linear Models

Up to now, we have been considering models when the data are normally distributed. Not all data are normally distributed. For example, when an outcome is measured as a success or failure, or when we count the number of events over a fixed period. Generalized linear models (GLM) are used to fit fixed effect data to certain types of data that are not normally distributed.

Generalized – not limited to normally distributed data
Linear – models use a linear combination of variables to 'predict' the response

Measurements are made in three scales

1. Nominal classification
    a. binary or dichotomous – only two categories: male, female; dead, alive
    b. more than two categories: red, green, blue; yes, no, do not know, not applicable.
2. Ordinal classification – natural order or ranking between categories
    a. young, middle-aged, old
    b. weight grouping: <100 kg, 101-150 kg, 151-200 kg, >200 kg.
3. Continuous measurement – where observation may fall anywhere on a continuum
    a. interval measurement
    b. ratio scale measurement – well defined zero

**Distributions**

Up to now, we have been using the following linear model form that assumes that **e** is identically and independently distributed (iid) with the Normal distribution $\mathbf{N}(0,\sigma^2)$.

$\mathbf{y = X\beta + e}$

The Normal distribution is from the Exponential family of distributions. Many of the "nice" properties of the Normal distribution are shared by other members of the Exponential family. Two other distributions in the Exponential family that are used extensively in GLM are the **binomial** and the **Poisson** distributions. These distributions are described as 'one-parameter' distributions because a single parameter describes the distribution.

The general form of the density function that any exponential family distribution can be written

$$f(y;\theta,\phi) = \exp\{[y\theta - b(\theta)]/a(\phi) + c(y,\phi)\}$$

where

$\theta$ = a location parameter (not necessarily the mean)
$\phi$ = a dispersion parameter(only in distributions with 2 parameters, N)

The general form for the one parameter distributions is

$$f(y;\theta) = \exp\{[y\theta - b(\theta)]/a + c(y)\}$$

where $a$ is now a constant.

The mean and variance can be written in terms of $a$ and $b$

mean(y)=μ= $b'(\theta)$
var(y)=$ab''(\theta)$

alternatively,

var(y)= $ab''(b'^{-1}(\mu))$

We'll now look at the specific Exponential distributions of interest.

**Binomial distribution**

The binomial distribution is used for binary data. The **Bernoulli** distribution is a special case of the binomial distribution that is used when the observations have one of two possible outcomes: 1="success" or 0="failure".

The Bernoulli distribution is a one parameter distribution, where from the general form of the distribution function

$a$=1
$b(\theta)$ =log(1+exp($\theta$))
$c(y)$=1

mean(y)=μ= $b'(\theta)$=$(1+\exp(\theta))^{-1}$

var(y)=μ(1-μ)

When the observations are recorded as the proportion of successes, then the binomial distribution is used.

**y=z/n**

where **z** is the number of successes and **n** is the number of tries.

The parameters for the general form of the distribution function for the binomial are

$a$=1/n
$b(\theta)$ =log(1+exp($\theta$))
$c(y)$=log[n!/((ny)!(n-ny)!)]

mean(y)=μ= $b'(\theta)=(1+\exp(\theta))^{-1}$

var(y)=μ(1-μ)/n

**Poisson distribution**

The Poisson distribution is used to model count data. The number of flies trapped in 24 hours.

y=z

where z is the count.

The parameters for the general form of the distribution function for the Poisson are

$a$=1
$b(\theta)$ =exp($\theta$)
$c(y)$=-log(y!)

mean(y)=μ= exp($\theta$)

var(y)=μ

If the underlying scale for the count data varies with each observation, then the data would be distributed Poisson with offset. Examples are observations made with varying time periods or size of geographical region. So the data must be "scaled" and the scale variable is called the offset ($t$).

y=z/t

mean(y)=μ

var(y)=μ/t.

**GLM and Canonical link functions**

For the GLM the model is written as follows

$\mathbf{y = \mu + e}$

$\boldsymbol{\mu = X\beta}$

var(**y**)=var(**e**)=**V**

This is to allow for the **μ** and the **Xβ** to be related by a "link function"

$g(\boldsymbol{\mu}) = \mathbf{X}\boldsymbol{\beta} = \theta$

where $\theta$ is the linear component of the General exponential density.

The link function can be thought of as a method of mapping the response data from their scale of observation to the real scale $(-\infty,+\infty)$.

The **canonical link function** is given by

$g = b^{-1}$

where $b$ is obtained from the general exponential density form. The following table presents the canonical link functions for the distributions covered

| Distribution | $g(\mu)=b^{-1}(\mu)$ | Name |
|---|---|---|
| Bernoulli | log(μ/(1-μ)) | Logit |
| Binomial | log(μ/(1-μ)) | Logit |
| Poisson | log(μ) | Log |
| Poisson with offset | log(μ) | Log |

The variance can be written in terms of μ and the canonical link function $g$

$\text{var}(y)=ag^{-1}(\mu)$

**The variance matrix**

var(**y**)=var(**e**)=**V**

Fixed effect model assumes that the observations are uncorrelated, therefore the variance matrix is diagonal.

Diagonal terms=variances of each observation

For six Bernoulli observations where $\mu = (\mu_1, \mu_2, \mu_3, \mu_4, \mu_5, \mu_6)'$

$$V = \begin{pmatrix} \mu_1(1-\mu_1) & 0 & 0 & 0 & 0 & 0 \\ 0 & \mu_2(1-\mu_2) & 0 & 0 & 0 & 0 \\ 0 & 0 & \mu_3(1-\mu_3) & 0 & 0 & 0 \\ 0 & 0 & 0 & \mu_4(1-\mu_4) & 0 & 0 \\ 0 & 0 & 0 & 0 & \mu_5(1-\mu_5) & 0 \\ 0 & 0 & 0 & 0 & 0 & \mu_6(1-\mu_6) \end{pmatrix}$$

The variances are different for each observation.

The variance can be written in the general form **V=AB**

where

**A**=diag{$a_i$}
**B** =diag{$g^{r1}(\mu_i)$}

So for the **V** for the Bernoulli, **A** is an identity matrix, while for the binomial distribution

$$A = \begin{pmatrix} 1/n_1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1/n_2 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1/n_3 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1/n_4 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1/n_5 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1/n_6 \end{pmatrix}$$

and **B** is the same as the **V** matrix for the Bernoulli.

For the Poisson distribution, **A** = I and

$$V = \begin{pmatrix} \mu_1 & 0 & 0 & 0 & 0 & 0 \\ 0 & \mu_2 & 0 & 0 & 0 & 0 \\ 0 & 0 & \mu_3 & 0 & 0 & 0 \\ 0 & 0 & 0 & \mu_4 & 0 & 0 \\ 0 & 0 & 0 & 0 & \mu_5 & 0 \\ 0 & 0 & 0 & 0 & 0 & \mu_6 \end{pmatrix}$$

**Dispersion and the dispersion parameter**

Count data are normally assumed to be from a Poisson distribution, where the mean and the variance are the same. However, this is probably not true in biological data. Birds, insects, and plants have a tendency to be more aggregated or clustered together than what would be expected from a Poisson model, so that the variance tends to be larger than the mean. Overdispersion is when the variance is larger than expected under a given model.

The dispersion parameter, $\phi$, can increase or decrease the variance in the model from the observation variances

$$V = \phi \mathrm{AB}$$

The implications of using a dispersion parameter will be discussed later.

**Logit link function**

log(μ/(1-μ)) which models the log of an odds,

is the canonical link function for Bernoulli and binomial distributions.

Using this link function referred to as **logistic regression** analyses. Because the results are obtained on a log scale, they are not easy to interpret directly.

Write the model

$\log(\mu/(1-\mu))=a+bx$

where $x$ is the indicator variable denoting treatment(=1, if treatment A, =0, if treatment B).

If the probability for success under treatment A is $p_A$ then

$\log(p_A/(1-p_A))=a+b$

Similarly, if the probability for success under treatment B is $p_B$ then

$\log(p_B/(1-p_B))=a$

If we want to look at the difference in response from treatments A and B

$\log(p_A/(1-p_A))-\log(p_B/(1-p_B))=a+b-a=b$

$$\log(p_A/(1-p_A))-\log(p_B/(1-p_B))=\log\frac{(p_A/(1-p_A)}{p_B/(1-p_B)}=b$$

If we exponentiate

$$\frac{p_A/(1-p_A)}{p_B/(1-p_B)}=e^b$$

The numerator is the odds of success on treatment A, the denominator is the odds of success on treatment B, therefore, this is called an **odds ratio** end $e^b$ is estimate of the odds ratio.

**The log link function**

log(μ) which models relative rates,

is the canonical link function for Poisson distributions.

Using this link function is referred to as loglinear. Just as with logistic regression, the results are not easy to interpret directly.

Write the model

$\log(\mu)=a+bx$

where $x$ is the indicator variable denoting treatment(=1, if treatment A, =0, if treatment B).

If the rate of an event under treatment A $r_A$ then

$\log(r_A)=a+b$

Similarly, if the rate of an event under treatment B is $r_B$ then

$\log(r_B)=a$

If we want to look at the difference in response from treatments A and B

$\log(r_A)- \log(r_B)=a+b-a=b$

$$\log(r_A)-\log(r_B)=\log\frac{(r_A)}{(r_B)}=b$$

If we exponentiate

$$\frac{(r_A)}{(r_B)}=e^b$$

which is the **relative rate or RR** or the ratio of two event rates.

If this log link function is used to analyze binary data, then the rates become risks and the RR stands for **relative risks**.

## Examples of Generalized Linear Models

### The logit function (or logistic regression)

### The Challenger Shuttle O-ring example (from SAS for Linear Models)

In 1986, the space shuttle Challenger blew up shortly after take off. Investigators focused on the suspected association between O-ring failure and cold temperatures. Data documenting the presence or absence of primary O-ring thermal distress on previous space shuttle launches prior to the disaster were reproduced in Agresti (1996) and are located in the dataset oring.dat.

The following SAS code reads in the data and prints it

```
options ps=80 ls=64;
data m;
 infile 'C:\users\kathy\statistics department\statistics 892- mixed
models\oring.dat';
 input temp td no_td total;
proc print;
run;
```

The output is

| Obs | temp | td | no_td | total |
|---|---|---|---|---|
| 1 | 53 | 1 | 0 | 1 |
| 2 | 57 | 1 | 0 | 1 |
| 3 | 58 | 1 | 0 | 1 |
| 4 | 63 | 1 | 0 | 1 |
| 5 | 66 | 0 | 1 | 1 |
| 6 | 67 | 0 | 3 | 3 |
| 7 | 68 | 0 | 1 | 1 |
| 8 | 69 | 0 | 1 | 1 |
| 9 | 70 | 2 | 2 | 4 |
| 10 | 72 | 0 | 1 | 1 |
| 11 | 73 | 0 | 1 | 1 |
| 12 | 75 | 1 | 1 | 2 |
| 13 | 76 | 0 | 2 | 2 |
| 14 | 78 | 0 | 1 | 1 |
| 15 | 79 | 0 | 1 | 1 |
| 16 | 81 | 0 | 1 | 1 |

Where temp is the temperature at launch, td are the number of launches where thermal distress occurred at that temperature, no_td are the number of launches where no thermal distress occurred at that temperature, and total are the total number of launches at that temperature.
It appears that the frequency of thermal distress is greater at lower temperatures, so we would want to fit a model for which the probability of thermal distress decreases as temperature increases. The probability is bounded by 0 and 1, so the logit link would be one reasonable alternative, so that the mean of the sample proportion is $y/N_i$ for the $\mu_i$ parameter that we want to estimate.

So the simplest model for these data is

$\log(\mu_i/(1-\mu_i))=a+bx_i$

This model can be fit with the following SAS code:

```
proc genmod;
model td/total=temp / link=logit dist=binomial type1;
run;
```

Note that the response variable is a ratio of the number of outcomes of interest (in our case, the number of thermal distressed O-rings) over the total number of outcomes. The independent variable TEMP is a direct regression variable. You can also have class variables. Just like with PROC GLM, if a variable appears in the CLASS statement, it is treated as a class variable. We have also specified the link and the distribution of the response variable. If the link and/or distribution are not specified, the defaults are logit and binomial. The type1 option gives the likelihood ratio test statistics for the hypotheses based on Type 1 estimable functions. This is an alternative hypothesis test for zero slope. Because we only have one independent variable, the same likelihood ratio test would occur no matter which type we used.

These are a selection of the results generated.

Model Information

| | |
|---|---|
| Data Set | WORK.M |
| Distribution | Binomial |
| Link Function | Logit |
| Response Variable (Events) | td |
| Response Variable (Trials) | total |
| Number of Observations Read | 16 |
| Number of Observations Used | 16 |
| Number of Events | 7 |
| Number of Trials | 23 |

This information tells you about the general model information and data information. Always check this to make sure that you have specified the model correctly and the data are what you expect.

```
          Criteria For Assessing Goodness Of Fit

  Criterion                 DF           Value        Value/DF

  Deviance                  14         11.9974          0.8570
  Scaled Deviance           14         11.9974          0.8570
  Pearson Chi-Square        14         11.1303          0.7950
  Scaled Pearson X2         14         11.1303          0.7950
  Log Likelihood                      -10.1576

                      Standard      Wald 95%      Chi-
Parameter DF Estimate   Error Confidence Limits Square Pr > ChiSq

Intercept  1  15.0429  7.3786   0.5810  29.5048  4.16    0.0415
temp       1  -0.2322  0.1082  -0.4443  -0.0200  4.60    0.0320
Scale      0   1.0000  0.0000   1.0000   1.0000
NOTE: The scale parameter was held fixed.
               LR Statistics For Type 1 Analysis

                                          Chi-
   Source         Deviance        DF     Square    Pr > ChiSq

   Intercept       19.9494
   temp            11.9974         1       7.95        0.0048
```

The goodness of fit information is how well the data fit the distribution used. The **deviance** is used to check the fit of the model. It is compared to a $\chi^2_{14}$. The table value for a $\chi^2_{14}$ at $\alpha$=.25 is 17.12. The scaled deviance and Pearson Chi-Square are relevant to lack of fit resulting from overdispersion, so not of interest here. Both the intercept and estimated slope for temperature were different from zero.

### The log link function (or loglinear model)

This experiment analyzes imperfection rates for two processes used to fabricate wafers for computer chips. There are two Treatments A and B, each applied to 10 wafers. The data are located in wafer.dat and are from Agresti (1996). The following SAS code reads in the data and prints it.

```
options ps=80 ls=64;
data m;
 infile 'C:\users\kathy\statistics department\statistics 892- mixed
models\wafer.dat';
 input treat $ count;
proc print;
run;
```

```
                  Obs    treat    count

                    1      A        8
                    2      A        7
                    3      A        6
                    4      A        6
```

| | | |
|---|---|---|
| 5 | A | 3 |
| 6 | A | 4 |
| 7 | A | 7 |
| 8 | A | 2 |
| 9 | A | 3 |
| 10 | A | 4 |
| 11 | B | 9 |
| 12 | B | 9 |
| 13 | B | 8 |
| 14 | B | 14 |
| 15 | B | 8 |
| 16 | B | 13 |
| 17 | B | 11 |
| 18 | B | 5 |
| 19 | B | 7 |
| 20 | B | 6 |

So the simplest model for these data is

$\log(\mu_i)=a+bx_i$

This model can be fit with the following SAS code:

```
proc genmod;
class treat;
model count=treat / dist=poisson type1;
run;
```

The GENMOD Procedure

Model Information

| | |
|---|---|
| Data Set | WORK.M |
| Distribution | Poisson |
| Link Function | Log |
| Dependent Variable | count |

| | |
|---|---|
| Number of Observations Read | 20 |
| Number of Observations Used | 20 |

Again, this information tells you about the general model information and data information.

Class Level Information

| Class | Levels | Values |
|---|---|---|
| treat | 2 | A B |

Criteria For Assessing Goodness Of Fit

| Criterion | DF | Value | Value/DF |
|---|---|---|---|
| Deviance | 18 | 16.2676 | 0.9038 |
| Scaled Deviance | 18 | 16.2676 | 0.9038 |
| Pearson Chi-Square | 18 | 16.0444 | 0.8914 |

| | | |
|---|---|---|
| Scaled Pearson X2 | 18 | 16.0444 | 0.8914 |
| Log Likelihood | | 138.2221 | |

The goodness of fit information tells us that the fit of the model is satisfactory.

Analysis Of Parameter Estimates

| Parameter | | DF | Estimate | Standard Error | Wald 95% Confidence Limits | |
|---|---|---|---|---|---|---|
| Intercept | | 1 | 2.1972 | 0.1054 | 1.9906 | 2.4038 |
| treat | A | 1 | -0.5878 | 0.1764 | -0.9335 | -0.2421 |
| treat | B | 0 | 0.0000 | 0.0000 | 0.0000 | 0.0000 |
| Scale | | 0 | 1.0000 | 0.0000 | 1.0000 | 1.0000 |

Analysis Of Parameter Estimates

| Parameter | | Chi-Square | Pr > ChiSq |
|---|---|---|---|
| Intercept | | 434.50 | <.0001 |
| treat | A | 11.11 | 0.0009 |
| treat | B | . | . |
| Scale | | | |

NOTE: The scale parameter was held fixed.

LR Statistics For Type 1 Analysis

| Source | Deviance | DF | Chi-Square | Pr > ChiSq |
|---|---|---|---|---|
| Intercept | 27.8570 | | | |
| treat | 16.2676 | 1 | 11.59 | 0.0007 |

From the parameter estimates we can obtain the odds ratio or relative rate using the following

$$\frac{(r_A)}{(r_B)} = e^b$$

where $b$ = -.5878, which is the estimate for Treatment A.

The ratio of imperfections for those wafers treated with A relative to those wafers treated with B is exp(-.5878)=.5555. The rate of imperfection for those wafers treated with A is about half that for those wafers treated with B and this difference is significant.