Lecture 14: GLM Estimation and Logistic Regression

Dipankar Bandyopadhyay, Ph.D.

BMTRY 711: Analysis of Categorical Data Spring 2011

Division of Biostatistics and Epidemiology

Medical University of South Carolina

Fitting GLMs

Suppose we have a GLM with a parameter vector

$$\vec{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix}$$

and we want the ML estimators of $\vec{\beta}$.

When we use GLMs, we typically have a non linear model.

For simplicity, denote $\hat{\boldsymbol{\beta}}$ as the vector of MLEs.

Iterative Solutions

Iterative Solutions to non-linear equations follow this algorithm:

1. A seed value is selected (initial guess for $\hat{\boldsymbol{\beta}}$)
2. Using a polynomial approximation of the likelihood, a second "guess" is obtained
3. The difference, C, between guess i and $i+1$ is calculated ($C = \beta^{(i+1)} - \beta^{(i)}$)
4. Once the difference $C < k$ where $k =$ "convergence criterion" (say 0.0001) then the estimate $\beta^{(i+1)} = \hat{\boldsymbol{\beta}}$

Note: when β is a vector, the difference $\beta^{(i+1)} - \beta^{(i)}$ yields a vector of c_i's where c_i is the convergence criterion for the i^{th} element of $\vec{\beta}$.

Convergence could be reached when all $|c_i| < k$ or when the $\sum_i |c_i| < k$

Iterative MLEs

In general, there are two popular iterative methods for estimating the parameters of a non-linear equations.

1. Newton-Raphson Method
2. Fisher's Scoring Method

Both take on the same general form and differ only in the variance structure.

Recall the Wald (non-null standard error) and the Score (null standard error).

The Wald and Score tests will be similar to the Newton-Raphson and Fisher's Scoring methods.

Score and Information

- An exponential class distribution can be written in the form

$$f(y_i; \theta) = \text{exp}[a(y_i)b(\theta) + c(\theta) + d(y_i)]$$

- Note: $a(\cdot), b(\cdot) \ldots$ are different functions than introduced in Lecture 11 (for example $c(\theta)$ (for this notation) equals $\log a(\theta)$ in Lecture 11 notation)
- So, $l(\cdot)$ can be written as

$$\begin{aligned} l &= \log L \\ &= \sum_{i=1}^{n} \log \left(\text{exp}[a(y_i)b(\theta) + c(\theta) + d(y_i)]\right) \\ &= \sum_{i=1}^{n} \{a(y_i)b(\theta) + c(\theta) + d(y_i)\} \end{aligned}$$

Score equations

- The “score” is $U = dl/d\theta$, so

$$\begin{aligned} U &= \sum_{i=1}^{n} a(y_i)\frac{d\ b(\theta)}{d\theta} + \frac{d\ c(\theta)}{d\theta} \\ &= \sum_{i=1}^{n} a(y_i)b'(\theta) + c'(\theta) \end{aligned}$$

$$Var(U) = E(U^2) = -E(U')$$

- where $Var(U)$ is the information.
- (for this class, assume these are definitions. Note that E(U) = 0)
- When Y is of the exponential class, the $\partial l/\partial\theta$ can be simplified.

Score Equations for Exponential Class Variables

$$U = \sum_{i=1}^{n} \frac{\partial E(Y_i|X_i)}{\partial \beta} \left[\frac{Y_i - E(Y_i|X_i)}{Var(Y_i|X_i)} \right]$$

For example,
Suppose $Y_i \sim Poi(\mu_i)$

$$E(Y_i \Big| X_i) = \mu_i = e^{X_i' \beta}$$

$$Var(Y_i\,|X_i) = \mu_i$$

$$\frac{\partial E(Y_i\,|X_i\,)}{\partial \beta} = X_i' e^{X_i' \beta}$$

$$= X_i' \mu_i .$$

So,

$$U = \sum_{i=1}^{n} X_i \mu_i \left[\frac{Y_i - \mu_i}{\mu_i} \right] = \sum_{i=1}^{n} X_i \left[Y_i - \mu_i \right]$$

Estimation

The MLE's are obtained by solving the score equations, U equal to zero.

$$U = \sum_{i=1}^{n} \frac{\partial E(Y_i|X_i)}{\partial \beta} \left[\frac{Y_i - E(Y_i|X_i)}{Var(Y_i|X_i)} \right] = 0$$

Note: U is actually a vector of the p parameters of β.

For the j^{th} parameter,

$$U_j = \sum_{i=1}^{n} \frac{\partial E(Y_i|X_i)}{\partial \beta_j} \left[\frac{Y_i - E(Y_i|X_i)}{Var(Y_i|X_i)} \right] = 0$$

Newton-Raphson vs. Fisher's Scoring

$$\widehat{\beta}^{(m)} = \widehat{\beta}^{(m-1)} - \left[\frac{\partial^2 l}{\beta_j \beta_k}\right]^{-1}_{\beta=\widehat{\beta}^{(m-1)}} U^{(m-1)}$$

What makes the Newton-Raphson unique is that $\left[\frac{\partial^2 l}{\beta_j \beta_k}\right]^{-1}_{\beta=\widehat{\beta}^{(m-1)}}$ is the variance estimated under the alternative (like a Wald test).

Fisher's Scoring uses the

$$E\left[\frac{\partial^2 l}{\beta_j \beta_k}\right]$$

Or the "expectation" of the "Hessian matrix".

Definition: $\left[\frac{\partial^2 l}{\beta_j \beta_k}\right]$ is called the Hessian.

For Fisher's Scoring, let

$$\iota_{jk} = E[U_j U_k] = E[\frac{\partial l}{\partial \beta_j}\frac{\partial l}{\partial \beta_k}]$$

With some work, it can be shown that

$$E[\frac{\partial l}{\partial \beta_j}\frac{\partial l}{\partial \beta_k}] = -E[\frac{\partial^2 l}{\partial \beta_j \beta_k}]$$

Therefore, Fisher's Scoring is similar to regular Score test, but it still plugs the estimates of $\widehat{\beta}^{(m-1)}$ into the iterative solutions.

Iterative Solutions by Hand

We will not be taking iterative solutions by hand.

In SAS,

1. SAS PROC GENMOD uses the Newton-Raphson method (by default)
2. SAS PROC LOGISTIC uses Fisher's Scoring method (by default)

Both give similar results. The parameter estimates will be close to identical, but in some cases, the standard errors may differ. In general, people do not lose sleep over the two methods.

Now, on to more about Logistic Regression

Logistic Regression for an $R \times 2$ tables

Consider the following toxicity dataset, with the rows fixed (or conditioned on) by design, i.e., the distribution of the observed data are a product of 4 binomials

```
                                 Toxicity

                         | SOME    | NONE   | Total
               ---------+--------+-------+
                       1 |       8 |     92 |  100
               ---------+--------+-------+
   Dose (mg)         10 |      15 |     85 |  100
               ---------+--------+-------+
                     100 |      22 |     78 |  100
               ---------+--------+-------+
                    1000 |      26 |     74 |  100
               ---------+--------+-------+
               Total             71     329    400
```

(note in a previous lecture, we looked at a similar data set - this one is different)

- The row margins $E(Y_j.) = y_j. = m_j.$ is fixed by design (or conditioned on), and the parameters of interest are the of the probabilities of ‘SOME’ toxicity, given the dose j.
- It makes sense to analyze the data as they arose, and to directly model

$$P(\text{Some Toxicity}|\text{dose level})$$

- In general, suppose we denote the column variable by

$$Y = \begin{cases} \text{1 if success (column 1)} \\ \text{0 if failure (column 2)} \end{cases} .$$

  and the row variable by X, where X can take on values $1, ..., R$.
- We are interested in modelling

$$P[Y = 1|X = j] = p_j$$

- For the i^{th} individual in row j, we let

$$Y_{ij} = \begin{cases} \text{1 if success} \\ \text{0 if failure} \end{cases},$$

  $i = 1, ..., n_j$.

- Then, the individuals in row j have independent bernoulli observations,

$$Y_{ij} \sim Bern(p_j)$$

  and the number of successes on treatment j is binomial:

$$Y_j = \sum_{i=1}^{n_j} Y_{ij} \sim Bin(n_j, p_j),$$

  for $j = 1, ..., R$.

Question of interest:

Does the probability of success vary with X ?

- Let x_j be the ordinal value or 'score' of level j of X (it could equal j or the dose level, or other values as described previously).
- The logistic regression model is

$$P[Y = 1|X = j] = p_j = \frac{e^{\beta_0+\beta_1 x_j}}{1 + e^{\beta_0+\beta_1 x_j}}$$

  where β_0 and β_1 are parameters.
- Note, if $\beta_1 = 0$, then

$$P[Y = 1|X = j] = \frac{e^{\beta_0}}{1 + e^{\beta_0}},$$

  for all x_j which is not a function of x_j, i.e.,

$$P[Y = 1|X = j] = P[Y = 1]$$

  does not change with x_j, and, Y and X are said to be independent.
- Thus, our main interest will be testing

$$\text{H}_0 : \beta_1 = 0.$$

Assigning ‘Scores’

- When looking for a ‘trend’ in the proportions, one may consider using different sets of scores for X

$$x_1 \leq x_2 \leq ... \leq x_R$$

- In this example

```
                          Toxicity

                   | SOME   | NONE  | Total
          ---------+--------+-------+
                 1 |      8 |    92 |  100
          ---------+--------+-------+
Dose (mg)       10 |     15 |    85 |  100
          ---------+--------+-------+
               100 |     22 |    78 |  100
          ---------+--------+-------+
              1000 |     26 |    74 |  100
          ---------+--------+-------+
          Total          71     329    400
```

Power of ‘Cochran-Armitage’ trend test

Two possible sets of scores are;

$$(1, 10, 100, 1000)$$

or

$$[\log_{10}(1), \log_{10}(10), \log_{10}(100), \log_{10}(1000)] = [0, 1, 2, 3]$$

- In general, when you assign scores and use the Cochran-Armitage trend test, a valid question is:
- 1. Will any set of scores

$$x_1 \leq x_2 \leq ... \leq x_R$$

be OK ?

- The answer is:
  Under the null

$$H_0 : p_j = \left(\frac{e^{\beta_0}}{1 + e^{\beta_0}} \right),$$

  any set of scores will give you a valid test (Type I error OK under the null).

- However, some scores are more powerful to detect departures from the null hypothesis in favor of the alternative

H$_A$: there is a trend in p_j with dose

- In particular, the most powerful ’scores’ to assign are the ones of the true model

$$p_j = \left(\frac{e^{\beta_0 + \beta_1 x_j}}{1 + e^{\beta_0 + \beta_1 x_j}} \right),$$

  i.e.,

$$x_1 \leq x_2 \leq ... \leq x_R,$$

- Suppose instead, you use the set of scores

$$z_1 \leq z_2 \leq ... \leq z_R$$

- The power of the test using the scores

  $$z_1 \le z_2 \le ... \le z_R$$

  approximately equals the squared Pearson correlation:

  $$[Corr(z_j, x_j)]^2 = \left(\frac{\sum_{j=1}^R n_j [z_j - \bar{z}][x_j - \bar{x}]}{\sqrt{\sum_{j=1}^R n_j [z_j - \bar{z}]^2 \sum_{j=1}^R n_j [x_j - \bar{x}]^2}} \right)^2$$

- Then, if z_j is a linear function of x_j, the correlation equals 1, and the efficiency equals 1.

Example

Recall the following toxicity dataset,

```
                          Toxicity

                   | SOME   | NONE  | Total
          ---------+--------+-------+
                 1 |      8 |    92 |  100
          ---------+--------+-------+
Dose (mg)       10 |     15 |    85 |  100
          ---------+--------+-------+
               100 |     22 |    78 |  100
          ---------+--------+-------+
              1000 |     26 |    74 |  100
          ---------+--------+-------+
          Total          71     329    400
```

- We want to fit the model

$$P(\text{Some Toxicity}|\text{dose level } j) = p_j =$$

$$\left(\frac{e^{\beta_0+\beta_1 x_j}}{1+e^{\beta_0+\beta_1 x_j}}\right),$$

- First, we will test for trend using
  $(x_1, x_2, x_3, x_4) = (1, 10, 100, 1000)$.
  and
  $(x_1, x_2, x_3, x_4) = (0, 1, 2, 3)$.

Cochran-Armitage Trend Tests (Score Test)

The null hypothesis for the Cochran-Armitage Trend test is that

$$H_0 = p_j = p \ \ \forall j$$

To test this in SAS, you need to specify the TREND option in the table statement.

```
data one;
input x y count;
cards;
   1  1  8
   1  0 92
  10  1 15
  10  0 85
 100  1 22
 100  0 78
1000  1 26
1000  0 74
;
run;
proc freq data=one;
 tables x*y/trend;
 weight count;
run;
```

For scores (1, 10, 100, 1000),

```
Cochran-Armitage Trend Test
---------------------------
Statistic (Z)       -2.6991
One-sided Pr <  Z    0.0035
Two-sided Pr > |Z|   0.0070

Sample Size = 400
```

Similarly, you could use the scores (0,1,2,3) to get

```
Cochran-Armitage Trend Test
---------------------------
Statistic (Z)       -3.5698
One-sided Pr <  Z    0.0002
Two-sided Pr > |Z|   0.0004

Sample Size = 400
```

Model	Scores	Chi-Square	$p-$value
(1)	(1,10,100,1000)	7.285 ($= -2.6991^2$)	0.0070
(2)	(0,1,2,3)	12.744 ($= -3.5698^2$)	0.0004

- Suppose $(1, 10, 100, 1000)$ are the correct scores, the efficiency when wrongly using $(0, 1, 2, 3)$ instead of
$$(1, 10, 100, 1000)$$
is
$$(Corr[x_j, \log_{10}(x_j)])^2 = 0.82414^2 = 0.67921$$
- Similarly, since the correlation coefficient is symmetric, the efficiency when wrongly using $(1, 10, 100, 1000)$ instead of $(0, 1, 2, 3)$ is
$$(Corr[x_j, \log_{10}(x_j)])^2 = 0.82414^2 = 0.67921$$

Notes on efficiency

- Suppose you have two tests, t_1 and t_2
- Suppose both tests are consistent (i.e., asymptotically converge to the true parameter)
- The asymptotic relative efficiency of t_2 to t_1 can be defined as

$$\text{ARE}_{21} = k$$

  where k is the value for the efficiency ($k = 0.679$ in our example)
- In terms of sample size, you would need k^{-1} subjects to reach the same critical value
- For example, we would need 1.5 ($= .679^{-1}$) times the number of subjects if we misspecified that ranks like we did

Using SAS Proc Corr

```
data one;
input x y count;
logx = log10(x);
cards;
   1  1 92
   1  0  8
  10  1 85
  10  0 15
 100  1 78
 100  0 22
1000  1 74
1000  0 26
;

proc corr ;
 var x logx y ;
 freq count;
run;
```

```
Pearson Correlation Coefficients / Prob > |R| under Ho: Rho=0 / N = 400
/ FREQ Var = COUNT

                     X              LOGX                 Y

X              1.00000           0.82414          -0.13495 = Corr
                0.0               0.0001            0.0070 = p-value

LOGX           0.82414           1.00000          -0.17849
                0.0001            0.0               0.0004

Y             -0.13495          -0.17849           1.00000
                0.0070            0.0004            0.0
```

Note that the $p-$value for $corr(x,y) = 0.0070$ and the $p-$value for the Cochran-Armitage test using scores (1,10,100, 1000) was also 0.0070.

This is not coincidence.

The Cochran-Armitage (CA) Trend Test is the same as

$$CA = n * [corr(x,y)]^2$$

```
data new;
 input var1 $ var2 $ corr;
 n = 400;
 CA = n*(corr**2);
 df=1;
 p = 1-probchi(CA,df);
cards;
x    y   -0.13495
logx y   -0.17849
;
proc print;
run;
```

```
OBS    VAR1    VAR2      CORR       N        CA      DF        P

 1     x        y      -0.13495    400     7.2846     1    .0069548
 2     logx     y      -0.17849    400    12.7435     1    .0003573


data new;
 input var1 $ var2 $ corr;
 eff = (corr**2);
cards;
x    logx 0.82414
;
proc print;
run;

OBS    VAR1    VAR2      CORR       EFF

 1      x      logx    0.82414    0.67921
```

Using SAS Proc Logistic

- Next, we will use SAS Proc Logistic, which also gives us the SCORE (Cochran-Armitage Trend) test as well as the Likelihood Ratio Test and Wald test for

  $$H_0 : \beta_1 = 0$$

  as well as the logistic regression estimates:

Proc Logistic

```
/* give x, y , and n for row binoimals */

data one;
input x y n_j;
cards;
   1  8 100
  10 15 100
 100 22 100
1000 26 100
;

proc logistic;
 model y / n_j =x;
run;
```

```
/*SELECTED OUTPUT */

        Testing Global Null Hypothesis: BETA=0

Test                    Chi-Square       DF     Pr > ChiSq

Likelihood Ratio            6.8146        1         0.0090(1)
Score                       7.2849        1         0.0070(2)
Wald                        7.0915        1         0.0077

(1) = Likelihood ratio test = Gˆ2
(2) = Cochran-Armitage Trend Test


* WALD Test significant (WALD Chi-Square approx equal to LR & Score)
```

```
The LOGISTIC Procedure

             Analysis of Maximum Likelihood Estimates

                               Standard          Wald
Parameter    DF    Estimate       Error    Chi-Square    Pr > ChiSq

Intercept     1     -1.7808      0.1692      110.8292        <.0001
x             1    0.000769    0.000289        7.0915        0.0077
```

Interpretation

Note, for a 990 unit increase in the dose (from 10 to 1000),

$$\begin{aligned} OR(1000:10) &= e^{\hat{\boldsymbol{\beta}}_1(1000-10)} \\ &= e^{.0007688(990)} \\ &= 2.14 \end{aligned}$$

the odds of some toxicity doubles.

Other Models:

Other possible models could include squared terms, cubic terms, etc. For example, the model including the squared terms is:

$$p_j = \left(\frac{e^{\beta_0 + \beta_1 x_j + \beta_2 x_j^2}}{1 + e^{\beta_0 + \beta_1 x_j + \beta_2 x_j^2}} \right),$$

SAS Proc Logistic

```
proc logistic data=one descending;
 model y = x x*x ;
 weight count;  /* number of individuals with y value */
run;

/* Selected Output */

              Analysis of Maximum Likelihood Estimates

                                Standard          Wald
 Parameter    DF    Estimate       Error    Chi-Square    Pr > ChiSq

 Intercept     1     -2.1088      0.2378       78.6214        <.0001
 x             1     0.00946     0.00385        6.0320        0.0140
 x*x           1     -8.4E-6    3.691E-6        5.1729        0.0229
```

Saturated Model

- Since the row margins are fixed, there is one free probability in each row, and the saturated model has a different probability for each level of x_j, i.e., the saturated model has R parameters.
- One way to get a saturated model is to use powers up to $R-1$, i.e.,

$$p_j = \left(\frac{e^{\beta_0+\beta_1 x_j+\beta_2 x_j^2+\ldots+\beta_{R-1} x_j^{R-1}}}{1+e^{\beta_0+\beta_1 x_j^2+\beta_{R-1} x_j^{R-1}}} \right),$$

- Alternatively, you get the same fit by fitting a separate probability for each row (separately maximizing each row binomial), giving the MLE

$$\widehat{p}_j = \frac{y_j}{n_j}$$

- Another way to fit the saturated model is to have a model with an intercept and $(R-1)$ row effects:

$$p_j = \left(\frac{e^{\beta_0 + \beta_j}}{1 + e^{\beta_0 + \beta_j}} \right),$$

  where we constrain $\beta_R = 0$, since there are only R free parameters. $X = R$ is often thought of as the 'reference group'.

- In particular,

$$p_1 = \left(\frac{e^{\beta_0 + \beta_1}}{1 + e^{\beta_0 + \beta_1}} \right),$$

$$p_2 = \left(\frac{e^{\beta_0 + \beta_2}}{1 + e^{\beta_0 + \beta_2}} \right),$$

$$p_{R-1} = \left(\frac{e^{\beta_0 + \beta_{R-1}}}{1 + e^{\beta_0 + \beta_{R-1}}} \right),$$

$$p_R = \left(\frac{e^{\beta_0}}{1 + e^{\beta_0}} \right),$$

- This model may be especially appropriate when the rows are not ordered, i.e., the rows may correspond to race, treatment, gender, etc...

Odds Ratios when rows are not ordinal

- Consider saturated model with $R = 3$, where

$$X = 1 = \text{Drug A},$$

$$X = 2 = \text{Drug B},$$

$$X = 3 = \text{Placebo (drug C)},$$

  and $Y = 1$ is a successful response.

- We fit the model

$$\text{logit}(p_j) = \beta_0 + \beta_j$$

  with $\beta_3 = 0$ for group 3 (placebo, the reference group).

- Then, for an individual on placebo,

$$\text{logit}(p_3) = \beta_0$$

- For an individual on drug A,

$$\text{logit}(p_1) = \beta_0 + \beta_1$$

- For an individual on drug B,

$$\text{logit}(p_2) = \beta_0 + \beta_2$$

- Then,

$$\beta_1 = \text{logit}(p_1) - \text{logit}(p_3) = \log\left(\frac{p_1/(1-p_1)}{p_3/(1-p_3)}\right)$$

  and

$$\beta_2 = \text{logit}(p_2) - \text{logit}(p_3) = \log\left(\frac{p_2/(1-p_2)}{p_3/(1-p_3)}\right)$$

- Thus, β_1 is the log-odds ratio for drug A relative to the placebo, and β_2 is the log odds ratio for drug B relative to the placebo.
- Suppose you want to compare drugs A and B. Then the log-odds ratio between A and B is

$$\begin{aligned}\beta_1 - \beta_2 &= [\text{logit}(p_1) - \text{logit}(p_3)] - [\text{logit}(p_2) - \text{logit}(p_3)] \\ &= [\text{logit}(p_1) - \text{logit}(p_2)] \\ &= \log\left(\tfrac{p_1/(1-p_1)}{p_2/(1-p_2)}\right)\end{aligned}$$

The estimate is

$$\hat{\boldsymbol{\beta}}_1 - \hat{\boldsymbol{\beta}}_2$$

and the variance can be estimated by

$$\widehat{Var}(\hat{\boldsymbol{\beta}}_1 - \hat{\boldsymbol{\beta}}_2) = \widehat{Var}(\hat{\boldsymbol{\beta}}_1) + \widehat{Var}(\hat{\boldsymbol{\beta}}_2) - 2\widehat{Cov}(\hat{\boldsymbol{\beta}}_1, \hat{\boldsymbol{\beta}}_2)$$

(the two are correlated because they both contain logit$(\widehat{p}_3))$.

Most computer packages will print out the covariances so that you can do it by hand, or, they will allow you to estimate the variance of a contrast of the form

$$\mathbf{c}\beta,$$

where **c** is a vector of constants.

Here

$$\mathbf{c} = [0 \;\; 1 \;\; -1]$$

and

$$\beta = [\beta_0 \;\; \beta_1 \;\; \beta_2]'$$

In particular, for this example,

$$\mathbf{c}\widehat{\beta} = [0 \;\; 1 \;\; -1]\widehat{\beta} = \widehat{\beta}_1 - \widehat{\beta}_2,$$

and

$$Var[\mathbf{c}\widehat{\beta}] = \mathbf{c}Var[\widehat{\beta}]\mathbf{c}' =$$
$$\widehat{Var}(\hat{\boldsymbol{\beta}}_1) + \widehat{Var}(\hat{\boldsymbol{\beta}}_2) - 2\widehat{Cov}(\hat{\boldsymbol{\beta}}_1, \hat{\boldsymbol{\beta}}_2)$$

Example

The toxicity data:

```
                          Toxicity

                   | SOME   | NONE  | Total
          ---------+--------+-------+
                 1 |      8 |    92 |  100
          ---------+--------+-------+
Dose (mg)       10 |     15 |    85 |  100
          ---------+--------+-------+
               100 |     22 |    78 |  100
          ---------+--------+-------+
              1000 |     26 |    74 |  100
          ---------+--------+-------+
          Total          71     329    400
```

We are not going to take the row ordering into account, and will fit the model,

$$\text{logit}(p_j) = \beta_0 + \beta_j$$

where we constrain $\beta_4 = 0$.
We are going to use the computer packages to test

$$\log[OR(100:10)] = \text{logit}(p_3) - \text{logit}(p_2) = \beta_3 - \beta_2 = 0$$

USING SAS PROC LOGISTIC

```
data one;
input x y count;

if x =   1 then x1=1;  else x1=0;
if x =  10 then x2=1;  else x2=0;
if x = 100 then x3=1;  else x3=0;

cards;
   1  0  8
   1  1 92
  10  0 15
  10  1 85
 100  0 22
 100  1 78
1000  0 26
1000  1 74
;
```

```
proc logistic data=one;
  model y = x1 x2 x3  ;
 freq count;
 contrast ’logOR for 100 vs 10’ x2 -1 x3 1;
run;

/* SELECTED OUTPUT */
             Analysis of Maximum Likelihood Estimates

                               Standard          Wald
Parameter    DF    Estimate       Error    Chi-Square    Pr > ChiSq

Intercept     1     -1.0460      0.2280       21.0495        <.0001
x1            1     -1.3962      0.4334       10.3785        0.0013
x2            1     -0.6886      0.3611        3.6364        0.0565
x3            1     -0.2197      0.3320        0.4378        0.5082
```

```
                 Contrast Test Results

                                     Wald
Contrast                 DF    Chi-Square    Pr > ChiSq

logOR for 100 vs 10       1        1.6086        0.2047
```

```
proc logistic data=one;
 class x /param=ref ; /* sets x=4 as reference group */
  model y = x   ;
 freq count;
 contrast 'logOR for 100 vs 10' x 0 -1 1 0;
run;
```

```
/* SELECTED OUTPUT */

                Analysis of Maximum Likelihood Estimates

                                    Standard          Wald
Parameter         DF     Estimate      Error    Chi-Square    Pr > ChiSq

Intercept          1      -1.0460     0.2280       21.0495        <.0001
x         1        1      -1.3962     0.4334       10.3785        0.0013
x         10       1      -0.6886     0.3611        3.6364        0.0565
x         100      1      -0.2197     0.3320        0.4378        0.5082


                 Contrast Test Results

                                     Wald
Contrast                 DF    Chi-Square    Pr > ChiSq

logOR for 100 vs 10       1        1.6086        0.2047
```

Gooodness-of-Fit

- The likelihood ratio statistic for a given model M_1 with estimates $\tilde{p}_j$ versus a 'saturated' model in which $\widehat{p}_j = y_j/n_j$, is often called the deviance, denoted by D^2,

$$
\begin{aligned}
D^2(\mathsf{M}_1) &= 2\{\log[L(\mathsf{Sat})] - \log[L(\mathsf{M}_1)]\} \\
&= 2\sum_{j=1}^{R}\left[y_j \log\left(\frac{y_j}{n_j\tilde{p}_j}\right) + (n_j - y_j)\log\left(\frac{n_j - y_j}{n_j(1-\tilde{p}_j)}\right)\right] \\
&= 2\sum_{j=1}^{R}\sum_{k=1}^{2} O_{jk}\log\left(\frac{O_{jk}}{E_{jk}}\right) \\
&\sim \chi^2_P
\end{aligned}
$$

under the null, where

$$P = \text{\# parameters in sat. model} - \text{\# parameters in } \mathsf{M}_1$$

- In general, the deviance D^2 is often used as a measure of overall goodness-of-fit of the model, and is a test statistic form terms **left out** of the model.

SAS Proc Logistic

```
data one;
data one;
input x y count;
cards;
   1  1  8
   1  0 92
  10  1 15
  10  0 85
 100  1 22
 100  0 78
1000  1 26
1000  0 74
;
proc logistic descending;
 model y = x /aggregate scale=d /* specify for deviance */ ;
 freq count;
run;
```

```
/* Selected Output */
       Deviance and Pearson Goodness-of-Fit Statistics

Criterion        DF          Value     Value/DF     Pr > ChiSq

Deviance          2         6.9618       3.4809         0.0308
Pearson           2         6.7383       3.3692         0.0344

Number of unique profiles: 4

             Analysis of Maximum Likelihood Estimates

                               Standard          Wald
Parameter    DF    Estimate       Error    Chi-Square    Pr > ChiSq

Intercept     1     -1.7808      0.3156       31.8392        <.0001
x             1    0.000769    0.000539        2.0373        0.1535
```

Here we would reject the null hypothesis of a “good fit”.

Likelihood Ratio Statistic for Nested Models

- Sometimes you can look at a broader model than the one of interest to test for 'Goodness-of-Fit'.
- For example, suppose you want to see if Model 1 fits,
  Model 1:

$$p_j = \left(\frac{e^{\beta_0 + \beta_1 x_j}}{1 + e^{\beta_0 + \beta_1 x_j}} \right).$$

- This model is nested in (model 1 nested in model 2)
  Model 2:

$$p_j = \left(\frac{e^{\beta_0 + \beta_1 x_j + \beta_2 x_j^2}}{1 + e^{\beta_0 + \beta_1 x_j + \beta_2 x_j^2}} \right),$$

- Recall, the deviance D^2 is sort of like a SUMS of SQUARES ERROR (error in the given model versus the saturated), and a smaller model will always have the same or more error than the bigger model.

- To test for significance of parameters in model 2 versus model 1, you can use

$$\Delta D^2(\mathsf{M}_2|\mathsf{M}_1) = D^2(\mathsf{M}_1) - D^2(\mathsf{M}_2)$$

  which is the ‘change in D^2’ for model 2 versus model 1.
- If the smaller model fits, in large samples,

$$\Delta D^2(\mathsf{M}_2|\mathsf{M}_1) \sim \chi^2_P,$$

  where P parameters are set to 0 to get the smaller model.

Using G^2

- As before, another popular statistic is G^2, which is the likelihood ratio test statistic for whether the parameters, except the intercept β_0, are 0 (i.e., the significance of parameters in the model).
- For G^2, the larger model always has bigger G^2 since it has more parameters (sort of like SUMS of SQUARES REGRESSION)
- Again, to test for significance of parameters in model 2 versus model 1, you can use

$$\Delta G^2(\mathsf{M}_2|\mathsf{M}_1) = G^2(\mathsf{M}_2) - G^2(\mathsf{M}_1)$$

  which is the 'change in G^2' for model 2 versus model 1.
- Thus, the likelihood ratio statistic for two nested models can be calculated using either ΔG^2 or ΔD^2.
- Note that $\Delta G^2 = \Delta D^2$ when testing the same two models (we will see this empirically in an example)

Residuals

- Sometimes you can look at residuals to see where the model does not fit well.
- The standard (or unadjusted) Pearson residuals

  $$e_j = \left(\frac{[y_j - n_j\widehat{p}_j]^2}{\sqrt{n_j\widehat{p}_j(1-\widehat{p}_j)}} \right)$$

  If the model fits, then, asymptotically,

  $$e_j \sim N(0,1)$$

  (as $n_j \to \infty$)
- Note that, the score statistic (Pearson's chi-square) versus the saturated model is

  $$X^2 = \sum_{j=1}^{R} e_j^2$$

- Another popular residual is the ‘Deviance residual’. The deviance residual is defined as

$$d_j = \pm\sqrt{\left[y_j \log\left(\frac{y_j}{n_j\widehat{p}_j}\right) + (n_j - y_j)\log\left(\frac{n_j - y_j}{n_j(1-\widehat{p}_j)}\right)\right]},$$

  where the sign (+ or -) is the same as $(y_j - n_j\widehat{p}_j)$. When $y_j = 0$ or $y_j = n_j$, the deviance residual is defined as

$$d_j = \left\{ \begin{array}{ll} -\sqrt{2n_j|\log(1-\widehat{p}_j)|} & \text{if } y_j = 0 \\ \sqrt{2n_j|\log(\widehat{p}_j)|} & \text{if } y_j = n_j \end{array} \right. .$$

- When none of the y_j equal 0 or n_j, then

$$D^2 = \sum_{j=1}^{R} d_j^2$$

The toxicity data:

```
                          Toxicity

                   | SOME   | NONE  | Total
          ---------+--------+-------+
                 1 |      8 |    92 |  100
          ---------+--------+-------+
Dose (mg)       10 |     15 |    85 |  100
          ---------+--------+-------+
               100 |     22 |    78 |  100
          ---------+--------+-------+
              1000 |     26 |    74 |  100
          ---------+--------+-------+
          Total          71     329    400
```

Summary of Models

	pars. in model		pars. not in model		
					$p-$value
Model	$df(G^2)$	G^2	$df(D^2)$	D^2	for D^2
(1) Null (β_0)	0	0	3	13.78	0.0032
(2) x	1	6.82	2	6.96	0.0308
(3) x, x^2	2	11.88	1	1.90	0.1682
(4) SATURATED	3	13.78	0	0	-

Overall, the model with linear and quadratic terms (x, x^2) appears to be the best fit.

Comparing Model 3 to 2

$$\Delta D^2 = 6.96 - 1.90 = 5.06$$

and

$$\Delta G^2 = 11.88 - 6.82 = 5.06$$

both on 1 degrees of freedom: Conclusion x^2 is needed in the model

```
                                    Standard          Wald
Parameter    DF    Estimate        Error    Chi-Square    Pr > ChiSq

Intercept     1     -2.1088       0.2378       78.6214        <.0001
x             1     0.00946      0.00385        6.0320        0.0140
x*x           1     -8.4E-6     3.691E-6        5.1729        0.0229
```

Note, the parameter estimate for the coefficient of x^2 is very small, but that is because x^2 is large, especially when $x = 1000$. Maybe I should have chosen $\log(x)$ as the covariate.

- For model (3), (x, x^2), the odds ratio for x_j versus $x_{j'}$ is

$$\begin{aligned} OR(x_j : x_{j'}) &= \frac{p_j/(1-p_j)}{p_{j'}/(1-p_{j'})} \\ &= \frac{e^{\beta_0+\beta_1 x_j+\beta_2 x_j^2}}{e^{\beta_0+\beta_1 x_{j'}+\beta_2 x_{j'}^2}} \\ &= e^{\beta_1(x_j-x_{j'})+\beta_1(x_j^2-x_{j'}^2)} \end{aligned}$$

- Then, the odds ratio for $x_j = 100$ versus $x_{j'} = 10$ is

$$\begin{aligned} OR(100 : 10) &= e^{.00946(100-10)-.0000084(100^2-10^2)} \\ &= 2.15 \end{aligned}$$

The observed OR for these two rows is

$$\frac{22 \cdot 85}{15 \cdot 78} = 1.6,$$

so the model overestimates this odds ratio by a little.

Residuals

The Pearson and Deviance residuals are

```
            x           y   Pearson   Deviance
  1.        1           1  -.9351095  -.9766505
  2.        1           0  -.9351095  -.9766505
  3.       10           1   1.004686   .9689428
  4.       10           0   1.004686   .9689428
  5.      100           1  -.0777304  -.0778652
  6.      100           0  -.0777304  -.0778652
  7.     1000           1   .0006648   .0006648
  8.     1000           0   .0006648   .0006648
```

Pearson’s chi-square for the model (x, x^2) versus the saturated model is the sum of squares of the Pearson residuals, and equals
1.89 (1 df) $p-$value = 0.1692.

This is similar to the Deviance for the model (x, x^2) versus the saturated model,
D^2 = 1.90 (1 df) $p-$value = 0.1682
The model (x, x^2) seems to fit OK.