CPH 931 — Fall 2008 — Dr. Charnigo

Lecture 7

Generalized linear models

A connection between linear and logistic regression. Recall that in linear regression we have

$$\mu_{\mathbf{x}} = \alpha + \beta_1 x_1 + \cdots + \beta_k x_k, \tag{1}$$

where $\mu_{\mathbf{x}}$ denotes the mean of a continuous response variable Y when the explanatory variables $X_1, \ldots, X_k$ have taken on the numerical values $x_1, \ldots, x_k$. On the other hand, in logistic regression we have

$$\text{logit}[p_{\mathbf{x}}] = \alpha + \beta_1 x_1 + \cdots + \beta_k x_k, \tag{2}$$

where $p_{\mathbf{x}}$ denotes the probability that a dichotomous response variable Y (coded as 0 or 1) equals 1 when the explanatory variables $X_1, \ldots, X_k$ have taken on the numerical values $x_1, \ldots, x_k$.

A key fact about a dichotomous variable Y is that the probability of Y equaling 1 is the same as the mean of Y. For instance, if the probability that Y equals 1 is 0.5, then half of the time Y will equal 1 and half of the time Y will equal 0; so, the expected value of Y will be 0.50. Using this fact, we can rewrite (2) as

$$\text{logit}[\mu_{\mathbf{x}}] = \alpha + \beta_1 x_1 + \cdots + \beta_k x_k, \tag{3}$$

where $\mu_{\mathbf{x}}$ denotes the mean of the dichotomous response variable Y when $X_1 = x_1, \ldots, X_k = x_k$.

The only explicit distinction between (1) and (3) is that the left side of (3) applies a nonlinear transformation to $\mu_{\mathbf{x}}$ before equating it to the right side. An implicit distinction is that the distribution of Y given $\mu_{\mathbf{x}}$ is normal for model (1) but binomial with one trial for model (3).

A general framework. Models (1) and (3) can be subsumed into the following general framework. Let Y be a response variable, and let $\mu_{\mathbf{x}}$ denote its mean when the explanatory variables $X_1, \ldots, X_k$ have taken on the numerical values $x_1, \ldots, x_k$. Consider relating Y to $X_1, \ldots, X_k$ via the following two-part statistical model. First, we specify that

$$g(\mu_{\mathbf{x}}) = \alpha + \beta_1 x_1 + \cdots + \beta_k x_k \tag{4}$$

for some increasing function g. Second, we specify a distribution for Y given $\mu_{\mathbf{x}}$. Such a model is called a "generalized linear model". Statisticians refer to g as a "link function" and to $\alpha + \beta_1 x_1 + \cdots + \beta_k x_k$ as a "linear predictor".

Example 1, linear regression. Take g to be the identity function, $g(\mu_{\mathbf{x}}) = \mu_{\mathbf{x}}$. Then (4) simplifies to (1). We specify that the distribution of Y given $\mu_{\mathbf{x}}$ is normal with mean $\mu_{\mathbf{x}}$ and unknown variance σ^2 that does not depend on $x_1, \ldots, x_k$.

Example 2, logistic regression. Take g to be the logit function, $g(\mu_{\mathbf{x}}) = \text{logit}[\mu_{\mathbf{x}}]$. Then (4) simplifies to (3). We specify that the distribution of Y given $\mu_{\mathbf{x}}$ is binomial with one trial and mean $\mu_{\mathbf{x}}$ (or, equivalently, success probability $\mu_{\mathbf{x}}$).

Example 3, Poisson regression. Take g to be the natural logarithm function, $g(\mu_{\mathbf{x}}) = \log[\mu_{\mathbf{x}}]$. We specify that the distribution of Y given $\mu_{\mathbf{x}}$ is Poisson with mean $\mu_{\mathbf{x}}$.

A brief review of the Poisson distribution follows, after which the Poisson regression model is developed in more detail and then exemplified using a data set on SARS incidence.

Poisson and negative binomial regression

Review of the Poisson distribution. We say that Y has the Poisson distribution with mean μ (> 0) if the probability mass function for Y is

$$P(Y = y) = e^{-\mu}\ \mu^y/y!, \tag{5}$$

where y is any nonnegative integer. The Poisson distribution is often used to model the total number of events occurring within some group of individuals over a specific period of time. If λ (> 0) is the incidence rate, then we have the relationship $\mu = \lambda t$, where t is the number of person-years.

To illustrate how the Poisson distribution can be used, let us suppose that the total number of colds experienced by the third-graders in a Lexington elementary school this year is a Poisson random variable. Suppose, moreover, that the incidence rate is $\lambda = 1.5$ and that there are 60 third-graders whom we are tracking for one year; hence, there are $t = 60$ person-years. If we want to know the probability that the total number of colds will be at least 100, we can calculate the probability as follows. First, we have $\mu = \lambda t = 1.5(60) = 90$. Second, if Y is Poisson with mean 90, then

$$P(Y \geq 100) = 1 - P(Y \leq 99) = 1 - 0.842 = 0.158.$$

Obviously I did not evaluate

$$P(Y \leq 99) = e^{-90}\ 90^0/0! + e^{-90}\ 90^1/1! + \cdots + e^{-90}\ 90^{99}/99!$$

using a desk calculator; rather, I used the Excel spreadsheet {PoissonCalc.xls} to find that $P(Y \leq 99) = 0.842$. Thus, the Poisson model suggests that there is a 15.8% chance of at least 100 colds being experienced this year by the third-graders in the Lexington elementary school.

Poisson regression. Suppose that a response variable Y is believed to have a Poisson distribution but that the mean $\mu_{\mathbf{x}}$ is believed to depend on the numerical values $x_1, \ldots, x_k$ assumed by some explanatory variables $X_1, \ldots, X_k$. Our goal is to quantify the dependence of $\mu_{\mathbf{x}}$ on $x_1, \ldots, x_k$.

As noted in **Example 3** on page 2, we can employ a generalized linear model with

$$\log[\mu_{\mathbf{x}}] = \alpha + \beta_1 x_1 + \cdots + \beta_k x_k. \tag{6}$$

We refer to (6), in conjunction with the specification that Y have a Poisson distribution, as a Poisson regression model.

We interpret β_1 in (6) as the additive increase in $\log[\mu_{\mathbf{x}}]$ corresponding to a one-unit increase in X_1 when $X_2, \ldots, X_k$ are fixed. Or, if you prefer, $\exp[\beta_1]$ is the multiplicative increase in $\mu_{\mathbf{x}}$ corresponding to a one-unit increase in X_1 when $X_2, \ldots, X_k$ are fixed.

If we want to express the incidence rate — rather than the mean — in terms of $x_1, \ldots, x_k$, then we may replace (6) by

$$\log[\mu_{\mathbf{x}}] = \log[t_{\mathbf{x}}] + \alpha^* + \beta_1^* x_1 + \cdots + \beta_k^* x_k, \tag{7}$$

as this simplifies to

$$\log[\lambda_{\mathbf{x}}] = \alpha^* + \beta_1^* x_1 + \cdots + \beta_k^* x_k. \tag{8}$$

Statisticians refer to $\log[t_{\mathbf{x}}]$ as an "offset".

If $t_{\mathbf{x}}$ is constant in our data set, then we anticipate obtaining essentially equivalent results from models (6) and (7) since $\log[t_{\mathbf{x}}] + \alpha^*$ can be identified with α (and $\beta_1^*, \ldots, \beta_k^*$ with $\beta_1, \ldots, \beta_k$).

To clarify what constant $t_{\mathbf{x}}$ looks like, here are a few records from a fictional data set with constant $t_{\mathbf{x}}$:

school	colds	third graders	avg daily high temp
A	92	60	58
B	83	60	62
C	107	60	52
D	124	60	53
E	65	60	69

And here are a few records from a fictional data set with nonconstant $t_{\mathbf{x}}$:

school	colds	third graders	avg daily high temp
A	137	90	58
B	97	70	62
C	89	50	52
D	62	30	53
E	11	10	69

We interpret β_1^* in (8) as the additive increase in $\log[\lambda_{\mathbf{x}}]$ corresponding to a one-unit increase in X_1 when $X_2, \ldots, X_k$ are fixed. Or, if you prefer, $\exp[\beta_1^*]$ is the (incidence) rate ratio associated with a one-unit increase in X_1 when $X_2, \ldots, X_k$ are fixed.

Example. The data set in {SARS.xls} presents information on SARS incidence in Hong Kong, Singapore, and Taiwan for each day in a period of approximately three months. In what follows, we confine our attention to the Hong Kong information.

The variable "Time" (Column C) is a time index, while "Time2" (Column D) is the corresponding quadratic. The variable "Dailyinf" (Column E) records the number of new SARS cases on a given day, while "Meantemp" (Column K) and "Meanhum" (Column L) record the weather conditions for that day. We take Dailyinf as the response variable and Time, Time2, Meantemp, Meanhum as explanatory variables.

The results of fitting model (6) in SAS are on page 1 of {SARSresults.rtf}. The expected number of new SARS cases as a function of Time, Time2, Meantemp, and Meanhum is estimated by

$$\exp[1.7930 + 0.0642 Time - 0.0013 Time2 + 0.0300 Meantemp + 0.0056 Meanhum].$$

We may interpret the 0.0300 by saying that a one-degree increase in mean temperature multiplies the expected number of new SARS cases by an estimated factor of $\exp[0.0300] = 1.0305$ at a fixed time and mean humidity. We may also say that a one-degree increase in mean temperature multiplies the incidence rate by an estimated factor of 1.0305 at a fixed time and mean humidity. The latter statement is valid since $t_{\mathbf{x}}$ would have been approximately constant: the number of people living in Hong King would not have varied much from day to day over this three-month period.

We can perform Wald tests on partial slope coefficients in exactly the same manner as for ordinal logistic regression (Cf. Lecture 6), and we can perform likelihood ratio tests in essentially the same manner, noting that SAS lists $\log L$ in the output (4084.4451 in this instance) rather than $-2 \log L$.

The overdispersion phenomenon. A common difficulty with Poisson regression is that the distribution of Y given $\mu_{\mathbf{x}}$ is not truly Poisson. If Y were Poisson with mean $\mu_{\mathbf{x}}$, then the variance of Y should also be $\mu_{\mathbf{x}}$. However, in practice we often find that the variance of Y is larger than $\mu_{\mathbf{x}}$. Such excess variability is called "overdispersion".

Overdispersion may occur because we have failed to incorporate some useful predictors into our statistical model:

1. We may have incorrectly judged a useful predictor to be irrelevant.
2. We may have neglected to measure a useful predictor because we did not realize that it would be useful.
3. We may have been unable to measure a useful predictor.
4. We may have conceded some realism to have a simple model.

In any case, a practical consequence of overdispersion is that the standard errors for parameter estimates are understated. This implies that the Wald chi-square statistics are overstated and, hence, that some declarations of statistical significance may be inappropriate.

We can diagnose overdispersion by allowing SAS to estimate a "scale parameter" σ. I emphasize that, in this setting, σ is not a standard deviation. Rather, σ indicates the degree to which the variance of Y is inflated above what would be consistent with a Poisson distribution. Explicitly, we have

$$Var[Y|\mathbf{x}] = \sigma^2 \; \mu_{\mathbf{x}}. \tag{9}$$

If $\sigma = 1$, then there is no overdispersion. If $\sigma > 1$, then there is overdispersion. We can address the overdispersion by multiplying all of the standard errors by (SAS's estimate of) σ. If $\sigma < 1$, then we say that we have "underdispersion". However, underdispersion does not occur often in practice.

Negative binomial regression. As noted above, one way to deal with overdispersion in Poisson regression is to estimate the scale parameter σ and adjust the standard errors accordingly; we will illustrate this with the SARS example on the next page. Before we do that, however, I want to describe another option for dealing with overdispersion. This option entails replacing the Poisson regression model by another generalized linear model in which the mean and variance of Y are not constrained to be equal.

Again suppose that

$$\log[\mu_{\mathbf{x}}] = \alpha + \beta_1 x_1 + \cdots + \beta_k x_k. \tag{10}$$

However, instead of assuming that Y is Poisson with mean $\mu_{\mathbf{x}}$ (and, hence, variance $\mu_{\mathbf{x}}$), assume that Y is negative binomial with mean $\mu_{\mathbf{x}}$ and variance $\mu_{\mathbf{x}} + k\mu_{\mathbf{x}}^2$. We refer to k (> 0) as a "dispersion parameter".

The mathematical form of the probability mass function for Y — i.e., the negative binomial analogue to (5) — is not of primary importance to us. I do not present it here because to do so would require me to define the gamma function; that would entail some calculus and would take us too far afield.

The only really important points for us to note about the negative binomial distribution are that: the variance of Y is greater than the mean; and, the dispersion parameter k determines precisely how much greater the variance is.

Finally, model (10) can be replaced by

$$\log[\lambda_{\mathbf{x}}] = \alpha^* + \beta_1^* x_1 + \cdots + \beta_k^* x_k \tag{11}$$

if our interest lies with $\lambda_{\mathbf{x}}$ rather than $\mu_{\mathbf{x}}$, just as (6) could be replaced by (8) in Poisson regression.

Example, continued. Refer to page 1 of {SARSresults.rtf}. I call your attention to the first row in the box on "Criteria for Assessing Goodness of Fit". The "Deviance" measures the extent to which the Poisson regression model does not fit the observed data. You can think of the Deviance as being analogous to the Residual Sum of Squares from linear regression. Since there were 92 observations for Hong Kong (corresponding to 92 days) and 5 parameters estimated, there are $92 - 5 = 87$ degrees of freedom. In the absence of overdispersion, the Deviance should be close to the degrees of freedom (or, put differently, the ratio of the Deviance to the degrees of freedom should be close to 1). Here we see that the ratio of the Deviance to the degrees of freedom is $307.68/87 = 3.5366 >> 1$, which suggests that there is overdispersion.

On page 2 of {SARSresults.rtf} are the results I obtained when I asked SAS to estimate the scale parameter in (9) and adjust the standard errors accordingly. The estimate of σ is $\sqrt{3.5366} = 1.8806$, and all of the standard errors in the "Analysis of Parameter Estimates" box have been multiplied by 1.8806 compared to their values on page 1. Note that Meantemp is no longer significant once the standard errors have been appropriately adjusted (Wald p-value = 0.1423 after adjustment). This illustrates that failing to correct for overdispersion in a Poisson regression model may lead to an erroneous declaration of statistical significance.

On page 3 of {SARSresults.rtf} are the results I obtained when I asked SAS to fit the negative binomial regression model (10) instead of the Poisson regression model (6). The estimates of the partial slope coefficients have changed somewhat, although the conclusion about the lack of significance for Meantemp is preserved (Wald p-value = 0.3440). Further, since the Deviance divided by the degrees of freedom is $98.231/87 = 1.1291 \approx 1$, switching to the negative binomial regression model is also a satisfactory mechanism for addressing overdispersion in this example.

Discussion questions

1. I stated that the total number of colds experienced by the third-graders in a Lexington elementary school this year could be regarded as a Poisson random variable. Why could it *not* be regarded as a binomial random variable, with any child who experienced a cold contributing a "success"?

2. On page 1 of {SARSresults.rtf} we can interpret 0.0300 and 0.0056 with relative ease. However, interpreting 0.0642 and -0.0013 requires more effort because we cannot increase Time by one unit without increasing Time2 (or vice versa). So, how can we interpret 0.0642 and -0.0013? Judging by these numbers, approximately when did the incidence of SARS peak in Hong Kong?