BIOS 6244 Analysis of Categorical Data
October 31, 2005 Lecture

Reading Assignment (will not be covered in class)
Sections 4.2.4 and 4.2.5, pp. 78-80. These pages will be posted to the website for those students who do not have a textbook.

Chapter 4 Generalized Linear Models

An alternative to the methods we have described so far for analyzing categorical data is to use a formal statistical model to relate the disease to the exposure. (Actually, *all* of the analyses we have described so far rely on an underlying model, but not of the form we will be discussing today.) Models that *accurately* describe the E-D association have several advantages:

(1) The structural form of the model provides a simple and interpretable way of representing the E-D association.

(2) Statistical inferences for the model parameters help us determine which explanatory variables affect the response, while controlling for the effects of possible confounding variables.

(3) The relative sizes of the estimated model parameters help us ascertain the strength and importance of associations between predictors and outcome.

(4) The predicted values obtained from the model help to "smooth" the data and provide improved estimates of the mean of the dependent variable.

Chapter 4 in our text focuses on *generalized linear models* (GLM's). "Generalized" means that there is some transformation g(Y) that can be applied to the response variable Y so that g(Y) is a linear function of the predictors. Traditional linear models such as multiple regression and ANOVA are special cases of GLM's since the transformation g(Y) = Y yields a linear relationship between the outcome and the predictors.

Components of a GLM (Sec. 4.1)

All GLM's have 3 components: the *random* component, the *systematic* component, and the *link*.

Random Component (Sec. 4.1.1)

Let $\{Y_{1,} Y_2, \ldots, Y_N\}$ denote the observations of the response variable Y, which will be assumed to be independent. The *random* component of the GLM is determined by the probability distribution that we select for Y.

In many of the categorical data examples we have discussed, $\{Y_{1,} Y_2, \ldots, Y_N\}$ represent binary outcomes, or perhaps the number of "successes" out a fixed number of trials. For data like this, we naturally assume that a binomial probability model might be appropriate. A possible model to consider if $\{Y_{1,} Y_2, \ldots, Y_N\}$ denote “counts” (as in observed frequencies in a contingency table) is the Poisson.

Systematic Component (Sec. 4.1.2)

As usual, we denote the expected value (or mean) of Y by $\mu = E(Y)$. In a GLM, the value of μ is assumed to depend on the observed values of the explanatory variables (i.e., there is a different expected value for every combination of values of the explanatory variables).

The *systematic* component of the GLM specifies all of the explanatory variables (predictors) that are to be used in the model. For all GLM’s, these explanatory variables are combined in a linear fashion. In other words, the right-hand side (RHS) of the model equation will always be

$$\alpha + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k, \qquad (11)$$

where the $\{x_j\}$ are the explanatory variables. The combination of explanatory variables given by Equation (11) is called a *linear predictor*. Note that some of the $\{x_j\}$ may be functions of other explanatory variables; for example, $x_3 = x_1x_2$ could be used to indicate the *interaction* between the exposure X_1 and a confounder X_2, and x_1^2 could be included in the model if we wanted to consider higher-order models in the relationship between Y and X_1.

Link (Sec.4.1.3)

The *link* between the random and systematic components specifies how $\mu = E(Y)$ is related to the linear predictor. One can model the mean μ directly, or, more commonly, as some monotone function $g(\mu)$ of the mean. Thus, in the most general case, a GLM is given by

$$g(\mu) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k .$$

The function $g(\cdot)$ is called the *link function*.

As mentioned before, multiple regression models are a special case of GLM’s since they can be obtained by using the *identity link* $g(\mu) = \mu$. Other links are used to represent non-linear relationships between the mean and the predictors.

If μ can be assumed to always be greater than 0 (as in the Poisson distribution), then a *log link* could be used, so that

$$\log(\mu) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k .$$

This particular GLM is called a *log-linear model*.

Another commonly used link is the *logit link* $g(\mu) = \log\left[\dfrac{\mu}{1-\mu}\right]$. It is appropriate when

$0 < \mu < 1$; for example, when μ is a probability. In this case, $\frac{\mu}{1-\mu}$ is the odds and $\log\left[\frac{\mu}{1-\mu}\right]$ is called the *logit* of μ, denoted logit(μ). A GLM using the logit link

$$\log\left[\frac{\mu}{1-\mu}\right] = \alpha + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k$$

is called a *logit model.*

Each probability distribution that we consider for the random component of the GLM is characterized by a *natural parameter.* For the normal distribution, the natural parameter is the mean, μ. For the Poisson, it is $\log \mu$. For the binomial, the natural parameter is $\log\frac{\pi}{1-\pi}$. The link function that uses the natural parameter for $g(\mu)$ in the GLM is called the *canonical link.* While other links are possible (e.g., the *log-log link*), it is the canonical links that are used most commonly.

Normal GLM (Sec. 4.1.4)

Random component: normal distribution

Link: identity

This is the same as the usual multiple regression model. In many applications, an effort is made to transform Y so that it has an approximately normal distribution with homogeneous variance, in order to satisfy the usual assumptions underlying multiple regression. In practice, finding the appropriate transformation is difficult. (The natural log is the most commonly used transformation by far.) Agresti says that finding an appropriate transformation is “usually not possible.” (However, massive amounts of research have been devoted to this topic.)

Finding an appropriate transformation is not necessary if one is willing to consider GLM’s, since they are not based on the normality and homoscedasticity assumptions. This is because fitting a GLM involves maximum likelihood estimation of model parameters using whatever probability distribution other than the normal that we specify in the random component. Furthermore, in using GLM’s, the choice of a link is separate from the choice of a random component. That is, we do not require that our “transformation” of Y via the link produce normality *and* constant variance. The main disadvantage of GLM’s is that least squares estimates of the model parameters generally do not perform very well and closed-form expressions for the estimated model parameters are typically not available. However, with modern statistical software and the availability of powerful desktop computers, this is no longer the issue it once was.

GLM's for Binary Data (Sect. 4.2)

In the usual setup for an Epi study, both the predictor ("exposure") and the outcome ("disease") are binary. The most commonly used probability model for a binary response is the *Bernoulli*. For a Bernoulli RV Y, we know that $\Pr(Y=1) = \pi$, $\Pr(Y=0) = 1 - \pi$, $E(Y) = \pi$, and $\mathrm{Var}(Y) = \pi(1-\pi)$. If we have n independent observations ("trials") of a Bernoulli RV, then their sum (i.e., the total # of succeses out of the n trials) has a binomial distribution with parameters n and π. In this section, we will use the binomial for the random component of the GLM. For simplicity's sake, we will consider only a single explanatory variable, X. We will model $\pi = \Pr(Y=1)$ as a function of x, so we will write $\pi(x)$.

Linear Probability Model (Sec. 4.2.1)

Random component: binomial

Link: identity

The equation for this GLM is

$$\pi(x) = \alpha + \beta x. \qquad (12)$$

As pointed by Agresti, this model has a "major structural defect." On the LHS of Equation (12), we know that $0 \le \pi(x) \le 1$ for all x. However, on the RHS, $-\infty < \alpha + \beta x < \infty$, depending on the values of α and β. Thus, the model given in Equation (12) could yield predicted values of $\pi(x)$ that are < 0 or > 1.

The linear probability model can be valid if the range of x is properly restricted, but practical modelling situations usually require a more complex form for the link.

Even though Equation (12) resembles a simple linear regression model, least squares estimation is not optimal. Since we are assuming that Y has a binomial distribution, we know that $\mathrm{Var}(Y) = \pi(x)[1-\pi(x)]$. Thus, the variance of the outcome variable is not the same for all values of the explanatory variable, and the least squares estimates can have larger variances than maximum likelihood (ML) estimates. ML is the preferred method of estimation for the linear probability model, and for all GLM's.

Example (snoring vs. heart disease) (Sec. 4.2.2)

Consider Table 4.1 in our text. It contains data from an investigation of snoring as a possible risk factor for heart disease.

(Table 4.1 is reproduced on the following page.)

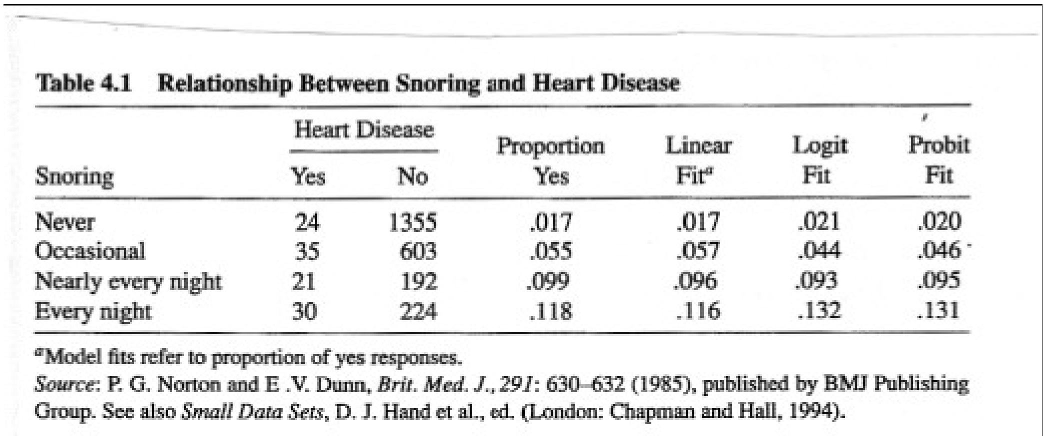

Table 4.1 Relationship Between Snoring and Heart Disease

Snoring	Heart Disease Yes	Heart Disease No	Proportion Yes	Linear Fit[a]	Logit Fit	Probit Fit
Never	24	1355	.017	.017	.021	.020
Occasional	35	603	.055	.057	.044	.046
Nearly every night	21	192	.099	.096	.093	.095
Every night	30	224	.118	.116	.132	.131

[a] Model fits refer to proportion of yes responses.

Source: P. G. Norton and E .V. Dunn, *Brit. Med. J., 291*: 630–632 (1985), published by BMJ Publishing Group. See also *Small Data Sets*, D. J. Hand et al., ed. (London: Chapman and Hall, 1994).

In the linear probability model, the rows of the 4x2 contingency table corresponding to Table 4.1 are treated as independent binomial samples with π(x) defined to be the probability of "success" in that row. Agresti scores the rows as x = 0, 2, 4, 5 so that the last 2 rows will be treated as if they are closer than the other adjacent pairs of rows.

Using PROC GENMOD in SAS, the fitted model using ML estimation is given by

$$\widehat{\pi(x)} = .0172 + .0198x .$$

So, for non-smokers (x = 0), the "fitted value" for the risk of heart disease is

$$\hat{\pi} = .0172 + .0198(0) = .0172 ,$$

compared with the observed risk of .0174. For the occasional snorers, the fitted value of the risk increases from about .02 to about .06, then jumps to .10 for those who snore nearly every night and then to .12 for those who always snore.

In assessing the goodness of fit (GOF) of any GLM, one should always compare the observed values with the fitted values. In Figure 4.1, the observed and fitted values are compared for the linear probability model (and for 2 other models as well). An "eyeball" comparison indicates that the linear probability model fits the data well. Formal methods for assessing the GOF of GLM's are covered in Sec. 5.4 of our text.

(Figure 4.1 is reproduced on the following page.)

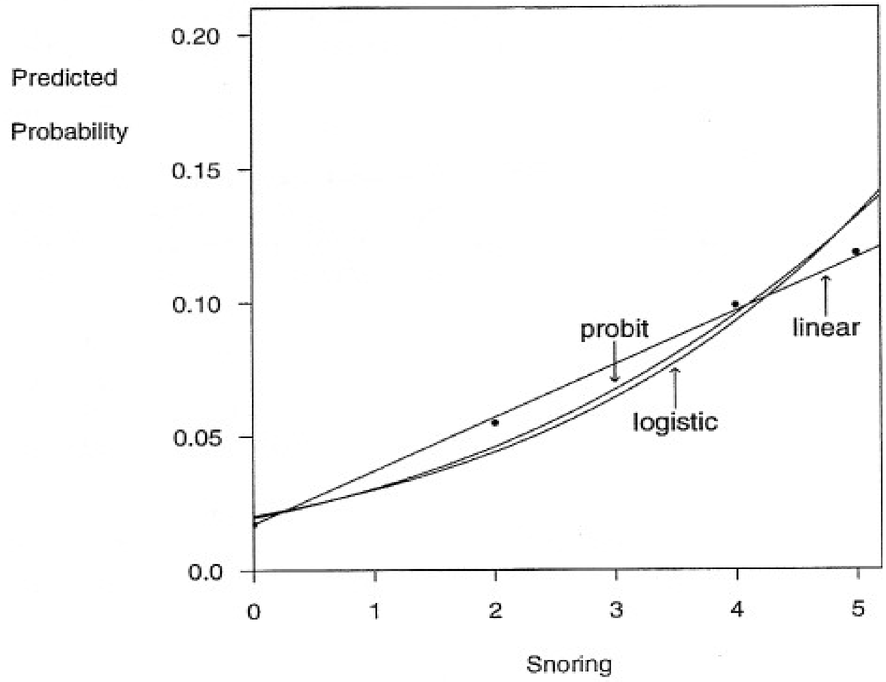

Figure 4.1 Fit of models for snoring and heart disease data.

Logistic Regression Model (Sec. 4.2.3)

In analyzing binary data, it is more common to assume a non-linear relationship between π(x) and x. A fixed change in X may have less impact when π is near 0 or 1 than when π is near .5. (For the linear probability model, a fixed change in X has the same impact regardless of the value of π.)

In practice, non-linear relationships between π(x) and x are often monotonic, with π(x) increasing (or decreasing) continuously as x increases. The "S-shaped" curves displayed in Figure 4.2 often provide a good model for the relationship between π(x) and x.

(Figure 4.2 is reproduced on the following page.)

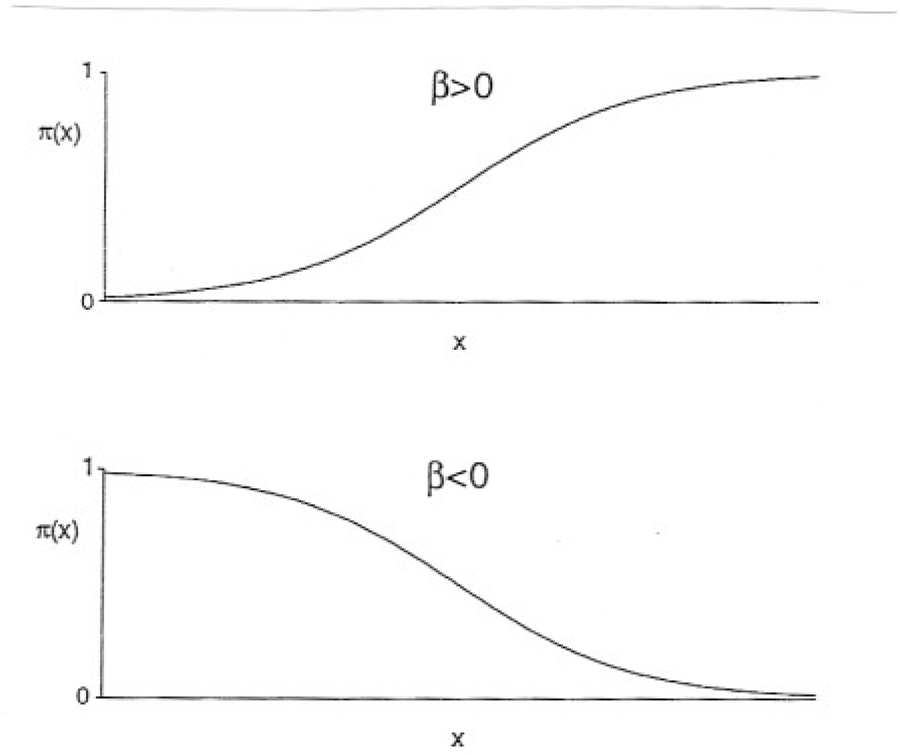

Figure 4.2 Logistic regression functions.

The most commonly used function having a shape like that in Figure 4.2 is the *logistic regression function*

$$\log\left[\frac{\pi(x)}{1-\pi(x)}\right] = \alpha + \beta x \, . \qquad (13)$$

The model represented by Equation (13) is a GLM with

Random component: binomial

Link: logit

Logistic regression models are sometimes called *logit models*. The logit is the natural parameter of the binomial, so the logit link is its canonical link. Notice that the "major structural defect" present in the linear probability model is not present here since both the LHS and RHS of Equation (13) can take on any values between $-\infty$ and ∞.

The parameter β in the model in Equation (13) determines the rate of increase or decrease of the relationship between π(x) and x. When $\beta > 0$, π(x) increases as x increases, and when $\beta < 0$, π(x) decreases as x increases. (See Figure 4.2.)

Example, cont. (snoring vs. heart disease)

PROC LOGISTIC in SAS produces the following ML parameter estimates for the logistic model for these data:

$$\log \mathrm{it}[\widehat{\pi(\mathrm{x})}] = -3.87 + .400\mathrm{x}\ .$$

The fitted values for this model are given in Table 4.1 and plotted in Figure 4.1. An "eyeball" comparison suggests that the logistic model does not fit the data quite as well as the linear probability model.

Probit Models (Sec. 4.2.5)

Another GLM that is very similar to the logistic regression model is the *probit* model. The random component of this model is binomial. The probit link is given by the inverse of the cumulative distribution function (cdf) of the standard normal. In other words, probit[π(x)] is the z-score producing a lower-tail probability of π(x) for the standard normal. So, probit(.05) = -1.645, probit (.975) = 1.96, etc. The probit model was widely used at one time, but the logistic model is now much more popular. The main reason for this is that the slope coefficients in a logistic regression can be converted very easily to odds ratios. (No such conversion is available for the probit model.) In most real-world applications, logit and probit models will yield almost identical results in terms of fitted values, as in the snoring vs. heart disease example. (See Table 4.1 and Figure 4.1.)