BIOS 6244 Analysis of Categorical Data
November 2, 2005 Lecture

GLM for Count Data: Poisson Regression (Sec. 4.3)

Many discrete outcome variables have counts as possible values. Examples include colony counts for bacteria, # of insurance claims, # of cases of disease, tumor counts, etc.

In Section 1.2 of our text, we considered the Poisson distribution as a sampling model for counts. In Chapter 6, Poisson GLM's are considered as models for counts in contingency tables. In this section, we consider Poisson models in the context of *regression analysis*, that is, modeling a single response variable as a function of one or more explanatory variables.

Poisson Regression (Sec 4.3.1)

In Section 1.2.1, we saw that the mean of a Poisson distribution is always positive. Although one can model a Poisson mean in a GLM using the identity link, it is more common to model the log of the mean. Thus, in the GLM

$$\log \mu = \alpha + \beta x \qquad (14)$$

the LHS and RHS can both take on any real value. Recall that $\log \mu$ is the natural parameter for the Poisson distribution and the log link is the canonical link for a GLM with a Poisson random component. A *Poisson log-linear model* is a GLM that uses a Poisson random component and the log link.

Note that the GLM represented by Equation (14) can be rewritten as

$$\mu = e^{\alpha+\beta x} = e^{\alpha}(e^{\beta})^{x}. \qquad (15)$$

Thus, a 1-unit increase in X has a multiplicative impact of e^{β} on μ, i.e., if we increase X by 1 unit in (15), we have

$$e^{\alpha}(e^{\beta})^{x+1} = e^{\alpha}(e^{\beta})^{x}e^{\beta} = (e^{\beta})\mu.$$

If $\beta = 0$, then $e^{\beta}=e^{0}=1$, so the mean of Y does not change as X changes. If $\beta > 0$, then $e^{\beta} > 1$ and the mean of Y increases as X increases. If $\beta < 0$, the mean of Y decreases as X increases.

Example (horseshoe crabs & satellites) (Sec 4.3.2)

Table 4.2 (pp. 82-83) presents data on a study of nesting horseshoe crabs. The entire data set is available on the course website.

Although not directly related to research in the health sciences, horseshoe crabs do have an important medical application. An extract of the horseshoe crab's blood is used by the pharmaceutical industry to ensure that their products, especially intravenous drugs and vaccines, are free of bacterial contamination. No other test works as easily or reliably for this purpose.

In the study for which the data are given in Table 4.2, each female horseshoe crab had a male crab attached to her in her nest. The study examined factors that affect whether the female crabs had any other males, called *satellites*, nearby. The outcome variable for this study was Y = # of satellites for each female crab. The explanatory variables included the female crab's color, spine condition, weight (kg),and carapace (shell) width (cm). In this section, we consider X = carapace width as the only predictor of the # of satellites.

Figure 4.3, p. 84, in our text presents a scatterplot of # of satellites vs. carapace width, with the plotting symbol equal to the # of observations at each point. It is difficult to discern any particular pattern from this plot.

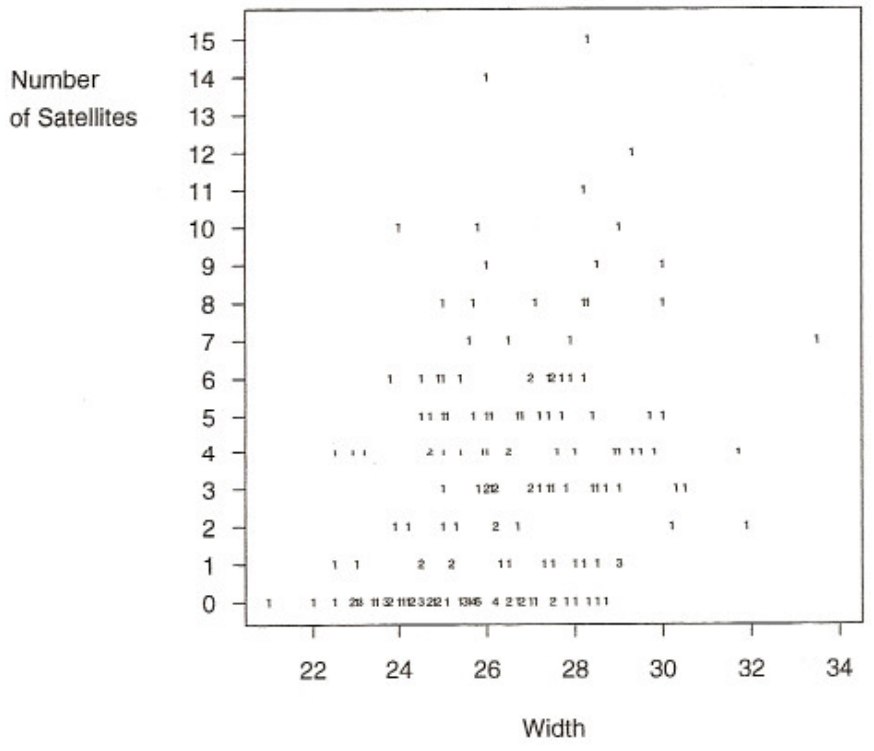

Figure 4.3 Number of satellites by width of female crab.

A more informative plot can be constructed if we group the female crabs according to width. Agresti uses the following intervals to group the data: ≤ 23.25, 23.25 – 24.25, 24.25 – 25.25, 25.25 – 26.25, 26.25 – 27.25, 27.25 – 28.25, 28.25 – 29.25, > 29.25. The mean # of satellites is then calculated within each category of carapace width, and plotted vs. the mean width in each category (Figure 4.4, p. 85).

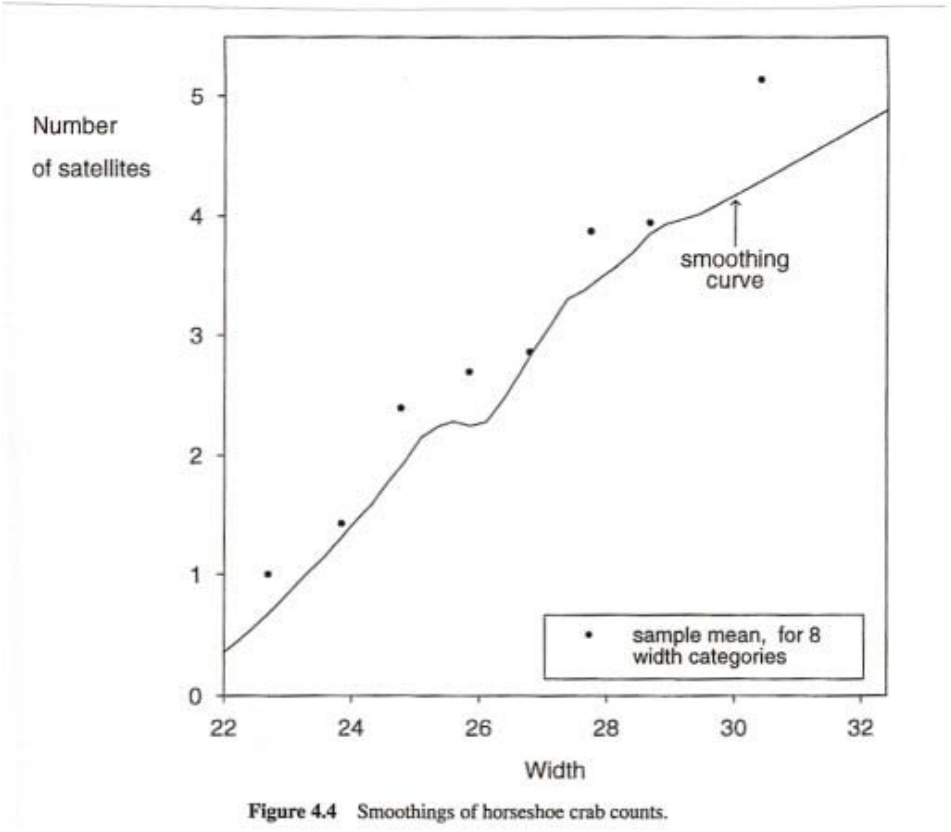

Figure 4.4 Smoothings of horseshoe crab counts.

One can also "smooth" the data by fitting a curve to the data points without assuming a particular functional form for the relationship between X & Y. (One example of such a curve is called a *loess plot*), One such curve is presented in Figure 4.4. Both the plot based on the grouped data and the smoothed curve indicate a strong increasing trend.

Let μ = expected # of satellites for a female crab and let X = carapace width of the crab. We first attempt to fit the log-linear model given by Equation (14) above. PROC GENMOD yields the following fitted model:

$$\log \hat{\mu} = \hat{\alpha} + \hat{\beta}x = -3.305 + 0.164x \, .$$

The ASE of $\hat{\beta}$ is 0.020, and an approximate significance test of H_0: $\beta = 0$ can be performed by calculating

$$z = \frac{.164}{.020} = 8.2 \, ,$$

which yields a highly significant p-value using the standard normal.

Thus, we conclude that there is a significant (positive) linear association between # of satellites & width of carapace.

For any given carapace width x, we can obtain the estimated # of satellites using the corresponding fitted value. So, for example, at the mean carapace width of 26.3 cm, we obtain the fitted value

$$\hat{\mu} = e^{\hat{\alpha}+\hat{\beta}x} = e^{-3.035+0.164(26.3)} = 2.74 .$$

In other words, based on the fitted GLM, our "best" estimate of the # of satellites for a crab with width 26.3 is 3. (One of the horseshoe crabs in the study had a carapace width of 26.3 cm and the observed # of satellites for her was 1.)

From this model, $e^{\hat{\beta}} = e^{.164} = 1.18$, which is the multiplicative effect on the fitted value of μ for each 1 unit (1 cm) increase in X. So, for example, the fitted value at x = 26.3 + 1 is 1.18(2.74) = 3.23 which can also be obtained by

$$\hat{\mu} = e^{-3.305+.164(27.3)} = 3.23 .$$

Another way to say this is that a 1 cm increase in carapace width yields an 18% increase in the estimated mean # of satellites.

The plot in Figure 4.4 suggets that a GLM with an identity link might also be used to describe the relationship between carapace width and # of satellites. PROC GENMOD in SAS yields

$$\hat{\mu} = \hat{\alpha} + \hat{\beta}x = -11.53 + 0.55x .$$

The ASE of $\hat{\beta}$ is 0.059, and once again an approximate test of H_0: $\beta = 0$ using the standard normal approximation to $\dfrac{\hat{\beta}}{\text{ASE}(\hat{\beta})}$ yields a highly significant result. In this model, the effect of X on μ is additive rather than multiplicative, i.e., a 1 cm increase in carapace width corresponds to a predicted increase of $\hat{\beta}$ = .55 in the estimated # of satellites. For example, the fitted value at the mean width of 26.3 cm is $\hat{\mu} = -11.53 + .55(2.63) = 2.93$; at x = 27.3, it is 2.93 + .55 = 3.48.

Figure 4.5, p. 86 in our text, compares the fitted values vs. observed values for both the log link and the identity link. For relatively larger and relatively smaller width categories, it appears that the GLM with the identity link fits the data a little better. Neither model fits the data particularly well in the middle of the range of carapace widths. Formal methods for determining if a Poisson regression model provides an adequate fit to the data are covered in Sec. 4.4.2 & 4.4.3. We will postpone our discussion of testing the GOF of GLM's until Chapter 5, when we consider logistic

(Figure 4.5 is given on the following page.)

regression (i.e., logit) models.

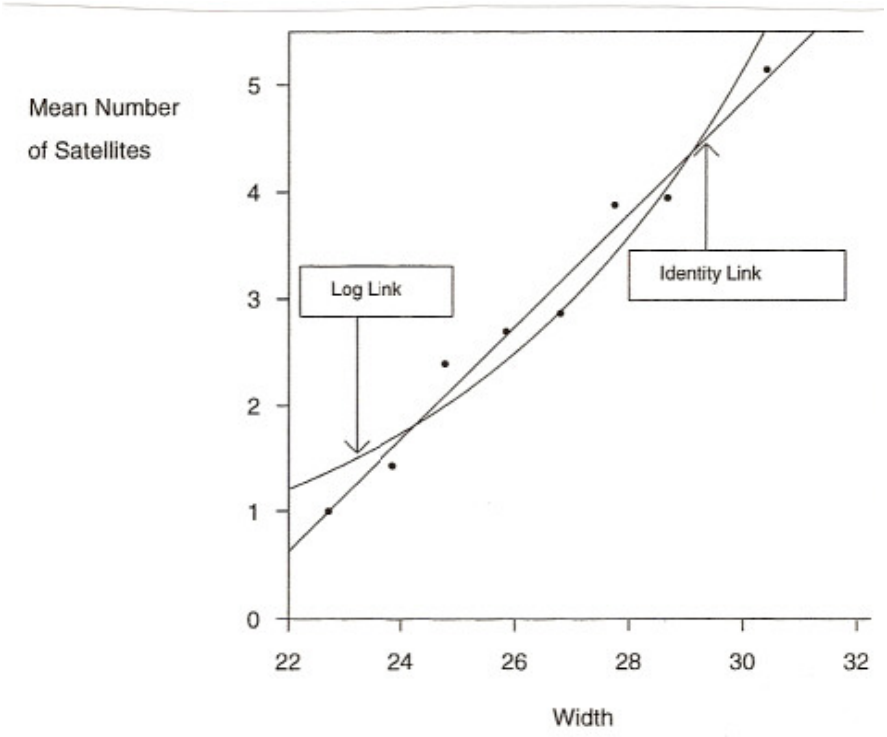

Figure 4.5 Estimated mean number of satellites for log and identity links.

Poisson Regression for Rate Data (Sec. 4.3.3)

In clinical research, Poisson regression is used most commonly when one wishes to model the incidence (or incidence density) of an event as a function of one or more explanatory variables.

For example, one could be interested in modeling the relationship between incidence of new melanoma cases per 100,000 population and age.

Let Y denote the "count" we are interested in (# of cases of disease, etc.) and let t denote the *index*; that is, the unit of space or time associated with Y. In Epi studies, t usually denotes the size of the "population at risk." Thus, the rate we are interested in modeling is $\frac{Y}{t}$, where t is a known constant, and the expected rate is $\frac{\mu}{t}$. A log-linear model for the expected rate is given by

$$\log\left(\frac{\mu}{t}\right) = \alpha + \beta x$$

or $$\log \mu - \log t = \alpha + \beta x. \qquad (16)$$

The ″adjustment term″ to the log link on the LHS of Equation (16), -log t, is called an *offset*. PROC GENMOD in SAS can fit GLM's that have offsets.

For the GLM in Equation (16), the expected # of events satisfies $\mu = te^{\alpha+\beta x}$, which implies that μ is proportional to t, with proportionality constant depending on the value of the explanatory variable X. So, under this model, in the melanoma example, for a particular age group, doubling the population size t would also double the expected number of cases of melanoma.

Example (MVA rates for elderly drivers) (Sec 4.3.4)

Agresti cites a cohort study published in the American Journal of Epidemiology in which a sample of n = 16,262 Medicaid enrollees aged 65-84 were followed for up to 4 years each. The total person-time for women in the sample was 17,300 person-years. The observed # of MVA's in which an injury was sustained by one of the women in the study was 175, yielding an incidence density of .01012 per year of driving, or 10.12 per 1,000 person-years of driving. The men were followed for a total of 21,400 person-years and received a total of 320 injuries in MVA's. The incidence density for men is therefore .01495 per year of driving, or 14.95 per 1,000 person-years of driving.

Let μ denote the expected number of MVA's in which an injury occurs during an observation period of t thousand person-years. To model the effect of gender on the MVA injury rate, we use the GLM in Equation (16) with X = 0 for females and X = 1 for males (i.e., X is a *dummy variable* for males). The log of the accident rate = α for females and $\alpha + \beta$ for males.

Our calculation of the estimated rates tells us that $\hat{\alpha} = \log(10.12) = 2.31$, $\widehat{\alpha+\beta} = \log(14.95) = 2.70$, and therefore $\hat{\beta}$ = 2.70 – 2.31 = .39. To test whether the injury rates are the same for males and females, we can test H_0: $\beta = 0$ using the results from PROC GENMOD, which yields $\hat{\beta}$ =.39 and ASE($\hat{\beta}$) = .09. An approximate test based on the standard normal is highly significant.

Thus, we reject H_0 and conclude that $\beta > 0$, i.e., that the true MVA injury rate is higher for males. Note that this analysis does not take into account the possibly different yearly driving levels for males and females.

Over-Dispersion in Poisson Regression (Sec. 4.4.4)

Recall that in any Poisson distribution, the population mean and variance are the same. Thus, in random samples from a Poisson distribution, we expect the sample mean & sample variance to be approximately the same. If this is not the case, then we say that *overdispersion* is present in the sample.

For example, Table 4.4 in our text shows the sample mean & variance for the # of satellites in each carapace width category for the horseshoe crab data. In each category, the variance is much larger than the mean, with $\frac{s^2}{\bar{x}}$ ratios ranging from 1.6 to 6.2.

Table 4.4 Sample Mean and Variance of Number of Satellites

Width	Number Cases	Number Satellites	Sample Mean	Sample Variance
< 23.25	14	14	1.00	2.77
23.25–24.25	14	20	1.43	8.88
24.25–25.25	28	67	2.39	6.54
25.25–26.25	39	105	2.69	11.38
26.25–27.25	22	63	2.86	6.88
27.25–28.25	24	93	3.87	8.81
28.25–29.25	18	71	3.94	16.88
> 29.25	14	72	5.14	8.29

Overdispersion is usually the result of heterogeneity among the study subjects. In the horseshoe crab example, suppose that weight, color, and spine condition, in addition to carapace width, all affect a female crab's # of satellites. If we consider width as the only explanatory variable, then we are ignoring the effect of the other variables that affect the # of satellites. In other words, crabs having the same carapace width (or those falling in each of the categories in Table 4.4) are actually a "mixture" of crabs having various weights, colors, and spine conditions. Thus, our sample of horseshoe crab carapace widths should rightfully be considered as a sample from a mixture of several Poisson populations. This mixture of populations has a variance that is greater than the variance of a single homogeneous Poisson population; hence, we may observe sample variances much larger than sample means when we ignore the other important explanatory variables. Many methods have been proposed for dealing with overdispersion in modeling discrete data when a Poisson (or binomial) sampling distribution is assumed. A discussion of these methods is beyond the scope of this course. Agresti describes and illustrates an elementary method that can be very effective on p. 93 of our text.

As mentioned previously in class, there are several objective methods that can be used to determine if the Poisson regression model provides a good fit to a set of count data. These are discussed by Agresti in Section 4.4. In this section, he also describes 3 procedures for testing general hypotheses about Poisson GLM parameters: the *Wald, likelihood-ratio,* and *score* tests. In Section 4.5, Agresti discusses some computational aspects of fitting GLM's, including use of the Newton-Raphson method, which is the most commonly used technique for finding MLE's when closed form expressions for the estimators are not available. He also discusses use of the likelihood function in performing statistical inference for GLM's and illustrates how the *deviance* between 2 possible models can be used in model fitting. We will consider similar methods for logistic regression models in Chapter 5.