12

Regression Models for Count Data

12.1. Introduction

In some epidemiological or clinical studies, the response of interest consists of a count, such as the number of cells that show definite evidence of differentiation, or the number of repeated infections experienced by a subject. The values recorded will be only non-negative integers.

In some instances, it may be possible to analyze observed data that are counts using the methods of multiple linear regression that we described in chapter 10. However, regression methods are available that are better suited to response measurements that are counts, and we discuss the most commonly used method, which is known as Poisson regression, in this chapter.

Because it often provides a satisfactory representation for the variability observed in count data, the Poisson distribution plays a role in their analysis that is similar to that of the normal distribution in multiple linear regression, and the binomial distribution in logistic regression. The first occasion when the Poisson distribution was used to characterize observations that were counts appears to have occurred at the end of the 19th century, when Ladislaus von Bortkiewicz [21] showed that, over a 20-year period, the annual number of deaths attributed to horsekicks suffered by corpsmen in each of 14 Prussian army corps could be fitted very convincingly by a Poisson distribution. However, the name Poisson derives from a French mathematician, Siméon-Denis Poisson, who derived the mathematical form of the distribution.

A more recent, slightly unusual, medical example in which the Poisson distribution was used to summarize the variation in observed counts was in the analysis of a randomized trial, conducted by Fallowfield et al. [22], that was

Table 12.1. The results of a Poisson regression analysis of the number of focussed and/or open questions asked by a physician during a patient consultation

Explanatory variable	Estimated regression coefficient	Estimated standard error	Test statistic	Significance level (p-value)
Course	0.24	0.05	5.06	<0.001
Physician sex	0.11	0.05	2.09	0.037
Seniority	−0.02	0.05	−0.45	0.651

designed to study the effect on physician communication skills of an intensive three-day training course.

Here, we ignore additional trial complexity, and consider a comparison of 80 doctors who were randomized to attend the three-day course with 80 doctors who were randomly chosen not to attend. We also restrict attention to a single outcome measure, namely the number of focussed and/or open questions asked by a physician during a patient consultation that occurred three months after the course ended, or three months after randomization for those physicians who did not receive any communication skills training. The course was designed to increase the frequency of such questions. For each physician, data were available from two consultations; to avoid undue complexity, we ignore the expected correlation between counts for the same physician and simply assume we have 160 observed counts for both the treatment group, i.e., those physicians who received training, and the control group.

Table 12.1 summarizes the results of a Poisson regression analysis of the number of focussed and/or open questions asked. The regression model included three explanatory variables that coded course attendance (yes = 1, no = 0), physician sex (female = 1, male = 0) and physician seniority (senior = 1, junior = 0). Readers will observe immediately that the format of this table is similar to those we first introduced in chapter 10 and subsequently encountered in chapter 11.

12.2. The Model for Poisson Regression

The theoretical formula from which we can calculate probabilities for counts that follow a Poisson probability function is characterized by a single parameter that is usually represented by the Greek letter λ. Conveniently, λ turns out to be the theoretical mean of the corresponding Poisson distribution,

so that if we have an estimated value for λ, we can immediately calculate the corresponding probability that a count equal to y is observed in a Poisson distribution with mean λ. Since λ is the only adjustable parameter in this Poisson model for the variation in observed counts, it is natural to link λ to the values of explanatory variables of interest. Because the mean, λ, of a Poisson distribution must be greater than zero, it would be unsuitable simply to assume that

$$\lambda = a + b_1X_1 + ... + b_kX_k = a + \sum_{i=1}^{k} b_iX_i,$$

where $X_1, \ldots, X_k$ represent the values of various explanatory variables, such as coding for the sex of a physician. Unless we restrict the values of a and the regression coefficients $b_1, ..., b_k$, the right-hand side of this equation for λ could sometimes be a negative value.

The logarithmic transformation is a remedy for this dilemma; the sign and magnitude of log λ is completely unrestricted, making the logarithm of the Poisson mean a natural choice to equate to the expression, $a + \Sigma_{i=1}^{k} b_iX_i$, the component of the Poisson regression model which is the same as that which occurs in other regression models. Thus, if $X_1, \ldots, X_k$ are potential explanatory variables whose values we wish to use to model variability in a response measurement, Y, that is thought to follow a Poisson distribution, then using the equation

$$\log \lambda = a + b_1X_1 + \ldots + b_kX_k$$

is a natural way of allowing the measured values of these explanatory variables to account for the variability in observed values of Y.

As in other regression models that we have previously considered, if a particular regression coefficient, say b_i, is zero, then the corresponding explanatory variable, X_i is not associated with the response, Y. Thus, if there is no evidence to contradict the hypothesis that b_i equals 0, then we probably can omit X_i from a Poisson regression model for the observed data. As we discussed in §11.2 for the case of logistic regression, a suitable statistic for testing the hypothesis that the regression coefficient, b_i, equals zero is

$$T = \frac{|\hat{b}_i|}{\text{est. standard error}(\hat{b}_i)}.$$

The results of an analysis may also be presented in terms of the ratio

$$\frac{\hat{b}_i}{\text{est. standard error}(\hat{b}_i)},$$

which is equal to T, apart from the sign. The latter is the ratio found in table 12.1. Whichever version of this test statistic is used, the conclusion regarding the associated explanatory variable, X_i, is the same.

The explanatory variables used in fitting the Poisson regression model summarized in table 12.1 are all binary ones that encode whether or not a physician in the study attended the training course, was a female, and was more senior.

There is considerable evidence in the study data that the estimated regression coefficient associated with attending the course is significantly different from zero, establishing a behavioural effect that is associated with the training provided. There is also some evidence of an effect associated with a physician's sex, but there is no evidence of different behaviour patterns between senior and junior physicians. The signs of the estimated regression coefficients for course attendance and physician sex each indicate that the estimated Poisson mean is larger if the physician is female or if he or she attended the communication skills training.

The results of an analysis based on a Poisson regression model can also be described in terms of a rate ratio or 'relative rate'. If b_j is the regression coefficient associated with a particular binary explanatory variable, such as course attendance, $\exp(b_j)$ represents the ratio of the rate at which the events of interest occur among physicians who received the skills training compared to those who did not. Thus, the key feature of the analysis that we can distill from table 12.1 is that the rate at which physicians asked focussed and/or open questions, adjusted for sex and seniority, is $\exp(0.24) = 1.27$ times greater after attending the training course. And if we use the estimated standard error for $\hat{b}_1$ of 0.05 to derive the 95% confidence interval $0.24 \pm 1.96(0.05)$, i.e., (0.14, 0.34), for b_1, then a corresponding 95% confidence interval for the relative rate is (exp(0.14), exp(0.34)) or (1.15, 1.40).

Of course, the importance of an effect may be linked to its absolute rather than relative size. In this communication skills study, the relative rate effect of 1.26 was associated with a mean number of 6.54 focussed and/or open questions asked during a patient consultation by a physician in the trained group compared with a mean of 5.14 in the control group. Providing such information, in addition to the Poisson regression results listed in table 12.1, is sensible and informative.

As in the case of logistic regression, the calculations involved in fitting a Poisson regression model to observed data are known as maximum likelihood estimation, the details of which are beyond the intended scope of this book. Even though many software packages that are now available will fit Poisson regression models, there are some aspects of these models that may require careful attention in any particular analysis. Thus, readers may wish to consult a statistician when the use of Poisson regression seems appropriate. However, we hope that our brief introduction to this regression model for count data has been informative, and will enable readers to understand the use of this statistical methodology in published papers.