Chapter 4

Modelling Counts - The Poisson and Negative Binomial Regression

In this chapter, we discuss methods that model counts. In a longitudinal setting, these counts typically result from the collapsing repeated binary events on subjects measured over some time period to a single count (e.g., number of episodes of diarrhea, as in the HIV/drinking water study). We start by discussing the most popular distribution for modelling counts, the Poisson distribution and Poisson regression (note, in the next chapter, we link Poisson regression directly to survival analysis). The chapter is finished by presenting a slightly bigger model, the negative binomial distribution, which handles some situations where the Poisson model is a poor fit.

4.1 Poisson Distribution

The Poisson distribution is often used to model information on *counts* of various kinds, particularly in situations where there is no natural "denominator", and thus no upper bound or limit on how large an observed count can be. This is in contrast to the Binomial distribution which focuses on observed proportions. Possible examples of count data where a Poisson model is useful include (i) the number of automobile fatalities in a given region over year intervals, (ii) the number of AIDS cases for a given risk group for a series of monthly intervals, (iii) the number of murders in Chicago by year, (iv) the number of server failures for a web-based company by year, and (v) the number of earthquakes of a certain magnitude in a seismically active region by decade. In each of these examples, there is no reasonable denominator associated with the counts–even in (i) and (iii) where population counts might be appropriate to capture incidence or event proportions, these totals may be difficult to ascertain or define in such a way that members of the population hold at least a similar level of risk of an event.

When a Poisson model is appropriate for an outcome Y, the probabilities of observing any specific count, y, are given by the formula:

$$Pr(Y = y) = \frac{\lambda^y e^{-y}}{y!}, \tag{4.1}$$

where λ is known as the population rate parameter, and $y! = y \times (y-1) \times \cdots \times 2 \times 1$. Such a Poisson random variable Y has expectation $E(Y) = \lambda$, and variance $Var(Y) = \lambda$. The fact that the expectation and variance agree provides a quick check on whether a Poisson model might be appropriate for a sample of observations. In the examples above, the parameter λ describes the (i) rate of automobile fatalities per year, (ii) the AIDS incidence rate per month, etc.

The number of pedestrian fatalities due to automobile accidents in Solana County, california was three in 1999. To illustrate the Poisson distribution, suppose that we believe that the annual rate for such fatalaties is two per year and that the distribution is Poisson. With $\lambda = 2$, Figure 4.1 shows the probability density associated with the Poisson distribution. The density indicates that the probability of observing a count of three fatalities in a specific year is just$\frac{2^3 \times e^{-3}}{3!} = 0.180$; the probability of no fatalities in a year is $2^0 \times e^{-2} = 0.135$. the probability of observing three or more pedestrian fatalities is 0.323, so that a year occurs about once out of every three on average assuming that we know $\lambda = 2$.

Suppose we observe m independent observations, $y_1, \ldots, y_m$, of a Poisson distribution with an unknown rate parameter λ. How do we use these observations to estimate λ? The average of the observed ys is always a reasonable estimator of the mean of Y, so this also seems appropriate to estimate λ. That is, we use the estimator

$$\hat{\lambda} = \frac{Y_1 + Y_2 + \cdots + Y_m}{m}. \tag{4.2}$$

This is, in fact, also the maximum likelihood estimator of λ.

Time–correlation. Of course, repeated observation of counts can also occur in series that are not arranged in any time sequence. The next subsection discusses such an example in detail.

Figure 4.1: Poisson probability distribution for $\lambda = 2$

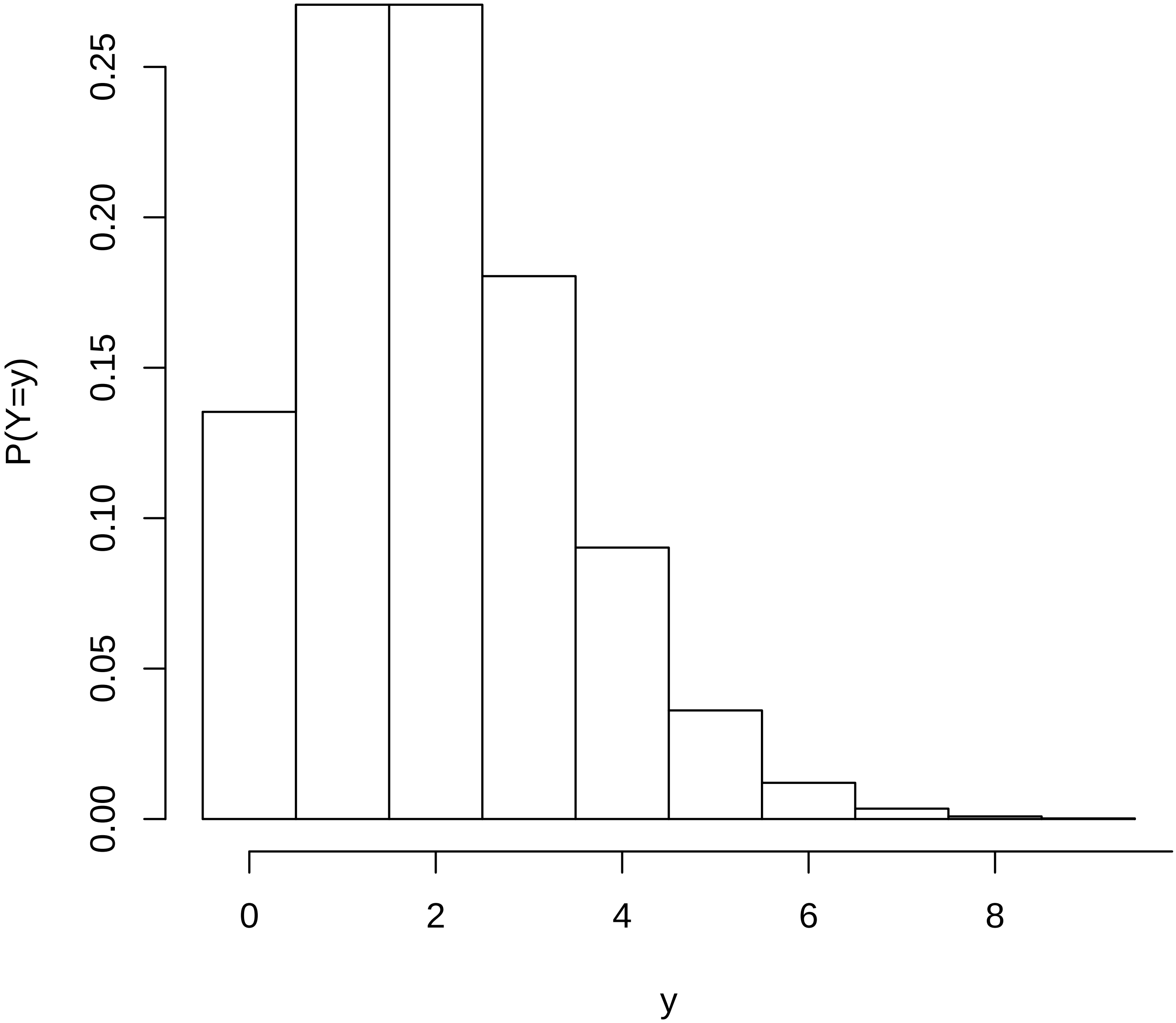


Table 4.1: Data on Individual Team Scores from the 2002 World Cup in Soccer

Number of Goals	Observed Number of Teams	Expected Number of Teams*
0	37	36.4
1	47	45.8
2	27	28.8
3	13	12.1
4	2	3.8
5+	2	1.2

* assuming a Poisson distribution

4.1.1 Example–The World Cup in Soccer

In the 2002 World Cup in soccer, held in Japan and South Korea, 64 soccer games were played. Suppose we wished to understand the chances that, in a given game, one team would score no goals, or one goal, or any other specific number of goals. Based on the scores of all the games played provides 128 data points on the number of goals a team scores in a single game. The distribution of these 128 observations is shown in the middle column of Table 4.1.

It is easy to calculate the average of the 128 observations, that is the average number of goals scored by a single team in a game, since it is simply the total number of goals scored, 161, divided by 128, namely 1.258. Similarly, the observed variance of the observations is 1.500. Although this is somewhat bigger than the mean, it suggests that it might be reasonable to use a Poisson model to describe the distribution of the observations, that is, to assume that the observations arise from sampling from a common Poisson distribution. The maximum likelihood estimate of the rate parameter is then just the average as noted in (x.x), yielding $\hat{\lambda} = 1.258$. Thus, the expected number of goals any single team scores in a game is a little more than one goal per game.

Using this estimate we can easily predict the expected number of times that a team score a specific number of goals in a game. For example, using $\hat{\lambda} = 1.258$, we calculate that the probability of no goals for a team is $1.258^0 \times e^{-1.258} = 0.284$, so that the expected number of occurrences of zero goals amongst 128 observayions is $128 \times 0.284 = 36.4$. This is close to the

Table 4.2: Data on Individual Team Scores from the 2002 World Cup in Soccer

Number of Goals	Observed Number of Teams	Expected Number of Teams*
0	37	36.4
1	47	45.8
2	27	28.8
3	13	12.1
4	2	3.8
5+	2	1.2

* assuming a Poisson distribution

observed number of zero goal occurrences (37) as shown in Table 4.1. Expected number of occurrences of other number of goals are also shown in Table 4.1, allowing easy comparison with the observed numbers. This shows that the Poisson model fits the obseved frequencies very closely. Note that the additional variation in the observations noted above—beyond what would be predicted by the Poisson model—might easily be explained that different teams do not share exactly the same λ, that is, scoring rate. We examine this possibility later in this chapter.

Not all count data is suitably modeled by a Poisson distribution. For example, Table 4.2 gives data on the number of homocides in Chicago for 31 years from 1965 to 1995. The average number of homocides per year is 768.2; the variance of the 31 observations is 16,505, morethan 20 times larger so that we would expect that the Poisson distribution would provide a very poor fit to these data as is amply demonstrated in Table 4.2.

4.2 Poisson Regression

In the World Cup example, we have already indicated that assuming a common goal scoring rate for all countries in every World Cup, dating back to 1930, may be implausible. That is, we wish to consider the possibility that the rate may differ across subgroups of the data, in this case defined by geography and time. In most disease investigations, understanding the difference between incidence rates across different groups is the primary focus. for example, in the Water Intervention Trial of Chapter 1.x, the goal of the study is to deternmine

Table 4.3: Yearly Data on Chicago Homocides from 1965 to 1995

Year	No. of Homocides	Year	No. of Homocides	Year	No. of Homocides
1965	397	1976	817	1986	750
1966	511	1977	827	1987	687
1967	551	1978	792	1988	661
1968	647	1979	853	1989	752
1969	724	1980	859	1990	849
1970	807	1981	879	1991	921
1971	824	1982	671	1992	939
1972	708	1983	732	1993	859
1973	867	1984	739	1994	926
1974	962	1985	667	1995	814
1975	822				

whether the rate of gastrointestinal events depends on whether participants have filtered or unfiltered water available. First, we look at the simplest case of two groups where the Poisson rate for the events of interest is allowed to different.

Suppose the two groups correspond to a binary covariate X. for example, $X = 1, 0$ corresponding to assignement to filtered and unfiltered water in the data of Chapter 1.x. Our model for the data now assumes that individuals with $X = 1$ experience events according to a Poisson distribution with rate parameter λ_1, with $X = 0$ indiviuals experiencing Poisson rate λ_0. We now want to estimate both λ_0 and λ_1, and examine evidence as to whether the two rates differ, allowing for sampling variation.

For estimation, we simply restrict the average of (x.x) to the two subgroups, for example, estimating λ_1 by the average event count for individuals with $X = 1$. This can be represented symbolically by

$$\hat{\lambda}_1 = \frac{\sum_{i=1}^{m} X_i Y_i}{\sum_{i=1}^{m} X_i} \qquad \hat{\lambda}_0 = \frac{\sum_{i=1}^{m} (1 - X_i) Y_i}{\sum_{i=1}^{m} (1 - X_i)}. \tag{4.3}$$

Now that we have described the Poisson distribution in the univariate case, the next task is to model how the distribution changes as a function of explanatory variables. Since the Poisson distribution has only one parameter, namely the mean rate, λ, we have little choice but to model λ as a function of x, or $\lambda(x)$. The general regression problem can be understood easiest in the two-sample case, that is when X represents two groups, such as the HIV water-intervention study described in section (??): X=1 (treatment), =0 (placebo). In this simple case, we need simply to estimate $\lambda(1)$ and $\lambda(0)$ to describe statistically the association of X (treatment) on Y (GI illnesses). We can parameterize this relationship as a simple log-linear model,

$$log(\lambda(X)) = a + bX, \tag{4.4}$$

which implies the distribution of GI counts among the placebo group is Poisson(e^a) and Poisson(e^{a+b}) among the treatment group. Further, the coefficient b has a convenient interpretation as the natural log of the incident rate ratio (IRR) comparing the treatment and the placebo groups:

$$IRR = e^b = \frac{\lambda(1)}{\lambda(0)} \tag{4.5}$$

In the simply two-sample case, the estimates of the coefficients are simply derived from the estimates of the mean rates, as discussed above, or, $\hat{b} = log(\lambda(1)/\lambda(0))$ and $\hat{a} = log(\lambda(0))$. Consider now the general case of several (possibly continuous) covariates, $X_1, X_2, \ldots, X_p$ and models of the form:

$$log(\lambda(X_1, X_2, \ldots, X_p)) = a + b_1X_1 + b_2X_2 + \cdots + b_pX_p \tag{4.6}$$

The interpretation of the coefficients are similar to (4.4); for example b_1 is the IRR for comparing a one unit increase in X_1, keeping all other variables $(X_2, \ldots, X_p)$ fixed.

4.2.1 Example—The World Cup in Soccer, Part 2

We illustrate a two group comparison first, addressing the question of whether the goal scoring rate was different in 2002 from 30 years earlier at the World Cup in 1970 in Mexico. This provides a quick look at the issue of whether modern soccer is more defensive leading to fewer goals on average. Assuming a Poisson distribution for the number of goals scored by a team in a single game as before, we now estimate a different rate for the two world Cups. The total number of goals in 1970 was 95, arising from 64 observations (that is, 32 games); this yields the estimate $\lambda_{1970} = \frac{95}{64} = 1.484$, as compared to $\lambda_{2002} = 1.258$. Thus, the goal scoring rate is indeed lower in 2002, but perhaps this merely reflects the chance effects of sampling variation.

4.3 Adapting Poisson Regression to Variation in Follow-up Periods

One concern in using the Poisson distribution to model the number of goals scored by a soccer team is that not all games last the same length of time. A few World Cup games—in the later stages of the competition—are extended to 120 minutes when a game is tied after 90 minutes. This was the case in early World Cup games but a further complication was introduced in 19xx. In that year and subsequent competitions, the ”Golden Goal” rule was introduced so that extra time was added in such tied games only until one or other team scored a goal. Either way, we expect a team to score more goals the more minutes they play, and this was ignored in Section 2.2.1.

A more epidemiological example of this problem is illustrated by the data from the Water Intervention Trial introduced in Chapter 1.x. In this case, there was significant variation in the amount of time the subjects were enrolled in the study and any reasonable estimate of the rates of diarrhea within treatment groups should account for this variation (see figure 4.2 for the distribution of total days in study). The solution to varying time at risk among individuals is quite simple and comes froms a property of the Poisson distribution: if λ is the mean rate for one unit of time (e.g., a year) then the rate for a general interval of time, say T is λT. One consequence is that if one is estimating the mean rate for one unit of time, then instead of using the simple average of counts, the estimate is:

$$\hat{\lambda} = \frac{Y_1 + Y_2 + \cdots + Y_m}{T_1 + T_2 + \cdots + T_m}$$

where Y_i is the count recorded for the ith object (e.g., subject) and T_i is the time the subject was at risk. In words, the estimated rate per unit of time is the total count divided by the total time at risk. In the regression context, if (4.6) is the log(rate) for unit of time given the covariates, $X_1, X_2, \ldots, X_p$, then the rate for an interval T is $log(\lambda(X_1, X_2, \ldots, X_p)) + log(T)$. Thus, one essentially treats $log(T)$ as a special additional covariate in the regression where the coefficient is fixed to be 1; this special covariate is sometimes referred to an *offset* and is entered as such in most statistical packages.

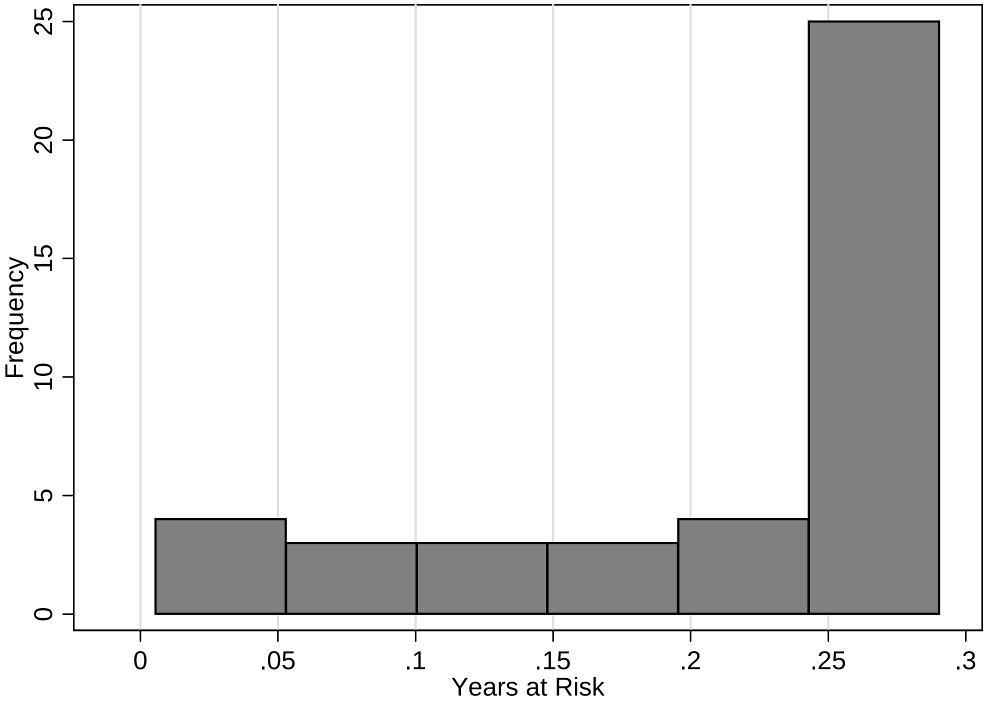
4.4 Negative Binomial Regression

The maximum likelihood procedure used to 1) derive the estimates and 2) provide the estimated variability (standard errors) of those estimates in Poisson regression makes a strong (and testable) assumption that every subject within a covariate group (a population that has all the same values for $X_1, X_2, \ldots, X_p$) has the same underlying rate of the outcome. As mentioned above, this also implies that the variability of counts within covariate group is equal to the mean, or:

$$var(Y(X_1, X_2, \ldots, X_p)) = exp(a + b_1 X_1 + b_2 X_2 + \cdots + b_p X_p). \tag{4.7}$$

If this fails to be true, the estimates of the coefficients can still be consistent using Poisson regression, but the standard errors can be biased and they will be too small. More typically

Figure 4.2: Distribution of time at risk in HIV water rrial



than not, we would not expect that we have measured every variable that contributes to the rates of events, so there will always be residual variation in the rates of events among people who all have the same covariate values. This is particularly true for the HIV Water Intervention trial, as we measure only one variable, treatment, and it seems unlikely that the underlying rate of GI illness is the same for all subjects within a treatment group. Specifically, there is variation in the severity of HIV disease among the participants in the study, and given treatment is randomized, we would expect this variation to exists in the treatment groups. Fortunately, there is an extension to Poisson regression, called negative-binomial regression, which can account for greater than Poisson variation and is based on the negative binomial distribution.

Consider again the univariate case - the negative binomial probability distribution of Y is:

$$P(Y=y) = \left(\frac{r}{r+\lambda}\right)^r \frac{\Gamma(r+y)}{\Gamma(y+1)\Gamma(r)} \left(\frac{\lambda}{r+\lambda}\right)^y, \qquad (4.8)$$

where Γ is the gamma function. The mean of the negative binomial distribution (like the Poisson) is λ but the variance is $\lambda + \lambda^2/r$, where r is called the dispersion parameter; figure 4.3 compares models with equivalent means, but increasingly small dispersion parameters (increasingly large variances). As the figure suggests, as r gets large (and λ is fixed), then the negative binomial converges to a a Poisson distribution, i.e., $var(Y) \rightarrow \lambda$. This means that the negative binomial model is a more general model than the Poisson and it can be motivated as a mixture of Poisson distributions. Specifically, if the underlying rate of events for subject i is Poisson with rate λ_i, which is gamma distributed with mean λ and variance λ^2/r then the marginal distribution of the counts is (4.8). Thus, fitting a negative binomial model to the HIV Water Intervention Trial implies the subjects in each group who have a mean rate of interest, but the rates of subjects within the treatment groups differ around their corresponding mean rate. This is a general description of a random effects model on which we will focus in later chapters. It often has reasonable theoretical justification, because we often expect that we have not collected all the explanatory variables relevant to explaining the variation in the underlying rates of outcomes among the study subjects. In addition, we can test whether the negative binomial distribution is a signficant improvement over the Poisson regression by fitting both models and performing a likelihood ratio test; a small p-value would imply the negative binomial model is a significantly better fit than the Poisson model.

Extraploting from the univariate case to the regression scenario is equivalent to Poisson regression, that is the regression (mean) model is equivalent (4.6), but now the likelihood is based on the negative binomial (4.8). Below, we compare the Poisson and negative binomial fits to the HIV Water Intervention Trial data.

Figure 4.3: The negative binomial probability distribution with $\lambda = 4$

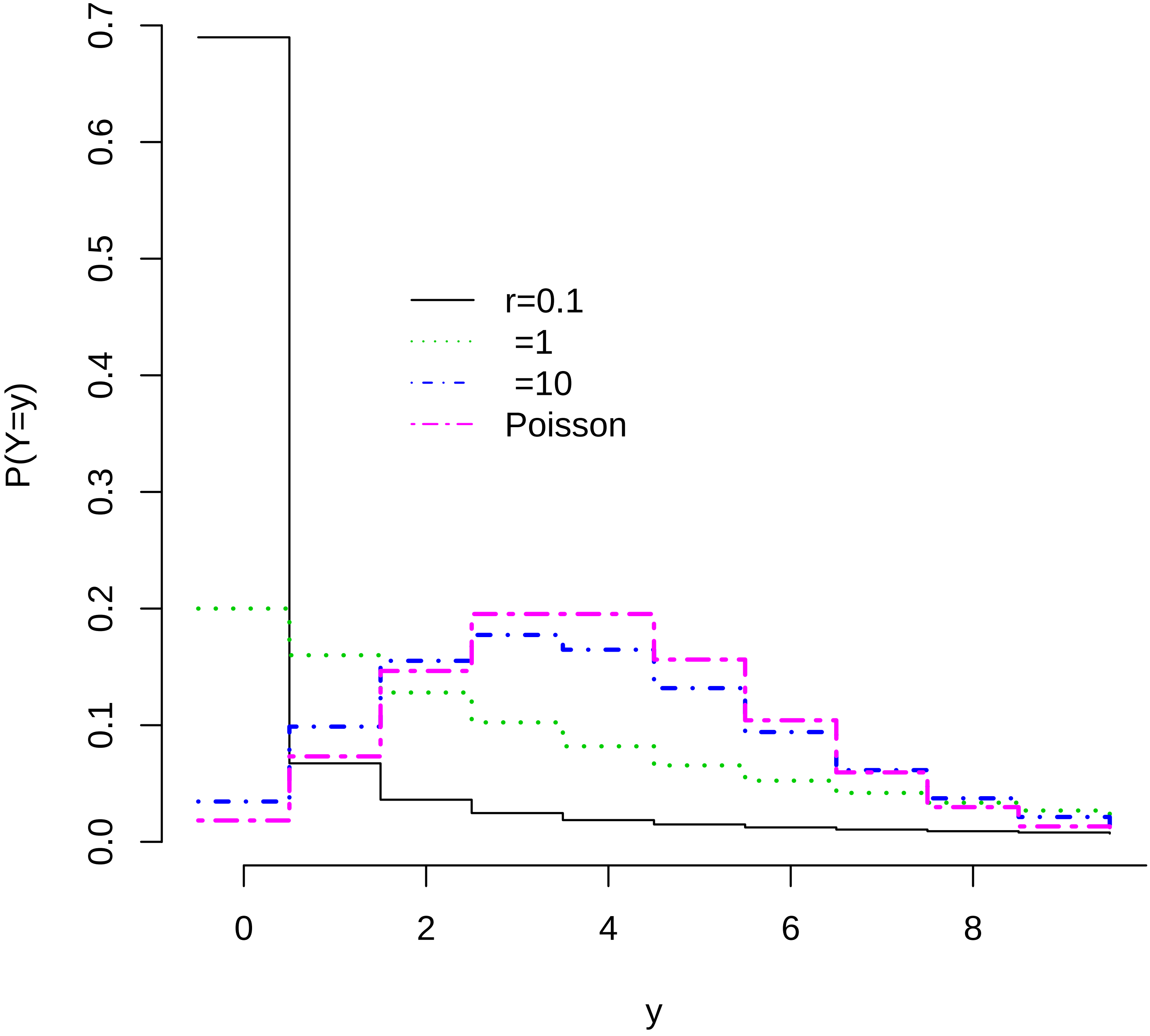


Table 4.4: Summary of data from HIV Water Intervention Trial

Group	Total Episodes	Total Follow-up Years	Incidence Rate
filtered	16	4.1	3.9
unfiltered	31	4.9	6.3

Table 4.5: Results from Poisson Regression of data from HIV Water Intervention Trial

Coefficient	Estimate	SE	p-value
β_0	1.37	0.25	< 0.001
β_1	0.46	0.31	0.13

4.5 Example—Water Intervention Trial

This analysis of this data is a simple two sample problem (filtered and unfiltered) with different follow-up times for each subject, so the estimate is precisely (ref???). We wish to fit a simple Poisson regression model,

$$\lambda(x) = exp(\beta_0 + \beta_1 x) \tag{4.9}$$

with x = 0 (filtered) and 1 (unfiltered). As discussed above, the maximum likelihood coefficient estimates use using only the summary information in table 4.3. Fitting the model

Table 4.6: Results from negative binomial regression of data from HIV Water Intervention Trial

Coefficient	Estimate	SE	p-value
β_0	2.05	0.57	< 0.001
β_1	0.29	0.72	0.68
log(dispersion)	-1.26	0.35	NA

provides the results in table 4.4, which translates to an IRR of 1.59, with an associated 95% confidence interval of $(0.87, 2.91)$. This suggests an increased, though not significant, rate of HCGI among those without the filters. As we just discussed, the standard errors of the coefficients are sensitive to the assumption that the data are Poisson distributed, which itself implies that the underlying rate of HCGI is the same for all subjects within the same treatment group; evidence that this is not true comes from the fact that the estimated coefficients of variation (CV) of the individual rates of HCGI for the treatment and sham group are 2.5 and 1.8, respectively, suggesting greater than Poisson variation (CV=1). Thus, a negative binomial model was fit to the same data (table 4.5), which results in an IRR of 1.34, with associated 95% confidence interval, $(0.33, 5.51)$. In addition, the likelihood ratio test comparing the relative fit of the negative binomial model to the Poisson model results in a very significant rejection of the Poisson model (p-value < 0.001). The negative binomial regression implies much wider confidence intervals and thus much less evidence for contribution of drinking water to HCGI.

4.6 Comments and Further Reading

4.7 Problems

Question 3.1 Fit a Poisson model to the data for the 1966 soccer World Cup. Using your estimated goal scoring rate, compute the expected frequencies for each goal count as in Table 3.1, and compare to observed frequencies. Comment on the adequacy of the fit of the model to the data. Assess whether the goal scoring rate in 1966 is different from that in 2002.

Question 3.2 Using Poisson regression, consider whether the crude continent measure for a country (South America, Europe, or the Rest of the World) influences goal scoring rate based on data for all the soccer World Cups. Allowing for this [possible effect, assess the evidence of a trend in goal scoring over time both overall, and then for each “continent” group separately.

References

Deeks, S. (1999). xxx. *xxx* **xx**, 750-762.