May 29, 2012

Generalized Linear Models (GLM)

In the previous chapter on regression, we focused primarily on the classic setting where the response y is continuous and typically assumed to have a normal distribution, at least approximately. However, in many environmental data analysis examples, the data to be modeled are clearly non-normal. For instance, we may record a binary response (e.g. survived or died, or infected or not infected, etc.) A zero-one variable is clearly non-normal. In many other common examples, the response of interest is a count, e.g. the number of eggs in a nest or number of vines on a tree. The Poisson distribution is often used to model count data and a Poisson regression can be used to relate count responses to predictors. This chapter introduces the *generalized linear model* to accommodate responses that follow non-normal distributions. The word "linear" appears here because the response is still modeled as a linear combination of predictors but the relation is not necessarily a direct relation as in the previous chapter. Instead, we need to introduce a *link* function that links the response to the linear predictor.

Before defining the generalized linear model, we start by introducing a special case – logistic regression.

1 Logistic Regression

In the classical regression framework, we are interested in modeling a continuous response variable y as a function of one or more predictor variables. Most regression problems are of this type. However, there are numerous examples where the response of interest is not continuous, but binary. Consider an experiment where the measured outcome of interest is either a success or failure, which we can code as a 1 or a 0. The probability of a success or failure may depend on a set of predictor variables. One idea on how to model such data is to simply fit a regression with the goal of estimating the probability of success given some values of the predictor. However, this approach will not work because probabilities are constrained to fall between 0 and 1. In the classical regression setup with a continuous response, the predicted values can range over all real numbers. Therefore, a different modeling technique is needed.

In the previous chapter on regression, one way to think about the model is in terms of conditional expectation: $E[y|x]$ is the conditional expectation of y given x. In a simple linear regression, we have $E[y|x] = \beta_0 + \beta_1 x$; that is, the conditional expectation is a linear function of x. If y is binary taking the values 0 and 1, then

$$E[y|x] = P(y = 1|x).$$

That is, in with a binary outcome, the regression of y on x is a conditional probability. If we label $y = 1$ as a "success", then the goal is to model the probability of success given x. The approach to this problem illustrated here is known as *logistic regression.* Note that other approaches are also possible, notably a *probit* regression.

In risk assessment studies, a *dose–response curve* is used to describe the relationship between exposure to a given dose of a drug or toxin say and its effect on humans or animals. Let us illustrate matters with an example.

Example. Beetles were exposed to gaseous carbon disulphide at various concentrations (in mf/L) for five hours (Bliss, 1935) and the number of beetles killed were noted. The data are in the following table:

Dose	# Exposed	# Killed	Proportion
49.1	59	6	0.102
53.0	60	13	0.217
56.9	62	18	0.290
60.8	56	28	0.500
64.8	63	52	0.825
68.7	59	53	0.898
72.6	62	61	0.984
76.5	60	60	1.000

If we let x denote the concentration of CS_2, then from the table it appears that the probability of death increases with increasing levels of x. Note that the mortality rate increases with increasing dose. Our goal is to model the probability of mortality as a function of dose x using a regression model.

If we consider for a moment the beetles exposed to the lowest concentration of 49.1 mf/L of CS_2, we see that a total of $n = 59$ beetles were exposed and 6 of them died. The *binomial probability* model is used to model outcomes of this type. A binomial experiment is one that consists of (i) n independent trials where (ii) each trial results in one of two possible outcomes (typically called success or failure which are recorded as 1 or 0 respectively), and (iii) the probability p of a success stays the same for each trial. In the beetle example at the lowest dose, we have $n = 59$ beetles. For each beetle, the outcome is either death (success) or alive (failure). We can let p denote the probability an individual beetle will die after five hours at this dose. If we let y denote the number of beetles that die, then y is a random variable with a binomial distribution. One can show that the probability that y assumes a value k where $k = 0, 1, \ldots, 59$, is given by the following formula:

$$P(y = k) = \binom{n}{k} p^k (1-p)^{n-k}. \tag{1}$$

In the current setup, the situation is more complicated than this because for each dose, we have a corresponding binomial distribution where the success probability p depends on x, the concentration of CS_2. Thus, our goal then becomes to model $p(x)$, the probability of death given an exposure to a CS_2 concentration equal to x. As mentioned above, the model

$$p(x) = \beta_0 + \beta_1 x$$

will NOT work since $p(x)$ must take values between 0 and 1. One of the standard ways of modeling the data in this situation is to use the *logistic regression function*:

$$p(x) = \frac{e^{\beta_0+\beta_1 x}}{1 + e^{\beta_0+\beta_1 x}}, \tag{2}$$

where β_0 and β_1 are the regression parameters of interest (similar to the intercept and slope parameters of a simple linear regression). Note that by construction, the values of $p(x)$ in (2) are constrained to lie between 0 and 1 as required.

Here are some notes about the logistic regression function in (2):

1. If $\beta_1 = 0$ then $p(x)$ is constant. That is, the probability of success will not depend on x.
2. $p(x)$ is an increasing function if $\beta_1 > 0$ and a decreasing function if $\beta_1 < 0$.
3. For a probability p, the function $p/(1-p)$ is called the *odds ratio*
4. One can show that

$$\log[p(x)/(1-p(x))] = \beta_0 + \beta_1 x.$$

The function $\log[p(x)/(1-p(x))]$ is called the *logit* of $p(x)$:

$$\text{logit}(p(x)) = \log[p(x)/(1-p(x))] = \beta_0 + \beta_1 x,$$

and is known as the *link* function for logistic regression. Note that the logit of $p(x)$ yields the usual linear regression expression.

The natural question now becomes – how do we use the data to estimate the parameters β_0 and β_1 in the logistic regression model? The answer is to use the method of *maximum likelihood estimation*. The logic behind maximum likelihood estimation is to determine the values of β_0 and β_1 what make the observed data most likely to have occurred. The method of maximum likelihood estimation is used very broadly in many statistical applications besides logistic regression. Maximum likelihood estimators often perform better than other types of estimation procedures in terms of being the most efficient use of data. Hence, maximum likelihood estimation is a very popular method of estimation in statistical practice. We now provide a brief introduction to maximum likelihood estimation.

1.1 A Brief Introduction to Maximum Likelihood Estimation

To illustrate ideas, we provide an example. Suppose a certain species of bird are found in two different locations, A and B say. 30% of the birds are infected with the West Nile virus in location A and 50% of the birds are infected with the virus in location B. Now suppose one has a sample of $n = 100$ birds from one of the locations, but it is not known which location. The $n = 100$ birds are tested and it is found that 40% of them have the West Nile virus. Let p represent the proportion of birds in

the population from which this sample of birds was obtained. Then either $p = 0.3$ or $p = 0.5$ depending on whether or not the birds are from location A or B respectively. Given the observed data (i.e. 40 out of 100 infected birds), which value of p is more likely? The idea of maximum likelihood is to find the value of the parameter(s) (in this case p) which makes the observed data most likely to have occurred.

If p is the probability that a randomly selected bird is infected, then $1 - p$ is the probability the bird is not infected. Using the binomial density (1), we have that the probability of observing k infected birds out of n birds is

$$L(p) = \binom{n}{k} p^k (1-p)^{n-k}.$$

We now think of L as a function of p, the success probability parameter and call it the likelihood. The goal of maximum likelihood estimation is to find the value of p that maximizes L, the likelihood. Since there are only two possible values for p, we can evaluate L at $p = 0.3$ and $p = 0.5$:

$$L(0.3) = \binom{100}{40} 0.3^{40}(1-0.3)^{60} = 0.00849$$

and

$$L(0.5) = \binom{100}{40} 0.5^{40}(1-0.5)^{60} = 0.01084.$$

Thus, $p = 0.5$ is the more likely value given that we observe 40 birds out of 100 with the infection. That is, the maximum likelihood estimate of p is $\hat{p} = 0.5$.

In most applications of maximum likelihood estimation applied to binomial probabilities, the probability of success is completely unknown. For instance, suppose in the previous example we simply have a sample of $n = 100$ birds from a population of interest and $k = 40$ of them are infected with the West Nile virus. Let p denote the proportion of the birds in the population that are infected. We can use maximum likelihood to estimate p in this modified example. Here, all we know is that $0 < p \leq 1$. Once again, the likelihood is

$$L(p) = \binom{n}{k} p^k (1-p)^{n-k} = \binom{100}{40} p^{40}(1-p)^{60}.$$

The goal is to find the value of p that maximizes $L(p)$ in this expression. Equivalently, we can maximize the natural logarithm of this expression which is easier to work with:

$$l(p) = \log[L(p)] = \log[\binom{100}{40}] + 40\log[p] + 60\log[(1-p)].$$

Taking the derivative of this expression with respect to p and setting it equal to zero allows us to find the maximum likelihood estimator (mle) of p. After doing the calculus, the mle is found to be

$$\hat{p} = k/n,$$

which is simply the sample proportion, a natural estimator of p. In this example, the mle is found to be $\hat{p} = 40/100 = 0.4$.

This example is a fairly elementary illustration of maximum likelihood estimation. The theory of maximum likelihood is used very extensively in much more complicated models with many parameters. From the theory of maximum likelihood, it follows that in most standard situations, the maximum likelihood estimators have approximate normal distributions provided the sample size is relatively large. Maximum likelihood estimators also tend to be nearly unbiased and they also tend to have smaller variances than other unbiased estimators which makes maximum likelihood estimation a very popular statistical estimation process.

1.2 Simple Logistic Regression

Now we can get an idea how maximum likelihood estimation is used in the logistic regression model. In the simple logistic regression model, the probability of success p depends on a regression variable x, and thus we write $p(x)$ as defined in (2). Suppose our data have been collected at values of $x_1, x_2, \ldots, x_m$. For each value x_i, suppose there were n_i trials with y_i successes, $i = 1, 2, \ldots, m$. In the beetle mortality example, the lowest dose of the gas was $x_1 = 49.1$. There were $n_1 = 59$ beetles exposed at this dose with $y_1 = 6$ deaths. Assuming whether or not the beetles survive are independent of one another, we can write the likelihood function as

$$L = \binom{n_1}{y_1} p(x_1)^{y_1}(1-p(x_1))^{n_1-y_1} \times \cdots \times \binom{n_m}{y_m} p(x_m)^{y_m}(1-p(x_m))^{n_m-y_m}.$$

Notice that this likelihood function L is actually a function of β_0 and β_1, the logistic regression parameters because

$$p(x_i) = \frac{e^{\beta_0+\beta_1 x_i}}{1+e^{\beta_0+\beta_1 x_i}}, \text{ for } i = 1, 2, \ldots, m.$$

Unfortunately, there do not exist formulas that give the estimates of β_0 and β_1 from a logistic regression in closed form as was the case in simple linear regression. Instead, iterative algorithms are needed to determine the maximum likelihood estimates of β_0 and β_1. Many software packages have the ability to fit logistic regression models. The two most popular algorithms for finding the maximum likelihood estimates are the *Newton–Ralphson algorithm* and the *iterated re–weighted least squares algorithm*. We will illustrate the fitting of a logistic regression model using the "glm" function in R which stands for *generalized linear model*. (In SAS you can fit the logistic regression model using PROC LOGISTIC.)

Example. We return to the beetle example above now. The R code for fitting the logistic regression model is

```
bliss=read.table("bliss.dat", header=T)
alive=bliss[,2]-bliss[,3]
deaths=bliss[,3]
concentration=bliss[,1]
phat=deaths/(deaths+alive)
fit <- glm(cbind(deaths,alive) ~ concentration, family=binomial)
summary(fit)
```

In the glm function, the command "family = binomial" tells R to fit a logistic regression model; glm can fit other models as well, as we will see. The above commands produce the following output:

```
Call:
glm(formula = cbind(deaths, alive) ~ concentration, family = binomial)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.2746  -0.4668   0.7688   0.9544   1.2990

Coefficients:
               Estimate Std. Error z value Pr(>|z|)
(Intercept)   -14.82300    1.28959  -11.49   <2e-16 ***
concentration   0.24942    0.02139   11.66   <2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

(Dispersion parameter for binomial family taken to be 1)

Null deviance:    284.2024   on 7  degrees of freedom
Residual deviance:   7.3849  on 6  degrees of freedom
AIC: 37.583

Number of Fisher Scoring iterations: 4
```

From the output we see that the maximum likelihood estimates for the intercept and slope are $\hat{\beta}_0 = -14.8230$ and $\hat{\beta}_1 = 0.2494$, which yields the following estimated logistic regression model:

$$\hat{p}(x) = \frac{e^{-14.8+0.25x}}{1 + e^{-14.8+0.25x}}.$$

Also, the slope estimate for insecticide concentration is highly significant which is no surprise. Plugging in a value for x, the CS_2 concentration, yields the estimated probability of a beetle dying at that dose after five hours of exposure. A plot of the sample proportions and the estimated logistic regression curve is plotted in Figure 1.

Inference for the Slope. The primary question of interest in a simple logistic regression (i.e. a logistic regression with only one predictor variable) is if the slope β_1 differs from zero. Recall that if $\beta_1 = 0$, then the predictor has no bearing on the probability of success or failure. In the beetle example, it is quite clear that as the CS_2 concentration increases, the probability of mortality increases as well. A simple method to test significance of β_1 is to use *Wald's* test. For large sample sizes, maximum likelihood estimators such as $\hat{\beta}_1$ from logistic regression typically follow normal distributions approximately. If we want to test the hypothesis

$$H_0 : \beta_1 = 0 \text{ versus } H_a : \beta_1 \neq 0,$$

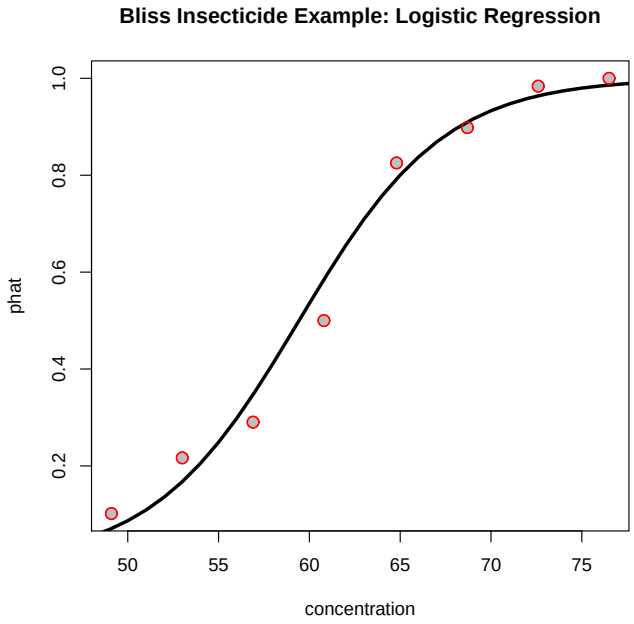

Figure 1: Proportion of beetle deaths versus Carbon disulphide concentration along with the estimated logistic regression curve.

then the statistic

$$z = \frac{\hat{\beta}_1}{\hat{se}(\hat{\beta}_1)},$$

will follow a standard normal distribution approximately when H_0 is true and the sample size is fairly large. If H_0 is false, then $|z|$ will tend to be large and we can compare z to critical values of the standard normal distribution. Note that if you square a standard normal random variable, you obtain a chi-squared random variable on one degree of freedom. Thus, alternatively, we could compare z^2 to critical values of the chi-squared distribution on one degree of freedom. If H_0 is false (i.e. $\beta_1 \neq 0$), the z^2 will tend to be large.

In the beetle example, $\hat{\beta}_1 = 0.2494$ and $\hat{se}(\hat{\beta}_1) = 0.0214$. This standard error is found using the theory of maximum likelihood (which involves taking second partial derivatives of the log-likelihood function). The z test statistic then is

$$z = \frac{\hat{\beta}_1}{\hat{se}(\hat{\beta}_1)} = \frac{0.2492}{0.0214} = 11.6449,$$

which is highly significant ($p < 2^{-16}$).

1.3 Interpreting the Logistic Regression Coefficients

The estimated slope in the beetle mortality example was $\hat{\beta}_1 = .2492$. What does this tell us about the relationship between dose of CS_2 and mortality? Recall that in a

simple *linear* regression, the slope β_1 measures the change in mean response y for a one unit change in x. The interpretation of the slope β_1 in a logistic regression however is not so straight forward. To help understand and interpret a logistic regression slope, we will first look at a very simple example where the predictor variable x is dichotomous, only two levels.

Prostate Cancer Example. Data on $n = 53$ prostate cancer patients (Collet, 1991) was collected. A laparectomy was performed on each patient to determine if the cancer had spread to surrounding lymph nodes. The goal is to determine if the size of the tumor can predict whether or not the cancer has spread to the lymph nodes. Define an indicator regressor variable x as

$$x = \begin{cases} 0, & \text{for small tumors} \\ 1, & \text{for large tumors} \end{cases}$$

and

$$y = \begin{cases} 0, & \text{lymph nodes not involved} \\ 1, & \text{lymph nodes involved} \end{cases}.$$

The maximum likelihood estimators from fitting a logistic regression of y on x are

$$\hat{\beta}_0 = -1.4351 \quad \text{and} \quad \hat{\beta}_1 = 1.6582.$$

The estimated probability that the cancer will spread to the lymph nodes is

$$\hat{p}(1) = \frac{e^{-1.435+1.658}}{1+e^{-1.435+1.658}} = .5555 \text{ for large tumor patients}$$

$$\hat{p}(0) = \frac{e^{-1.435}}{1+e^{-1.435}} = .1923 \text{ for small tumor patients}$$

ODDS: For large tumor patients (i.e., $x = 1$), the (estimated) *ODDS* that the cancer will spread to the lymph nodes is

$$\frac{\hat{p}(1)}{1-\hat{p}(1)} = .5555/(1-.5555) = 1.25.$$

The interpretation of the odds is: For *large tumor patients* the probability the cancer spreads to lymph nodes is about one and a quarter times higher than the probability the cancer will not spread to the lymph nodes. For small tumor patients (i.e. $x = 0$), The (estimated) odds that the cancer has spread to the lymph nodes is

$$\frac{\hat{p}(0)}{1-\hat{p}(0)} = .1923/(1-.1923) = .238.$$

Thus, for small tumor patients, the probability the cancer spreads is only about 1/4 ($\approx .238$) of the probability the cancer will not spread. Or, inverting things, we could say that the probability the cancer will not spread is about 4 times higher than the probability it will spread for small tumor patients.

The **ODDS RATIO** is defined as the ratio of the odds for $x = 1$ to the odds for $x = 0$:

$$\text{ODDS RATIO } = \frac{p(1)/(1-p(1))}{p(0)/(1-p(0))}.$$

One can do some algebra to show:

$$\text{ODDS RATIO } = e^{\beta_1}. \tag{3}$$

The logistic regression slope β_1 is related to the odds ratio by:

$$\textbf{LOG–ODDS } = \log(\text{odds ratio}) \ = \beta_1.$$

For the prostate cancer example we found that

$$\hat{\beta}_1 = 1.658.$$

Thus, the odds–ratio (of those with large tumor to those with small tumors) is estimated to be

$$e^{\hat{\beta}_1} = e^{1.658} = 5.25.$$

The interpretation of the odds–ratio in this example is that the odds the cancer spreads is about 5 times greater for those with large tumors compared to those with small tumors.

Caution: This does not mean that those with large tumors are five times more likely to have the cancer spread than those with small tumors (see *relative risk*).

To help understand the odds ratio, suppose there are 100 patients with small tumors. The odds that the cancer spreads for small tumor patients is approximately 1/4. Thus, we would expect that for every 20 patients who have had the cancer spread, there will be 80 patients for whom the cancer has not spread: $20/80 = 1/4$. Now the odds ratio is approximately equal to 5. That is, the odds the cancer has spread is about 5 times higher for large tumor patients than for small tumor patients:

$$\begin{aligned} \text{Odds for Large Tumor Patients } &= 5 \times (\text{Odds for Small Tumor patients}) \\ &= 5 \times \frac{20}{80} \\ &= \frac{100}{80} \end{aligned}$$

which can be interpreted as: out of 180 patients with large tumors, we would expect to see 100 patients who have had the cancer spread compared to only 80 patients who have not had the cancer spread. If we interpret this for a total of 100 patients with large tumors (instead of 180 patients) we would expect to see about 55 patients where the cancer has spread compared to about 45 for whom the cancer has not spread.

Independence. Note that if the odds ratio equals 1, then the odds of the cancer spreading is the same for large and small tumor patients. In other words, the odds of the cancer spreading does not depend on the size of the tumor. Recall

$$\text{ODDS RATIO } = e^{\beta_1}.$$

Thus, if the odds ratio equals 1, this implies that the logistic regression slope $\beta_1 = 0$ since $e^0 = 1$.

Relative Risk. The relative risk is defined as:

$$\text{Relative Risk} = \frac{p(1)}{p(0)}.$$

In the prostate example, the relative risk is estimated to be

$$\frac{\hat{p}(1)}{\hat{p}(0)} = \frac{0.5555}{0.1923} = 2.89$$

which is quite a bit less than the odds ratio of 5.25. The relative risk (2.89) means that the probability of the cancer spreading for the large tumor patients is almost 3 times higher compared to the small tumor patients.

Interpreting the slope when x is continuous. Returning to the beetle mortality example, recall that $\hat{\beta}_1 = 0.2494$. If we increase the dose by one unit, then the odds ratio for a unit change in x can be shown to be (after some algebra):

$$\frac{p(x+1)/(1-p(x+1))}{p(x)/(1-p(x))} = e^{\beta_1}.$$

The estimated odds ratio for an increase of one unit of CS_2 concentration is

$$e^{\hat{\beta}_1} = e^{0.2494} = 1.283,$$

indicating that the odds of dying is about 1.283 times greater for each additional unit increase in CS_2.

1.4 Multiple Logistic Regression

In the classical regression setting we have seen how to include more than one regressor variable in the regression model. Multiple regressors can also be incorporated into the logistic regression model as well. Suppose we have p regressor variables $x_1, x_2, \ldots, x_p$. Then we can generalize (2) and define a multiple logistic regression function:

$$p(x_1, \ldots, x_p) = \frac{e^{\beta_0+\beta_1 x_1+\cdots+\beta_p x_p}}{1 + e^{\beta_0+\beta_1 x_1+\cdots+\beta_p x_p}} \tag{4}$$

and the logit of $p(x)$ is

$$\text{logit}(p(x)) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p.$$

Maximum likelihood is generally used to estimate the β_j's and their standard errors for the multiple logistic regression model as was done for the *simple* logistic regression.

Tests for Significance for Multiple Logistic Regression. Just as in the case for multiple regression, we can also perform statistical tests to determine if subsets of the

the regression coefficients differ from zero. The testing procedures for both multiple regression and multiple logistic regression are based on the same principal: fit the full model and the reduced model and compare the two fits. If the reduced model does nearly as good a job as the full model, then the reduced model is preferred. The actual mechanics of the testing procedure in multiple logistic regression differ from that of multiple regression though which we now discuss.

Likelihood Ratio Test. The logistic regression model is a special case of a *generalized linear model*. For generalized linear models, a statistic called the **deviance** is computed which measures how close the predicted values from the fitted model match the actual values from the raw data. Maximum likelihood is generally used to estimate the parameters for generalized linear models. The *likelihood* is simply the probability density computed from the observed data values with the parameters replaced by their estimates. An extreme case is to fit a *saturated* model where the number of parameters equals the number of observations. One of the fundamental goals of statistics is to determine a simple model with as few parameters as possible. The saturated model has as many parameters as observations and hence it provides no simplification at all. However, we can compare any proposed model to the saturated model to determine how well the proposed model fits the data. The deviance D is defined as

$$D = 2[\text{log-likelihood(saturated model)} - \text{log-likelihood(proposed model)}].$$

If the proposed model is a good approximation to the truth, then the deviance should be relatively small and the sampling distribution of the deviance D follows a chi-squared distribution on $n - p - 1$ degrees of freedom approximately. This is an asymptotic result meaning that it is valid as the sample size n goes to infinity. However, this asymptotic result is usually not of much use. Instead, interest lies in comparing *nested* models – comparing reduced models to a full model. The principal is the same as it was for multiple regression. Consider the full model

$$\text{logit}(p(x_1, x_2, \ldots, x_p)) = \beta_0 + \beta_1 x_1 + \cdots + \beta_q x_q + \beta_{q+1} x_{q+1} + \cdots + \beta_p x_p,$$

and we want to test the null hypothesis

$$H_0 : \beta_{q+1} = \cdots = \beta_p = 0,$$

versus the alternative hypothesis that at least one of these coefficients differs from zero. If H_0 is true, then the regressor variables $x_{q+1}, \ldots, x_p$ are redundant in the full model and can be dropped. In order to test H_0 in practice, all one needs to do is fit the full model and the reduced model and compare their respective deviances. The test statistic is:

$$X^2 = D_{\text{reduced}} - D_{\text{full}}.$$

Note that the deviances in the above equation both involve the evaluation of the log-likelihood for the saturated model and when we take the differences of the deviances (reduced - full), the log-likelihood for the saturated model cancels out. Thus, the test statistic X^2 can be written as:

$$X^2 = 2[\text{log-likelihood(full model)} - \text{log-likelihood(reduced model)}]. \tag{5}$$

If H_0 is true, then the test statistic X^2 has an approximate chi-squared distribution (provided the sample size is sufficiently large) whose degrees of freedom is equal to the difference in the number of parameters between the full and reduced models: $p - q$. If H_0 is false, then the test statistic tends to be too large to be considered as deriving from the chi-squared distribution on $p - q$ degrees of freedom. That is, if X^2 is too big, reject H_0. If we are testing at a level of significance α, then we reject H_0 if $X^2 > \chi_{\alpha,p-q}$, the α critical value of the chi-squared distribution on $p - q$ degrees of freedom.

The test statistic given by (5) is based on the notion of a *likelihood ratio test*. This test compares the observed likelihood of the full and reduced models. Since the reduced model is a constrained version of the full model, the likelihood of the reduced model will be less than or equal to the likelihood for the full model. The two models are then compared in terms of the ratio (reduced/full) of their observed likelihoods. Taking the logarithm of the ratio turns it into a difference of the log-likelihoods. If the log-likelihood is multiplied by -2, its sampling distribution under the null hypothesis is chi-square asymptotically with degrees of freedom equal to the difference in the number of parameters between the full and reduced models. The test statistic (5) is sometimes referred to as the log-likelihood ratio statistics or LLRS.

Additionally, significance tests can be performed for individual regression coefficients (i.e. $H_0 : \beta_j = 0$) by computing *Wald* statistics which are similar to the partial t-statistics from classical regression:

$$w_j = \frac{\hat{\beta}_j}{\hat{\text{se}}(\hat{\beta}_j)}.$$

Under the null hypothesis that $\beta_j = 0$, the Wald test statistic w_j follows approximately a standard normal distribution (and its square is approximately a chi-square on one-degree of freedom). In order to illustrate these testing ideas, we now illustrate with an example.

Example. (*open inbreeding.r*) Inbreeding occurs naturally within plant populations. A study was conducted to study the effect of plant inbreeding on the resistance and tolerance of the plant to native herbivores in Napa County, California using the plant Yellow Monkeyflower. The response variable y of interest in this study was whether or not the plant produced flowers (0 for no and 1 for yes). Flower production is needed for reproduction. The primary explanatory variable of interest was the indicator variable indicating whether the plant was inbred (value of 1) or cross-bred (with a value of 0). Two other covariates were also recorded: Herbivore damage to the plants due to spittlebugs, adult and larval beetles, slugs, deer, etc. was recorded as a percentage on a discretized scale (12 categories); and the dried above ground biomass of the plant (in grams). The first few lines of the data (compliments of C. Ivey) are shown in the following table:

Damage	inbred	y	Biomass
0.2138	0	1	0.0525
0.2138	0	1	0.0920
0.2138	1	1	0.0239
0.2138	0	0	0.0149
0	0	0	0.0410
0.3907	0	0	0.0264
0.2138	0	1	0.1370
0.2138	1	1	0.0280
0.2138	1	1	0.0248
0.2138	1	1	0.0892
0.2138	1	0	0.0123
0	0	0	0.0105
⋮	⋮	⋮	⋮

The regressor variables biomass and damage are thought to explain much of the variability in the response and they are included as covariates. The regressor of primary interest is the indicator for whether the plant is inbred or not. The logit for the full model is

$$\text{logit}(p(\text{INBRED}, \text{DAMAGE}, \text{BIOMASS})) = \beta_0 + \beta_1 \text{INBRED} + \beta_2 \text{DAMAGE} + \beta_3 \text{BIOMASS}.$$

(Note that this model does not contain any interaction terms – none of them were significant.) Here is R code for fitting the multiple logistic regression:

```
dat = read.table("inbreeding.dat")
y = dat[,3]
damage=dat[,1]
inbred = dat[,2]
biomass = dat[,4]
fitfull = glm(y ~ inbred+biomass+damage, family=binomial)
summary(fitfull)
```

The R output from running this "full" model is given below:

```
 Call:
glm(formula = y ~ inbred + biomass + damage, family = binomial)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -1.1230     0.6795  -1.653   0.0984 .
inbred        0.3110     0.6821   0.456   0.6484
biomass      41.3349    15.9374   2.594   0.0095 **
damage       -1.8555     2.1420  -0.866   0.3864
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

```
(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 69.235  on 49  degrees of freedom
Residual deviance: 52.463  on 46  degrees of freedom
AIC: 60.463

Number of Fisher Scoring iterations: 6
```

The minus two times the log-likelihood for this full model fit is 52.463 which the output calls the Residual deviance. Note that the Wald test for significance of the coefficients for inbred and damage yield p-values of $p = 0.6484$ and $p = 0.3864$ indicating that both of these regressors appear to be redundant in the full model. To illustrate how to simultaneously test for redundancy of a set of predictors, we can fit a reduced logistic regression model without inbred and damage to test $H_0 : \beta_1 = \beta_3 = 0$:

```
fitreduced = glm(y ~ biomass, family=binomial)
summary(fitreduced)
x2= 2*(logLik(fitfull)-logLik(fitreduced)) # log-likelihood ratio test statistic
as.numeric(x2)
pval=1-pchisq(x2,2)
as.numeric(pval)
```

Selected output from running this reduced model and computing the likelihood ratio statistic is given by

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -1.2556     0.5129  -2.448   0.0144 *
biomass      37.9157    14.7390   2.572   0.0101 *

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 69.235  on 49  degrees of freedom
Residual deviance: 53.579  on 48  degrees of freedom
AIC: 57.579

> x2= 2*(logLik(fitfull)-logLik(fitreduced)) # log-likelihood ratio test statistic
> as.numeric(x2)
[1] 1.116378
> pval=1-pchisq(x2,2)
> as.numeric(pval)
[1] 0.5722445
```

The minus two times the log-likelihood for this reduced model is 53.579 and the value of the log-likelihood ratio statistic is $X^2 = 1.1164$. Since the full and reduced models differ by 2 parameters, we can compare this test statistic to a chi-squared distribution on 2 degrees of freedom. The p-value for this test is $p = 0.5723$. Thus we conclude

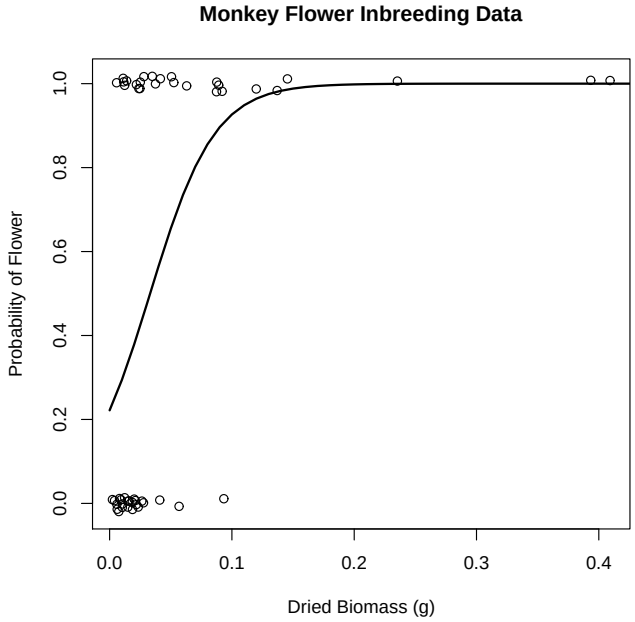

Figure 2: A logistic regression plot for the Yellow Monkey flower plant. The response is 1 if the plant produced at least one flower and 0 if it did not produce a flower. The response is plotted against plant biomass. Also shown is the estimated logistic regression curve. Note that the 0-1 responses were "jittered" so that they would show up better.

that there is insufficient evidence that the coefficients β_1 and β_3 differ from zero. This allows us to settle on a logistic regression model involving only the biomass as a predictor. The estimated probability of a plant producing a flower for a given biomass is

$$\frac{\exp\{-1.2556 + 37.9157\text{Biomass}\}}{1 + \exp\{-1.2556 + 37.9157\text{Biomass}\}}.$$

From this analysis, it appears that whether or not this particular plant species was inbred does not affect its ability to reproduce, even when accounting for the plant's size and the damage sustained from herbivores. It is useful to note that when a logistic regression is run using only the indicator for inbreeding (x_1) in the model, the regressor x_1 is still not significant ($p = 0.9442$). A plot of the response y versus biomass along with the fitted logistic regression curve is shown in Figure 2. Note that in order to better distinguish between the zero and one response values, the "jitter" command was used in the plot statement:

```
plot(biomass,jitter(y,.1), main="Monkey Flower Inbreeding Data",
      ylab="Probability of Flower", xlab="Died Biomass (g)")
```

Because the estimated slope is positive, we see that the probability of reproducing increases as the size of the plant increases.

Note that the estimated slope is $\hat{\beta}_1 = 37.9157$ and using this to compute the odds

ratio will yield an astronomically huge number. Since the dried biomass of the plants range in values from 0 to 0.4, it makes better since to compute the odds ratio, not for a unit increase in biomass, but for a much smaller increases, say an increase of 0.1 grams. In this case, the odds-ratio will be

$$e^{0.1\hat{\beta}_1} \approx 44.3$$

indicating that for each additional 0.1 increase in dried biomass, the odds of producing at least one flower increase by a *factor* of 44.3.

2 Generalized Linear Models: The General Setup

The classical linear model which encompasses linear regression models as well as the usual analysis of variance models for comparing treatment means can be expressed as

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p + \epsilon. \tag{6}$$

Factorial ANOVA models result when the regressor variables are indicator variables and we obtain regression models when the regressors are continuous or they are a mix of continuous/indicator variables. The response y in (6) is a linear function of the model coefficients, the β_j's and hence the name *linear model*. The error in (6) is often assumed to have a normal distribution. (6) can be generalized to handle situations where the error is non-normal. In the generalized setting, the response variable y is still related to the regressors through a linear combination $\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$, except that the relation may not be direct as in (6). Instead, the response is linked to the linear combination $\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$ by a *link* function.

In the logistic regression model, the link function was the logit. Recall that in the logistic regression setup, the response y is a zero-one Bernoulli random variable whose success probability $p(x_1, \ldots, x_p)$ depends on a set of regressors $\boldsymbol{x} = (x_1, \ldots, x_p)$ by means of

$$p(x_1, \ldots, x_p) = \frac{e^{\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p}}{1 + e^{\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p}}, \tag{7}$$

and the logit link function is defined as:

$$\text{logit}(p(x_1, \ldots, x_p)) = \log[p(x_1, \ldots, x_p)/(1 - p(x_1, \ldots, x_p))] = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p.$$

In the classical regression model (6), the link function is simply the identity function.

Logistic regression and (6) are two examples of *generalized linear models*. Other examples of generalized linear models are Poisson regression, log-linear models (for example count data in contingency tables), survival analysis models. The three main principles of generalized linear models are:

1. We have a sample $Y_1, \ldots, Y_n$ of independent response variables from an *exponential family*. Basically, the exponential family of probability distributions are those whose density function can be expressed as an exponential function and the *support* of the density (i.e., the points where the density is greater than

zero) does not depend on the parameters of the model (the normal and binomial distributions are both in the exponential family).

2. There exists a linear predictor $\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$ of the response variable Y.

3. The mean response depends on the linear predictor through a link function g. This mean response that depends on the values of x_j's is known as a conditional expectation: $\mu_{y|\boldsymbol{x}} = E[Y|x_1, \ldots, x_p]$ and

$$g(\mu_{y|\boldsymbol{x}}) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p.$$

   For (6), the function g is the identity $g(x) = x$ and for logistic regression, $p(x)$ is the conditional mean of the Bernoulli response Y given values of the regressor variables $\boldsymbol{x}$.

The statistical estimation and testing for generalized linear models is usually done using likelihood theory, as was the case for logistic regression. Once a distribution for the response is decided upon, the likelihood equations can be written out and the usual maximum likelihood estimation procedure is then followed. Additionally, model testing is typically carried out using likelihood ratio tests based on differences in deviances as we saw with the multiple logistic regression example. Using the theory of maximum likelihood estimation, the test statistics under the null hypotheses based on deviances follow chi-square distributions for large sample sizes.

2.1 Overdispersion

In the classic linear models (e.g. regression or ANOVA), the assumption is made that the variance of the error is constant across observations. This assumption is often violated in practice. In regression, one can fit a *weighted least-squares* to adjust for the unequal variances. In generalized linear models, we do not expect the responses to have equal variances. For example, in the beetle example at the beginning of the chapter, the binomial responses have a variance equal to $n_i p(1-p)$ where n_i is the number of beetles exposed to the pesticide. Different values of n_i yield different variances. Nonetheless, it is quite common in generalized linear models for the variability of the response to exceed what is expected by the model. The term for this is *overdispersion*.

Overdispersion can be caused by several factors. For instance, in logistic regression, lack of independence between observations. A positive covariance between Bernoulli (i.e. zero-one) responses could inflate the variance. If overdispersion is a problem, an overdispersion parameter can be estimated and the log-likelihood ratio test would need to be adjusted by dividing by this estimated overdispersion variance which would yield an approximate F-test.

We have only briefly introduced the topic of generalized linear models. Many of the issues in regression such as goodness of fit, model building etc. are also issues with generalized linear models. Details on these topics can be found in books that focus on

generalized linear models. The classic text on the subject is McCullagh and Nelder's 1989 book *Generalized Linear Models* (McCullagh and Nelder, 1989).

We close this chapter with another very useful example of a generalized linear model.

3 Poisson Regression

The Poisson probability distribution is perhaps the most commonly used discrete distribution for modeling count data. For example, a study was done on occurrences of endemic gastroenteritis as measured by hospitalizations, emergency room, physician visits, and long-term care visits. One of the interests lies in associating gastroenteritis with water quality measures. If we let Y denote the number of occurrences of this illness over a specific period of time, then a Poisson distribution may be a reasonable way of modeling the distribution of Y. Note that if Y equals the number of cases of gastroenteritis, then Y can assume values $0, 1, 2, \ldots$. The Poisson distribution is parameterized by a rate parameter $\mu > 0$ and the probability density function $f(y)$ is given by

$$f(y) = P(Y = k) = e^{-\mu}\frac{\mu^k}{k!}, \text{ for } k = 0, 1, \ldots. \tag{8}$$

The mean and variance of a Poisson random variable equals μ. The Poisson probability model can be derived theoretically in situations where the following three conditions hold:

1. The occurrences of the event of interest in non-overlapping "time" intervals are independent.
2. The probability two or more events in a small time interval is small, and
3. The probability that an event occurs in a short interval of time is proportional to the length of the time interval.

One can show that the Poisson distribution belongs to the exponential class of probability distributions and hence, a generalized linear model approach can be used to relate a Poisson response to predictor variables.

Suppose that a Poisson response y depends on a set of regressor variables $x_1, \ldots, x_p$. Because the mean must be positive, a natural way of modeling the conditional expectation of the response of y given the predictors is

$$\mu = E[y|x_1, \ldots, x_p] = e^{\beta_0+\beta_1 x_1+\cdots+\beta_p x_p}. \tag{9}$$

If we take the natural logarithm of each side of this equation we obtain

$$\log(\mu) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p,$$

which indicates that the natural logarithm is the link function for a Poisson regression. This is an example of a *log-linear* model.

If a data set consists of measurements on a response variable y that correspond to counts as well as measurements on a set of regressor variables, one may want to simply model this using the classical regression model. However, if a standard regression is used to model count data, one will often encounter problems with *heteroscedasticity* which means unequal error variances. Recall that one of the assumptions in the standard regression setup is that the variance of the error distribution is constant, regardless of the values of the regressors. However, as noted above, the variance of a Poisson random variable is μ. The Poisson distribution is unique in that its mean and variance are equal to each other. If the Poisson mean is related to regressors as in (9), then

$$\text{var}(Y) = e^{\beta_0+\beta_1 x_1+\cdots+\beta_p x_p},$$

clearly indicating that the variance depends on the values of the regressors as well and hence the equal variance assumption will not hold for Poisson count data.

Maximum likelihood estimation is typically used to estimate the parameters of a Poisson regression model. To illustrate, let y_i be a Poisson random variable that depends on a predictor x_i for a random sample $i = 1, \ldots, n$. Then the likelihood is

$$L(\beta_0, \beta_1) = \Pi_{i=1}^{n} e^{-\mu_i} \mu_i^{y_i} / y_i!$$

where μ_i is given by (9). Taking the logarithm of the likelihood gives the log-likelihood function:

$$l(\beta_0, \beta_1) = \sum_{i=1}^{n} [-\mu_i + y_i \log(\mu_i) - \log(y_i!)],$$

and the goal is to find the values of β_0 and β_1 that maximize this function. We can do this in R using "glm" and specifying "family=poisson". Alternatively, one can use PROC GENMOD in SAS.

The following example will be used to illustrate Poisson regression.

Example. *C. dubia Zooplankton* were exposed to different doses of Nickel (μgram/liter) at the Little Miami River. One of the doses was zero which acted as a control and the highest dose was 34 $\mu g/l$. The response variable is the number of offspring produced over the seven days. There were 10 females exposed at each dose. The data set *zooplanktonLMR.dat* has the following columns: column 1: dose, columns 2-9: an indicator for whether or not the female was alive or dead on each given day (0=dead, 1=alive) and column 10: Number of offspring (data source: Custer, K.W. and Burton, G.A.). Figure 3 shows a scatterplot of the raw data of number of offspring versus dose of nickel. We could ignore the fact that the response is a count and just fit an ordinary least-squares regression. Clearly from Figure 3, a straight line model will not provide a suitable approximation for the relation between count and dose. If we fit a quadratic linear model, we get the residual plot shown in Figure 4. Fitting this quadratic model is clearly not a good strategy, but it helps to illustrate a couple points. From the residual plot in Figure 4 we see structure in the plot indicating that the quadratic model is not a good fit. Since the number of offspring appear to be declining exponentially with nickel dose, it is not surprising that the quadratic model performs poorly. Additionally however, the residual plot shows a very clear unequal variance in the residuals which is to be expected for Poisson type data.

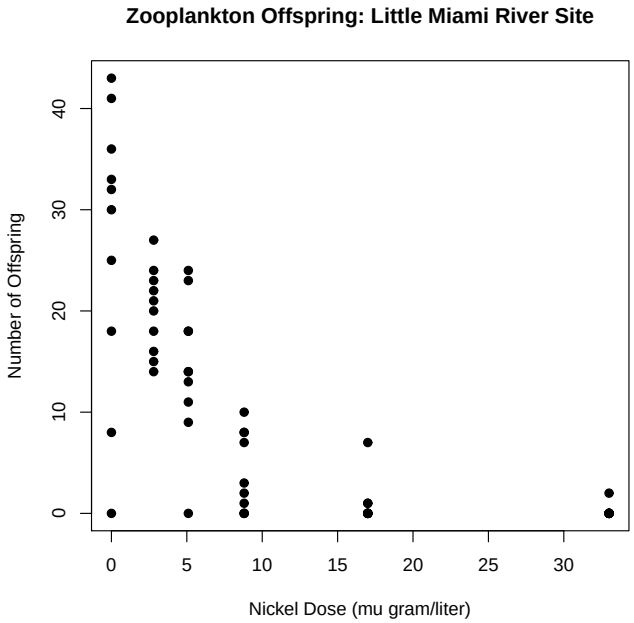

Figure 3: Scatterplot of the number of zooplankton offspring after 7 days versus the level of nickel to which the females were exposed.

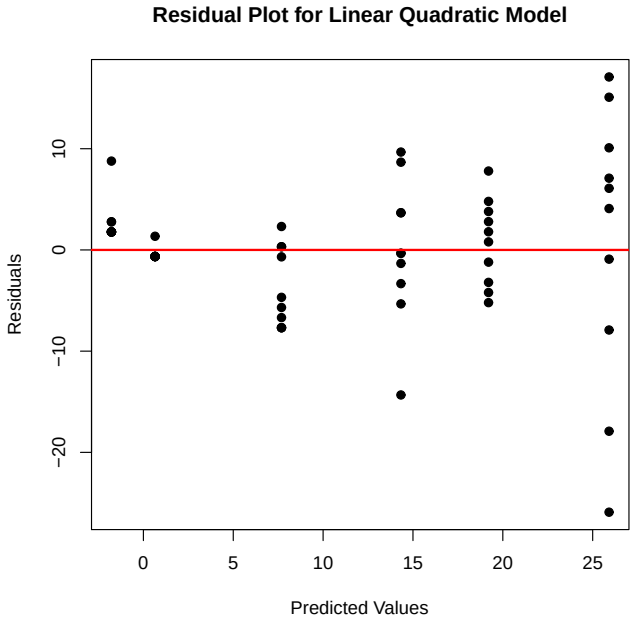

Figure 4: Residual plot from fitting a least-squares quadratic model

We can fit a simple Poisson regression by regressing the number of offspring on level of nickel using the following R code:

```
# lmr stands for Little Miami River
lmr <- read.table("zooplanktonLMR.dat", header=F)
y <- lmr[,10]
dose <- lmr[,1]
fit <- glm(y ~ dose, family=poisson)
summary(fit)
plot(dose, y, xlab="Nickel Dose (mu gram/liter)", pch=19,
     ylab="Number of Offspring",
     main="Zooplankton Offspring: Little Miami River Site")
nickel=seq(0,40, by=1)
yhat1=exp(coef(fit)[1]+coef(fit)[2]*nickel)
lines(nickel, yhat1, lwd=2, col=2)
```

The output from this model is given by the summary command:

```
Call:
glm(formula = y ~ dose, family = poisson)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  3.38150    0.04959   68.19   <2e-16 ***
dose        -0.18225    0.01092  -16.69   <2e-16 ***

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 872.94  on 59  degrees of freedom
Residual deviance: 251.32  on 58  degrees of freedom
AIC: 424.80

Number of Fisher Scoring iterations: 5
```

From this output, we have the following estimated model:

$$\hat{y} = e^{3.38-0.182x},$$

where x is the dose of nickel. The coefficient for dose (x) is highly significant as indicated by the near-zero p-value. A plot of the raw data and the fitted Poisson regression curve is shown in Figure 5.

Unlike the linear model, in order to interpret the slope coefficient in a Poisson regression, it makes better sense to look at the ratio of predicted responses (instead of the difference) for a unit increase in x:

$$\frac{e^{\beta_0+\beta_1(x+1)}}{e^{\beta_0+\beta_1 x}} = e^{\beta_1}.$$

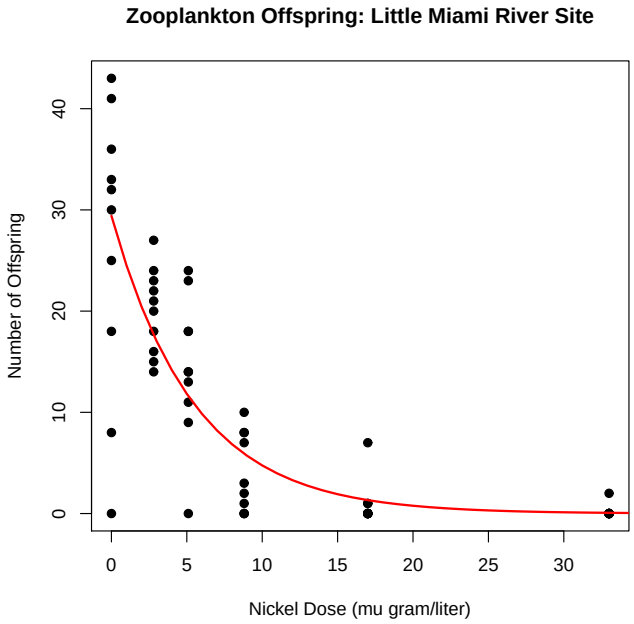

Figure 5: Zooplankton data with an estimated Poisson regression curve.

For the zooplankton example with $\hat{\beta}_1 = -0.182$, we have

$$e^{\hat{\beta}_1} = e^{-0.182} = 0.8334.$$

Thus, for a unit increase in the dose of nickel, we would expect to see the number of offspring to decrease by a *factor* of 0.8334.

As in the classic regression and logistic regression models, the Poisson regression model easily generalizes to multiple Poisson regression with more than one predictor.

3.1 Overdispersion

Recall that the variance of a Poisson random variable is equal to the mean of the Poisson random variable. A common problem with Poisson regression is that the response is more variable than what is expected by the model. This is called *overdisperson*. One way to diagnose overdispersion is to look at the deviance statistic which is -2 times the difference in the log-likelihood of the model under consideration and the saturated model. In the saturated model, the predicted response is just the actual observed response y_i. Letting $\hat{y}_i$ denote the predicted response from the Poisson model, one can show that the deviance is

$$D = -2[\text{log-likelihood(model)} - \text{log-likelihood(saturated)}] = 2\sum[y_i \log(y_i/\hat{y}_i) - (y_i - \hat{y}_i)]. \quad (10)$$

(Note that we take $y_i \log(y_i) = 0$ if $y_i = 0$.) If the model is a good fit to the data, then it follows that the deviance should be roughly equal to the deviance degrees of freedom which is the sample size n minus the number of estimated coefficients:

$df = n - p - 1$ where p is the number of predictors. Asymptotically, the deviance follows a chi-square distribution on these degrees of freedom if the model is correctly specified. In R, the "glm" function calls this the "Residual deviance". If the residual deviance greatly exceeds the residual degrees of freedom, then that is an indication of an overdispersion problem.

From the output on the simple Poisson regression, we see that the residual deviance is 251.32 on 58 degrees of freedom. Since 251.32 is much greater than 58, this indicates that there may be an overdispersion problem with this simple model. When overdispersion is present, the estimated standard errors of the regression coefficients are not reliable and inferences based on them (e.g. Wald tests) are unreliable.

One fix for the inference problem is to estimate an overdispersion parameter based on comparing actual counts to predicted counts:

$$\phi = \frac{1}{(n-p-1)} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 / \hat{y}_i. \tag{11}$$

In R, we can compute ϕ and adjust the model by typing the following:

```
> phi=sum((y-fit$fitted)^2/fit$fitted)/fit$df.residual
> summary(fit, dispersion=phi)
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  3.38150    0.10013  33.771   <2e-16 ***
dose        -0.18225    0.02205  -8.264   <2e-16 ***
```

Note that the parameter estimates are identical to the original fit, but the standard errors, after adjusting for overdispersion, are larger (e.g. the estimated standard error for the dose coefficient is about double its original value).

The problem of overdispersion can be caused by the specification of the model. For instance, important predictor variables may have been left out of the model. Another source of overdispersion is correlations amongst the observations, perhaps due to clustering (e.g. observations obtained from the same plant or plot). Random effects can be introduced in these cases to account for correlations.

Another reason overdispersion can occur is that a different count distribution is needed to model the data. For instance, the *negative binomial* distribution can be used to model count data, see Section 3.2.

Logarithm Transformation. As with any other regression model, sometimes transformations of the predictor variables can lead to more appropriate models. For example, suppose a Poisson count y is modeled via a regression on a predictor x. Instead of considering

$$E[y|x] = e^{\beta_0 + \beta_1 x},$$

one could instead use a logarithm transformation of the predictor and consider:

$$E[y|x] = e^{\beta_0 + \beta_1 \log(x)},$$

but this model reduces to

$$E[y|x] = \beta_0^* x_1^{\beta}, \tag{12}$$

where $\beta_0^* = \log(\beta_0)$. The model (12) stipulates that the mean response is a "polynomial" function of x instead of an exponential function of the predictor.

3.2 Negative Binomial Model

The Poisson model does not always provide a good fit to a count response. An alternative model is the negative binomial distribution. Consider the usual set up for a binomial experiment where the probability of success is p and we observe independent trials that are either successes or failures. Suppose we observe a process of trials until we have r successes. Then the total number of trials W required to observe r successes has the following probability distribution:

$$P(W = n) = \binom{n-1}{r-1} p^r (1-p)^{n-r}.$$

To fit a regression using the negative binomial, instead of the Poisson, one can use the "glm.nb" function in R. This will generate estimated coefficients that are related to the mean response μ by

$$e^{\beta_0 + \beta_1 x} = \frac{\mu}{\mu + r},$$

where r is the negative binomial distribution function parameter (and is called theta in the output from glm.nb in R).

Looking back at the zooplankton example, recall that as the dose of nickel got larger, more female zooplankton died off which would naturally lead to less offspring. Let Y denote the number of offspring and for a given value of the rate parameter μ, Y has a Poisson distribution. If the rate parameter is also a random variable related to the likelihood of the female surviving that has a gamma distribution with density $g(\mu)$, then the resulting distribution for Y is an *infinite mixture* of a Poisson and a gamma distribution given by

$$\begin{aligned} P(Y = y) &= \int_0^\infty f(y|\mu) g(\mu) d\mu \\ &= \int_0^\infty \frac{e^{-\mu}\mu^y}{y!} g(\mu) d\mu. \end{aligned}$$

Solving this integral yields a negative binomial distribution. From this point of view, the negative binomial provides an elegant generalization of the Poisson regression model.

3.3 Rate Models

In many count regression problems, the actual count may be related to the size of the unit from which observations are taken. For instance, the number of egg masses found

on a plant will be related to the number of leaves on the plant. In the zooplankton example, it stands to reason that the number of offspring over the seven day period will depend on the number of days the female survives. For instance, if days equals the number of days the female survives, we could write

$$\log(y/\text{days}) = \beta_0 + \beta_1 x.$$

Exponentiating, we can consider a model

$$E[y|x] = e^{\text{days}+\beta_0+\beta_1 x}.$$

In this case, the variable days would have a fixed coefficient of one. In Poisson regression, this is called the *offset*.

For the example presented here with the zooplankton, note that in the original data set, for each female, there was an indicator variable (0 or 1) for whether or not the female was alive at each day (days 0–7). We could model this data by using ideas from the branch of statistics known as *survival analysis*. Let $f(y|x)$ denote the probability density function for the number of offspring of a female at a given dose x. Let d denote the number of days the female survives. Then we can write (with a slight abuse of notation):

$$\begin{aligned} f(y|x) &= \frac{f(y,x)}{f(x)} \\ &= \sum_{d=0}^{7} \frac{f(y,d,x)}{f(x)} \\ &= \sum_{d=0}^{7} f(y|x,d)\frac{f(x,d)}{f(x)} \\ &= \sum_{d=0}^{7} f(y|x,d)f(d|x). \end{aligned}$$

A Poisson regression model could be used to estimate $f(y|x,d)$ and a survival analysis model could be used to model the probability a female survives for d days at a dose x, given by $f(d|x)$ in the above equation. Note that in this data set, many observations are *right censored* meaning that the experiment ended before many of the females died and therefore we do not know the value d for many females – this is a common occurrence in survival analysis. We will not go into further details on this subject here but the interested reader can refer to one of several reference books on survival analysis (e.g. Hosmer and Lemeshow, 2008).

3.4 Zero-Inflated Poisson Regression

Another very common occurrence when working with count data is that there will be an overabundance of zero counts which is not consistent with the Poisson model. For example, in one survey, trees in a forest were observed and the number of vines on the trees were counted. The number of vines varied from tree to tree from zero

vines up to 30 vines. However, there were many trees with no vines which led to an *zero-inflated* Poisson model.

One way to model a zero-inflated Poisson model is via a two-component finite mixture that stipulates that the population consists of two sub-populations: one subpopulation is degenerate with counts of zero and the other subpopulation follows a Poisson distribution (e.g. Lambert, 1992; Min and Agresti, 2001) with density

$$f(y; p^*, \lambda^*) = \begin{cases} (1 - p^*) + p^* e^{-\lambda^*} & \text{if } y = 0 \\ p^* e^{-\lambda^*} (\lambda^*)^y / y! & \text{if y = 1, 2, } \ldots . \end{cases}, \tag{13}$$

where p^* is the proportion of the population in the Poisson sub-population.

One can fit a zero-inflated Poisson regression model using the R package *pscl* which has the function *zeroinfl*.

References

Bliss, C. I. (1935). The calculation of the dosage–mortality curve. *Annals of Applied Biology* **22**:134–167.

Collet, D. R. (1991). *Modeling Binary Data*. Chapman & Hall, London.

Hosmer, D. W. and Lemeshow, S. (2008). *Applied Survival Analysis, 2nd Edition*. Wiley, New York.

Lambert, D. (1992). Zero-inflated poisson regression, with an application to defects in manufacturing. *Technometrics* **34**:1–14.

McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models*. Chapman & Hall, London.

Min, Y. and Agresti, A. (2001). Random effect models for repeated measures of zero-inflated count data. *Statistical Modelling* **5**:1–19.