SS9055B: Generalized Linear Models

Section 8: Poisson Log-Linear Models II

Objectives

The objective of this lecture is to explore methods to account for overdisperion in GLM. In particular, we will consider overdispersion in the Poisson model. By the end of the lecture you should be able to:

- explain the mean-variance relationship in GLM,
- model overdispersed count data using negative binomial regression,
- explain the quasi-likelihood method,
- fit and interpret quasi-Poisson and quasi-binomial models, and
- critique the use of quasi-likelihood.

Introduction

Up to this point in the class we have only considered models in which the dispersion, parameter, ϕ, is known. In both the logistic regression model and the Poisson regression model $\phi = 1$ and in the gamma model we considered for the brain size data I set $\phi = .25$. In this section of the course we will consider models for count data in which the dispersion parameter is unknown and must be estimated from the data.

Mean-Variance Relationship

One important feature of generalized linear models that we have not discussed so far is that every GLM produces a fixed relationship between the mean and the variance. In the logistic regression model, $E(Y_i) = p_i$ and $\text{Var}(Y_i) = p_i(1-p_i)/n$ so that $\text{Var}(Y_i) = E(Y_i)(1-E(Y_i))/n$. In the Poisson log-linear model, $\text{Var}(Y_i) = E(Y_i)$. This is not an accident, and it should not be too surprising based on the theory we developed. Recall that both the mean and variance of Y_i can be computed from the cumulant generating function:

$$E(Y_i) = b'(\theta_i) \text{ and } \text{Var}(Y_i) = \frac{\phi b''(\theta_i)}{\omega_i}.$$

Assuming that $b'(\theta_i)$ is invertible we can solve the first expression to write θ_i as a function of μ_i, say $\theta_i = h(\mu_i)$, and then substitute this into the second equation to obtain a new function

$$v(\mu_i) = \text{Var}(Y_i) = \frac{\phi b''(h(\mu_i))}{\omega_i}.$$

This function is called the variance function and relates the mean and the variance for each member of the dispersion exponential family. Note that the variance for a specific observation may still depend on the weights as well, but these values are treated as known so they do not need to be estimated from the data.

While the variance function is fixed for any specific response distribution, the dispersion parameter ϕ allow us to have some control over this relationship. For example, $v(\mu_i) = \mu_i$ for the the Poisson log-linear model and so the variance must be proportional to the mean. However, the dispersion parameter allows some flexibility by letting the ratio between the variance and the mean to be something other than 1. Most commonly, the variance is greater than the mean. This is termed overdispersion and is the situation that we will focus on.

Example

As an example we will consider data from the study of horseshoe crabs originally published in Brockmann (1996). Horseshoe crabs are not really crabs, but are the only existing species in a very old and primitive group of organisms. They live in the Atlantic Ocean and are quite ugly[1]. Each spring thousands of horseshoe crabs come ashore along the US East coast to mate. Females, which are bigger, are swarmed by the males as they crawl up the beaches to dig nests in the sand and lay their eggs. The interest in this study is to know if some females are more attractive to males. Every female that comes onto the beach has at least one male attached, but some have extra males (called satellites). We will model the relationship between the weight and of a female and the number of satellites she has attached. Weight is measured in kilograms and colour is assigned to categories labelled 3, 4, or 5 (I don't know why there is no category 1 or 2). The data is in the file `crabs.csv` and full code for the analysis is provided in the accompanying `R` file.

Figure 1 shows the original data. The output from fitting the log-linear Poisson model with weight as the predictor and the number of satellites as the response is provided in Table 1. The results strongly suggest that the mean number of satellites increases with the weight of the females ($p < .001$). Specifically, the model indicates that the log of the mean increases by .59 (95%CI=.46,.71) for every 1 kg increase in the female's weight. More intuitively, this implies that the mean number of satellites increases $e^{.46} = 1.80$ (95%CI=1.58, 2.03) times for every 1 kg increase in the female's weight.

Assessing the fit of this model presents the same problem that arose when testing the goodness-of-fit of the binomial model with a continuous predictor. In this case, the counts per individual

[1] That is a subjective statement, but my prior probability is very close to 1.

Table 1: Output from initial Poisson log-linear model of the number of satellites as a function of a female's weight.

```
Call:
glm(formula = satellites ~ weight, family = poisson(), data = crabs)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.42841    0.17893  -2.394   0.0167 *
weight       0.58930    0.06502   9.064   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 632.79  on 172  degrees of freedom
Residual deviance: 560.87  on 171  degrees of freedom
AIC: 920.16

Number of Fisher Scoring iterations: 5

                2.5 %      97.5 %
(Intercept) -0.7771762 -0.07591025
weight       0.4597002  0.71449835
```

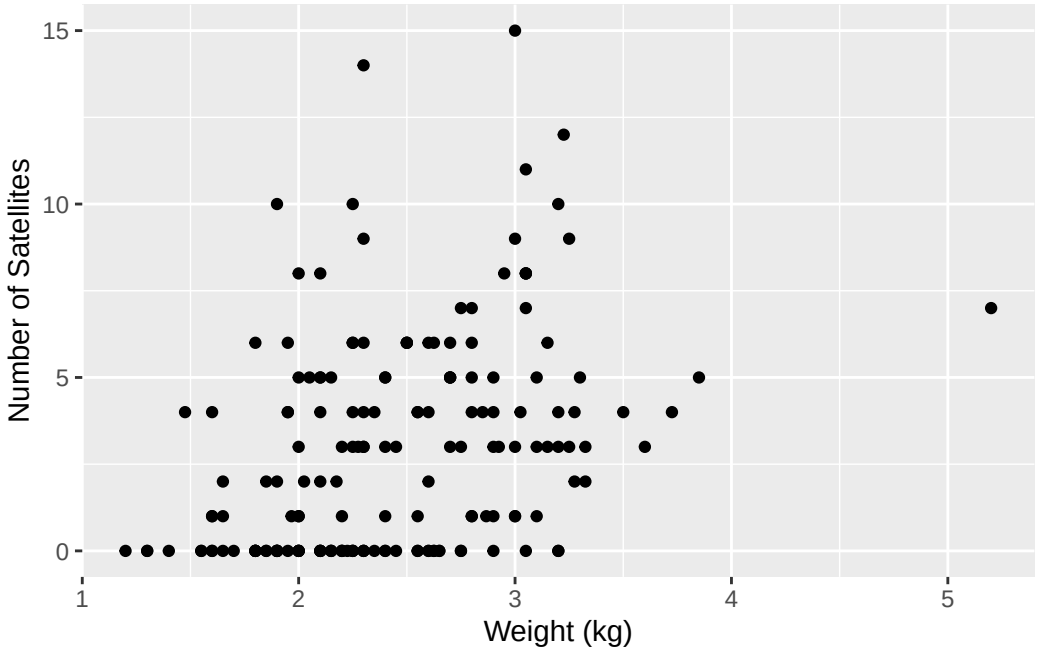

Figure 1: Number of satellites per female vs the female's weight.

are often very small (more than 72% of the observations are less than 5). This means that the residual deviance will not be well approximated by a chi-square distribution and the deviance goodness-of-fit test may be misleading. To solve this problem, we can break the weight into discrete categories and assess the fit of the surrogate model using group weight as the predictor instead. Table 2 shows the results from fitting the model with weight broken into 10 categories according to the quantiles of its distribution. This yields approximately 17 individuals in each category. Unfortunately, the results of the deviance goodness-of-fit test suggest that the model does not fit the data well. The residual deviance is 557.178 on 171 DF which yields $p < .001$. Moreover, the plot of the standardized residuals in Figure 2 shows that some of the errors are very large: 53 (30.6%) of the standardized residuals are bigger than 2 in absolute value and 3 observations have residuals bigger than 4. This is very unlikely to happen by chance, and suggests that there are problems with the model.

One solution would be to add covariates to try an improve the fit. The other covariates in the data are the width of the females and the categorical colour variable. Not surprisingly, the width and weight of a female are highly correlated, so the fit is not improved by adding this predictor into the model. Table 3 provides output from modelling the mean number of satellites as a function of both grouped weight, color, and their interaction. Comparing the two models with the likelihood ratio test shown in Table 4 provides strong evidence in favour of the more complicated model ($p = .003$), but the deviance goodness-of-fit test still suggests that the model is not sufficient ($\chi^2 = 537.51$ on 165 DF, $p < .001$).

Table 2: Output from initial Poisson log-linear model of the number of satellites as a function of a female's weight.

```
Call:
glm(formula = satellites ~ grp_weight, family = poisson(), data = crabs2)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.68929    0.21593  -3.192  0.00141 **
grp_weight   0.69195    0.07966   8.687  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 632.79  on 172  degrees of freedom
Residual deviance: 557.17  on 171  degrees of freedom
AIC: 916.46

Number of Fisher Scoring iterations: 6

                2.5 %     97.5 %
(Intercept) -1.116995 -0.2703068
grp_weight   0.536034  0.8483875
```

Table 3: Output from initial Poisson log-linear model of the number of satellites as a function of a female's weight (grouped) and colour.

```
Call:
glm(formula = satellites ~ grp_weight * color, family = poisson(),
    data = crabs2)

Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)         2.1276     0.8843   2.406 0.016134 *
grp_weight         -0.2736     0.3346  -0.818 0.413463
color3             -2.5291     0.9303  -2.719 0.006555 **
color4             -3.4499     0.9937  -3.472 0.000517 ***
color5             -3.6477     1.1243  -3.244 0.001177 **
grp_weight:color3   0.8829     0.3506   2.518 0.011788 *
grp_weight:color4   1.1387     0.3751   3.036 0.002401 **
grp_weight:color5   1.2439     0.4358   2.854 0.004317 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 632.79  on 172  degrees of freedom
Residual deviance: 537.51  on 165  degrees of freedom
AIC: 908.81

Number of Fisher Scoring iterations: 6
```

Table 4: Comparison of Poisson log-linear models.

```
Analysis of Deviance Table

Model 1: satellites ~ grp_weight
Model 2: satellites ~ grp_weight * color
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1       171     557.17
2       165     537.51  6   19.652 0.003192 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

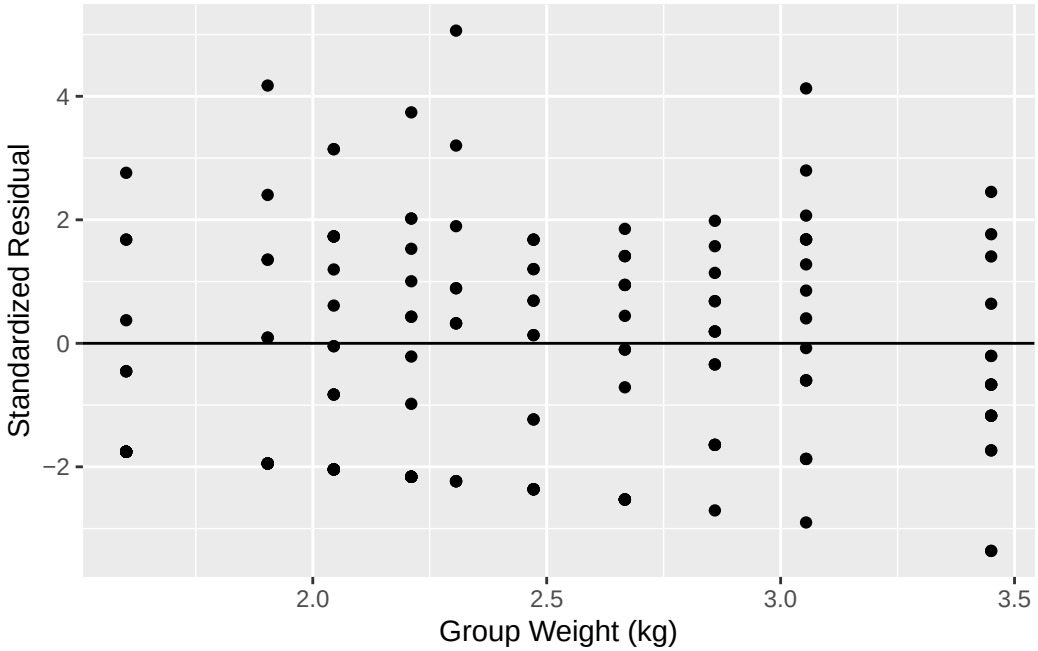

Figure 2: Residuals from the Poisson log-linear model versus grouped weight.

Overdispersion

The problem with these models is that the variance of the counts is simply much bigger than would be expected under the Poisson assumption. As noted introduction, the mean and variance of the Poisson distribution are equal. This means that if the model with the grouped weights and color as predictors fits the data then the mean and variance of the observations within each weight-by-colour class shoudl be about the same. Figure Figure 3 compares the mean and variance for of the counts for the groups we created earlier, and you can see that the variance is larger in every group. Fitting a linear regression model to the mean and variance (forced to go through the origin) suggests that the variance is roughly 2.9 times as large as the mean, on average (see Table 5. Again, this is a very strong indication that the data are overdispersed. Without any more covariates, there are two possible solutions to this problem. The first is to model the excess variability directly. The second is to fake the analysis (more technically called quasi-likelihood).

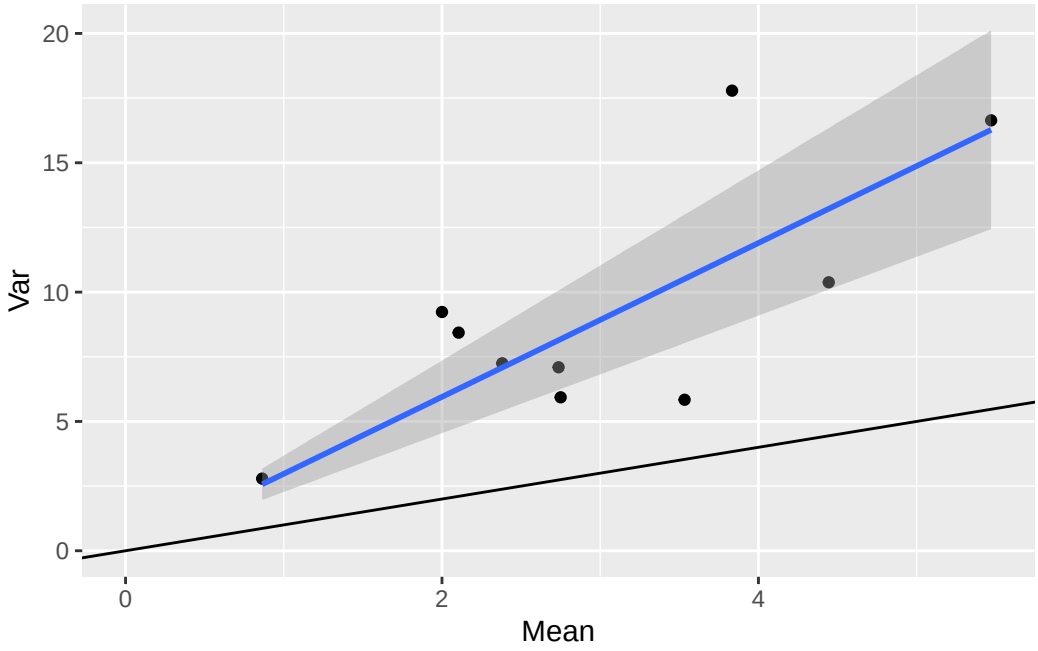

Figure 3: Variance as a function of the mean for the horseshoe crab data.

Table 5: Output from modelling the variance as a function of the mean for the horseshoe crabs data.

```
Call:
lm(formula = Var ~ Mean - 1, data = summarize(group_by(crabs2,
    Group), Mean = mean(satellites), Var = var(satellites)))

Residuals:
    Min      1Q  Median      3Q     Max
-4.6728 -1.9475  0.1929  1.7190  6.3846

Coefficients:
     Estimate Std. Error t value Pr(>|t|)
Mean   2.9748     0.3103   9.587 5.08e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.206 on 9 degrees of freedom
Multiple R-squared:  0.9108,    Adjusted R-squared:  0.9009
F-statistic: 91.91 on 1 and 9 DF,  p-value: 5.077e-06
```

Modeling Dispersion Explicitly

Negative Binomial Model

To model the overdispersion explicitlym we need to introduce further parameters that allow more flexibility in the relationship between the mean and the variance. The most common example of this for count data is the negative binomial model. Suppose that

$$Y_i|\mu_i \sim \text{Poisson}(\mu_i), \quad i = 1, \ldots, N$$

as in the usual Poisson log-linear model. However, we will now introduce extra variability into μ_i. Instead of assuming that μ_i is completely systematic and determined directly by the linear predictor we will assume that μ_i follows a gamma distribution whose mean is determined by the linear predictor. This allows the values μ_i to vary between individuals even though they have the same values of the predictors. Explicitly, we will model

$$\mu_i|\gamma, x_i \sim \text{Gamma}(1/\gamma, \gamma\mu(x_i))$$

where

$$\log(\mu(x_i)) = \eta_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}$$

as in the usual log-linear model. It is straightforward to show that the marginal distribution of $Y|\gamma, x_i$ is negative binomial:

$$Y|\gamma, x_i \sim \text{Neg. Bin.}\left(1/\gamma, \frac{\gamma\mu(x_i)}{1+\gamma\mu(x_i)}\right)$$

It follows that:

$$E(Y_i) = \mu(x_i)$$

but:

$$\text{Var}(Y_i) = \mu(x_i) + \gamma\mu(x_i)^2.$$

If $\gamma = 0$ then $\text{Var}(Y_i) = \mu(x_i)$ for all individuals sharing the same values of x_i, and we get the Poisson model back. However, if $\gamma > 0$ then the variance is greater than the mean, which allows for the overdispersion.

It is important to note that the negative-binomial distribution does not fit into the framework for GLMs because the distribution is not in the exponential dispersion family if both γ and $\boldsymbol{\beta}$ are unknown. It is part of the exponential family when γ is known, but that is generally not the case (This is akin to knowing σ^2 for a linear regression). However, most of the methods we have studied can still be applied if we can estimate γ. In particular, we can still apply Wald and LRT based inference to compute hypothesis tests and construct confidence intervals and

the LRT is still determined by the deviance. Negative binomial models can be fit in `R` using the function `glm.nb()` which is part of the `MASS` package.

Table 6 and Table 7 provide output from fitting negative binomial models to the crabs data with just weight or both weight and colour as predictors. Notice that the output contains the same elements that we had previously, except for the addition of the value `Theta` and its standard error. This quantity corresponds to the value $1/\gamma$ in the equations above. Figure Figure 4 shows the relationship between the mean and variance estimated from the more complicated model ($\text{Var}(Y_i) = \mu(x_i) + \mu(x_i)^2/.9844$) overlayed on the empirical estimates of the mean and variance. You can see that the curve fits the points fairly well.

Once again, we need to assess the fit of the model using the grouped weight in place of the continuous predictor. Results for the model including the grouped weight and color are shown in Table 8. When we compute the goodness-of-fit test now we find that the deviance goodness-of-fit statistic is 197.65 on 165 degrees of freedom which translates to a p-value of .042. This still provides evidence that the model does not fit the data well, suggesting that we need to find other covariates to improve the fit. However, the fit of the negative binomial model is still much better than the fit of the Poisson model.

Comparing the summaries of the models to the Poisson log-linear models we fit before you will see that the estimates are similar, but the standard errors are bigger and the confidence intervals are wider. This is a result of the extra uncertainty introduced by allowing the variance to be larger. Another consequence is that the tests are also less powerful, so the p-values are not as small. In particular, the LRT comparing the two models, shown in Table 9, provides no evidence at all to support the more complicated model ($p = .56$).

In the end, my conclusion is that the data only provide evidence that the weight of the female is an important predictor of the number of satellites. However, the data is much more variable than expected under the Poisson assumption suggesting that there are other important predictors of a female horseshoe crab's attractiveness that have not been included.

Beta-Binomial

Similar models can also be constructed for binomial data if we allow $n_i Y_i \sim \text{Binomial}(\pi_i)$ but assume that $\pi_i|x_i$ is random rather than deterministic. The most common choice is to assume that $\pi_i|x_i$ follows a beta distribution so that:

$$\begin{aligned} n_i Y_i|\pi_i &\sim \text{Binomial}(n_i, \pi_i) \\ \pi_i|x_i &\sim \text{Beta}(\alpha_i, \beta_i) \end{aligned}$$

Table 6: Summary of negative binomial model fit to number of satellites as a function of the female's weight.

```
Call:
glm.nb(formula = satellites ~ weight, data = crabs, init.theta = 0.9310592338,
    link = log)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -0.8647     0.4048  -2.136   0.0327 *
weight        0.7603     0.1578   4.817 1.45e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for Negative Binomial(0.9311) family taken to be 1)

    Null deviance: 216.43  on 172  degrees of freedom
Residual deviance: 196.16  on 171  degrees of freedom
AIC: 754.64

Number of Fisher Scoring iterations: 1

              Theta:  0.931
          Std. Err.:  0.168

 2 x log-likelihood:  -748.644
```

Table 7: Summary of negative binomial model fit to number of satellites as a function of the female's weight and color the horseshoe crab data.

```
Call:
glm.nb(formula = satellites ~ weight * color, data = crabs, init.theta = 0.984428301,
    link = log)

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)     1.9104     2.3179   0.824    0.410
weight         -0.1925     0.8746  -0.220    0.826
color3         -2.3750     2.3746  -1.000    0.317
color4         -3.6724     2.4720  -1.486    0.137
color5         -3.0384     2.6833  -1.132    0.257
weight:color3   0.8213     0.8959   0.917    0.359
weight:color4   1.2438     0.9419   1.320    0.187
weight:color5   1.0036     1.0580   0.949    0.343

(Dispersion parameter for Negative Binomial(0.9844) family taken to be 1)

    Null deviance: 223.05  on 172  degrees of freedom
Residual deviance: 197.02  on 165  degrees of freedom
AIC: 761.79

Number of Fisher Scoring iterations: 1

              Theta:  0.984
          Std. Err.:  0.181

 2 x log-likelihood:  -743.792
```

Table 8: Summary of negative binomial model fit to number of satellites as a function of the female's weight (grouped) and color for the horseshoe crab data.

```
Call:
glm.nb(formula = satellites ~ grp_weight * color, data = crabs2,
    init.theta = 0.9888543545, link = log)

Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)         2.1095     1.9934   1.058   0.2899
grp_weight         -0.2667     0.7424  -0.359   0.7194
color3             -2.7482     2.0746  -1.325   0.1853
color4             -3.5950     2.1518  -1.671   0.0948 .
color5             -3.1175     2.3915  -1.304   0.1924
grp_weight:color3   0.9667     0.7739   1.249   0.2116
grp_weight:color4   1.1984     0.8117   1.477   0.1398
grp_weight:color5   1.0154     0.9404   1.080   0.2803
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for Negative Binomial(0.9889) family taken to be 1)

    Null deviance: 223.59  on 172  degrees of freedom
Residual deviance: 197.65  on 165  degrees of freedom
AIC: 761.97

Number of Fisher Scoring iterations: 1

              Theta:  0.989
          Std. Err.:  0.183

 2 x log-likelihood:  -743.972
```

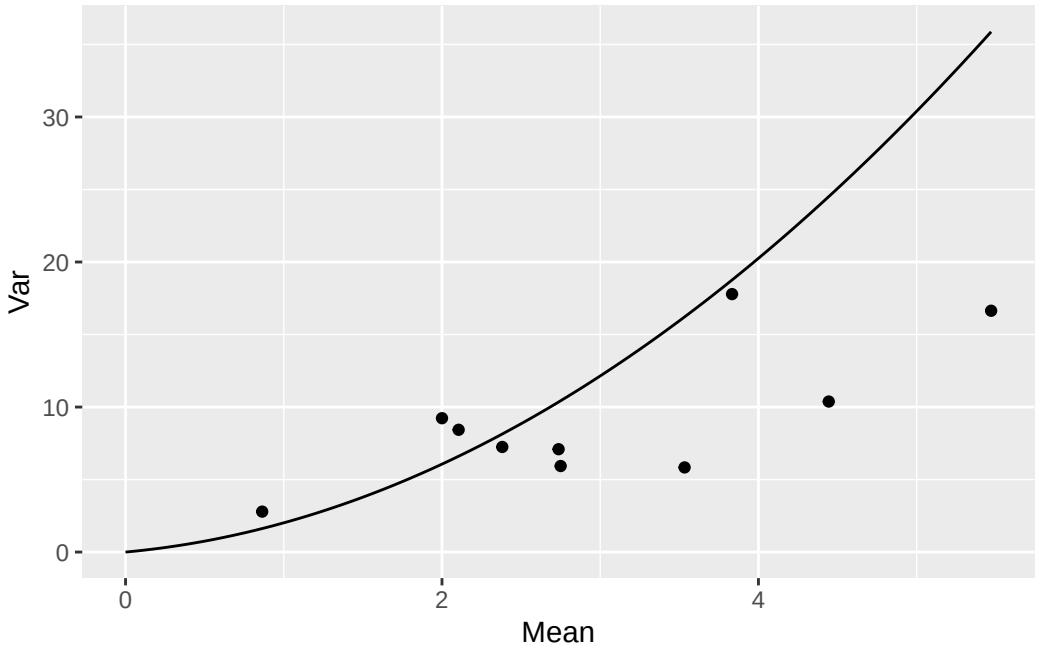

Figure 4: Variance as a function of the mean for the horseshoe crab data. The relationship provided by the negative binomial model is overlayed.

Table 9: Comparison of the negative binomial models including weight alone or both weight and colour as predictors.

```
Likelihood ratio tests of Negative Binomial Models

Response: satellites
           Model     theta Resid. df    2 x log-lik.   Test    df LR stat.
1         weight 0.9310592       171       -748.6437
2 weight * color 0.9844283       165       -743.7922 1 vs 2     6 4.851545
  Pr(Chi)
1
2 0.56299
```

The marginal distribution of n_iY_i then follows a beta-binomial. The expected value and variance are:

$$\begin{aligned} E(n_iY_i) &= n_iE(\pi)_i = n_i\frac{\alpha_i}{\alpha_i+\beta_i} = n_i\mu_i \\ \text{Var}(n_iY_i) &= \frac{n_i\alpha_i\beta_i(\alpha_i+\beta_i+n)}{(\alpha_i+\beta_i)^2(\alpha_i+\beta_i+1)} \\ &= n_i\mu_i(1-\mu_i)\left(\frac{\alpha_i+\beta_i+n}{\alpha_i+\beta_i+1}\right) \\ &= n_i\mu_i(1-\mu_i)\gamma_i \end{aligned}$$

where μ_i is the mean of the beta distribution, $n_i\mu_i(1-\mu_i)$ is the variance of the binomial and

$$\gamma_i = \left(\frac{\alpha_i+\beta_i+n}{\alpha_i+\beta_i+1}\right)$$

is called the variance inflation factor.

Quasi-Likelihood

The second method to account for overdispersion in a generalized linear model is through quasi-likelihood. Recall that we have two ways to conduct inference for a generalized linear model. The first is Wald based inference which depends only local information at the maximum of the likelihood and on the asymptotic normal distribution of the estimators. The second is likelihood ratio based inference which depends on the broad shape of the likelihood. We have provided several reasons for preferring likelihood based inference – the tests are more powerful and confidence intervals are narrower and guaranteed to stay within the parameter space. However, Wald based inference has a significant advantage in that it depends on only a few pieces of information about the distribution of the data. Consider the following facts we know about inference for a GLM in which Y_i, $i = 1, \ldots, N$, are independent random variables belonging to some member of the exponential dispersion family with mean $\mu_i = g^{-1}(\eta_i)$, linear predictor $\eta_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}$, and variance $\text{Var}(Y_i)$.

1) The maximum likelihood estimates are obtained by solving the likelihood equations

$$\sum_{i=1}^{N} \frac{(y_i-\mu_i)x_{ij}}{\text{Var}(Y_i)}\frac{\partial\mu_i}{\partial\eta_i} = 0, \quad j = 0, \ldots, p. \tag{1}$$

2) The maximum likelihood estimators are approximately normally distributed with mean $\boldsymbol{\beta}$, the true parameter values, and covariance matrix $\text{Cov}(\boldsymbol{\beta}) = (X'WX)^{-1}$ where W is diagonal with i^{th} element

$$w_i = \frac{1}{\text{Var}(Y_i)}\left(\frac{\partial\mu_i}{\partial\eta_i}\right)^2. \tag{2}$$

If we assume that the mean and variance of the distribution are linked through the variance function, so that $\text{Var}(Y_i) = \phi v(\mu_i)$ then the likelihood equations become

$$\sum_{i=1}^{N} \frac{(y_i - \mu_i)x_{ij}}{v(\mu_i)} \frac{\partial \mu_i}{\partial \eta_i} = 0, \quad j = 0, \ldots, p. = \tag{3}$$

These equations are fully determined by the relationship between the mean and the linear predictor, which is determined by the link function, and the relationship between the mean and the variance, which is determined by the variance function. If you know these two things then you can compute the maximum likelihood estimators, construct the approximate normal distribution and conduct Wald based inference even if you don't know what the actual likelihood is. This is the basis for quasi-likelihood estimation.

To generate a quasi-likelihood we simply specify the mean variance relationship, plug this into the likelihood equations for the GLM, and then proceed with inference. This method was first suggested by Wedderburn (1974) and is called quasi-likelihood because the method never defines an explicit likelihood function. In fact, it is possible to specify variance functions which lead to quasi-likelihoods for which no corresponding distribution exists.

Quasi-Poisson Model

The simplest quasi-likelihood is constructed by starting with some initial model in the GLM framework and then inflating the variance by a constant multiple. The variance of the Poisson is equal to the mean, and so the new variance function becomes $v(\mu_i) = \phi\mu_i$ where ϕ is now a parameter to be estimated. For the Poisson model $\partial\mu_i/\partial\eta_i = \mu_i$ and the likelihood equations for the quasi-Poisson model become:

$$\sum_{i=1}^{N} \frac{(y_i - \mu_i)x_{ij}}{\phi\mu_i} \mu_i = 0, \quad j = 0, \ldots, p$$

which implies:

$$\sum_{i=1}^{N} (y_i - \mu_i)x_{ij} = 0, \quad j = 0, \ldots, p$$

exactly as we had before. This means that the maximum likelihood estimates for this model will be exactly the same as for the standard Poisson model. However, the i^{th} entry of the W matrix becomes:

$$w_i = \frac{(\partial\mu_i/\partial\eta_i)^2}{v(\mu_i)} = \frac{\mu_i^2}{\phi\mu_i} = \frac{\mu_i}{\phi}.$$

Hence:

$$\text{Cov}(\hat{\boldsymbol{\beta}}) = \phi(X'\text{Diag}(\mu_1, \ldots, \mu_N)X)^{-1}$$

which is ϕ times the asymptotic covariance matrix of the standard Poisson model. This makes perfect sense. If we inflate the variance of the observations by a factor ϕ then there is more uncertainty in the parameter estimates and our standard errors increase.

The fact that the dispersion parameter, ϕ, cancelled out of the likelihood equations should not have come as a too big of a surprise. In the case of a normal linear regression model the estimates of the regression coefficients are independent of the variance estimate. The only time you need to estimate the residual variance, σ^2, is to construct the sampling distribution of the estimators in order to compute standard errors and confidence intervals. The same thing happens here.

We do have to find some way to estimate ϕ, and the method that Wedderburn suggested is based on the magnitude of Pearson's chi-squared statistics, or equivalently, the sum of standardized squared errors. Pearson's chi-squared statistic is

$$X^2 = \sum_{i=1}^{N} \frac{(y_i - \mu_i)^2}{\phi v(\mu_i)}$$

and, under the regularity conditions we've been assuming, $X^2 \dot{\sim} \chi^2_{N-(p+1)}$. In particular

$$E(X^2) \approx N - (p+1).$$

Based on this, Wedderburn suggested estimating the dispersion parameter by equating the method of moments estimator

$$\hat{\phi} = \frac{1}{(N-(p+1))} \sum_{i=1}^{N} \frac{(y_i - \hat{\mu}_i)^2}{v(\hat{\mu}_i)} = \frac{1}{(N-(p+1))} X^2$$

where X^2 is the chi-squared statistic for the standard model with $\phi = 1$.

This suggests a very simple approach for fitting quasi-likelihood models with variance functions of the form $v(\mu) = \phi v^*(\mu)$, where $v^*(\mu)$ is the variance function for a standard GLM:

1) Fit the standard GLM to obtain estimates of the coefficients, $\hat{\beta}_0, \ldots, \hat{\beta}_p$.
2) Estimate the dispersion parameter as:

$$\hat{\phi} = \frac{1}{(N-(p+1))} X^2$$

   where X^2 is the Pearson chi-square statistic for the model fit in part 1.
3) Estimate $\text{Cov}(\hat{\beta})$ by multiplying the asymptotic covariance matrix from the fitted model by $\hat{\phi}$.

Note that for a normal regression model $v(\mu_i) = 1$ and $\phi = \sigma^2$ so that this estimate is exactly equal to the usual unbiased estimate of the residual variance:

$$\hat{\sigma}^2 = \frac{1}{(N-(p+1))} \sum_{i=1}^{N} (y_i - \hat{\mu}_i)^2.$$

For a Poisson log-linear model the estimated dispersion parameter is

$$\hat{\phi} = \frac{1}{(N-(p+1))} \sum_{i=1}^{N} \frac{(y_i - \hat{\mu}_i)^2}{\mu_i}.$$

Table 10: Summary of quasi-Poisson model fit to number of satellites as a function of the female's weight for the horseshoe crab data.

```
Call:
glm(formula = satellites ~ weight, family = quasipoisson(), data = crabs)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -0.4284     0.3168  -1.352    0.178
weight        0.5893     0.1151   5.120 8.17e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for quasipoisson family taken to be 3.13414)

    Null deviance: 632.79  on 172  degrees of freedom
Residual deviance: 560.87  on 171  degrees of freedom
AIC: NA

Number of Fisher Scoring iterations: 5

                2.5 %    97.5 %
(Intercept) -1.0430913 0.1978485
weight       0.3569932 0.8078453
```

Example: Horseshoe Crabs

We will turn once again to the example of the horseshoe crabs to illustrate the quasi-Poisson model. To fit this model in `R` you simply need to define the argument `family=quasipoisson()` within the call to `glm()`. The results of the analysis are shown in Table 10.

Notice that the estimate of the effect of weight on the mean number of satellites is exactly the same for the quasi-Poisson model as for the original Poisson model shown in Table 1. However, most of the rest of the output has changed. In particular, where the dispersion parameter has always been set equal to 1 before `R` now tells us that the dispersion parameter has been estimated ("taken to be") 3.13. The confidence intervals are also wider than before. Here the 95% confidence interval for the effect of weight extends from .35 to .81, where as before the interval went from .46 to .81. It shouldn't surprise you that the relative width of the confidence intervals is $1.77 = \sqrt{3.13}$. The z-value has also decreased and, correspondingly, the p-value has increased.

The residual deviance is also exactly the same for the Poisson and quasi-Poisson models, but we have to be a little careful here. Recall that the likelihood ratio test statistic is actually equal to the difference in the *scaled* deviance. The test statistic to assess the goodness-of-fit for the quasi-Poisson model is $560.87/3.13 = 179.19$. This is slightly more than the degrees of freedom, 171, and the computed p-value would be .32 suggesting that we would not reject the fit of the model.

Technically, we should discretize the covariate to assess the fit of the model, but there's no reason to do this. You will never reject the fit of a quasi-Poisson model (or any other quasi-likelihood model) because the dispersion parameter has essentially been introduced to account for any lack of fit. It's possible to show that if N is large enough then the Pearson chi-square statistic and residual deviance are approximately equal for any GLM. This means that $X^2 \approx G^2$ and so, using Wedderburn's estimator,

$$\hat{\phi} \approx \frac{1}{(N-(p+1))} G^2.$$

Simply rearranging this shows that

$$\frac{G^2}{\hat{\phi}} \approx (N-(p+1))$$

which is its degrees of freedom. This means that the scaled deviance of a quasi-likelihood model will always be close to the mean of the null distribution and the p-value will always be large.

Quasi-binomial Model

Quasi-likelihood models can also be constructed in the case of logistic regression. Once again, we will consider the quasi-likelihood where the variance is simply inflated by a constant. The variance of the binomial model is $\text{Var}(Y_i) = \mu_i(1-\mu_i)/n_i$, and so the new variance function becomes $v(\mu_i) = \phi\mu_i(1-\mu_i)/n_i$ where ϕ is again a parameter to be estimated. For the binomial model $\partial\mu_i/\partial\eta_i = \mu_i(1-\mu_i)$ and the likelihood equations for the quasi-binomial model are simply:

$$\sum_{i=1}^{N} \frac{n_i(y_i-\mu_i)x_{ij}}{\phi\mu_i(1-\mu_i)} \cdot \mu_i(1-\mu_i) = 0, \quad j = 0, \ldots, p$$

which implies again that:

$$\sum_{i=1}^{N} n_i(y_i-\mu_i)x_{ij} = 0, \quad j = 0, \ldots, p$$

exactly as we had before. As for the Poisson model, the parameter estimates are exactly the same as for the standard model. However, the i^{th} entry of the W matrix becomes:

$$w_i = \frac{(\partial\mu_i/\partial\eta_i)^2}{v(\mu_i)} = \frac{[\mu_i(1-\mu_i)]^2}{\phi\mu_i(1-\mu_i)/n_i} = \frac{n_i\mu_i(1-\mu_i)}{\phi}.$$

Hence:

$$\text{Cov}(\hat{\beta}) = \phi(X'\text{Diag}(n_1\mu_1(1-\mu_1), \ldots, n_N\mu_N(1-\mu_N))X)^{-1}$$

which is ϕ times the asymptotic covariance matrix of the standard binomial model.

The model can fit in exactly the same way as before:

1) Fit the standard GLM to obtain estimates of the coefficients, $\hat{\beta}_1, \ldots, \hat{\beta}_p$.
2) Estimate the dispersion parameter as:

$$\hat{\phi} = \frac{1}{(N-(p+1))} X^2$$

   where X^2 is the Pearson chi-square statistic for the model fit in part 1.
3) Estimate $\text{Cov}(\hat{\beta})$ by multiplying the asymptotic covariance matrix from the fitted model by $\hat{\phi}$.

While this seems to make sense mathematically, there are some problems with this model. Suppose, for example, that the data are Bernoulli so that $n_i = 1$ for all $i = 1, \ldots, N$. You could still follow the procedure above to fit the quasi-Binomial model and obtain estimates and standard errors for the beta parameters, and it's likely that $\hat{\phi}$ won't be exactly equal to 1. But, this is mathematically impossible. If Y_i takes the value 0 with probability $1-\mu_i$ and 1 with probability μ_i then the variance of Y_i has to be $\mu_i(1-\mu_i)$, it simply cannot be larger or smaller. There is no distribution on the values 0 and 1 that would produce the correct mean and variance if $\phi \neq 1$. This is one example of a quasi-likelihood model that is not associated with any real probability model.

Conclusion

Quasi-likelihood is like a giant broom for cleaning up messes. With one fell swoop you will sweep all of the problems under the rug and your model will appear to fit the data properly. However, your model probably won't have any biological or physical justification, though it is easy to paint stories afterward, and it's possible that your apparent model doesn't actually define a distribution for the data, which is simply bizarre.

While it is important to make sure that your model fits the data, quasi-likelihood should only be considered as a last resort. You should try your hardest beforehand to figure out why your model isn't fitting the data. Do you need to transform one of the predictors or add a polynomial term to your model? Have you forgotten an important predictor that explains the results? Are there outliers that might be inflating the deviance? Could there be a source of dependence between the observations that you haven't considered? If you've checked all of these (and more) and still can't fit the data well then you should try to model the overdispersion. Can you account for the extra variance in counts by using the negative binomial model instead of the Poisson model? Does a beta-binomial explain appropriately model the extra variation in proportions? Only after you've explored all of these possibilities should you throw up your

hands and fit a quasi-likelihood model to artificially inflate the dispersion parameter to account for unknown variability in your data.

References

Brockmann, Jane H. 1996. "Satellite Male Groups in Horseshoe Crabs, Limulus Polyphemus." *Ethology* 102 (1): 1–21.