ANALYSIS OF PROPORTIONS FROM SPLIT-PLOT AND REPEATED MEASURES EXPERIMENTS

Kenneth J. Koehler

Recommended Citation

ANALYSIS OF PROPORTIONS FROM SPLIT-PLOT AND REPEATED MEASURES EXPERIMENTS

Kenneth J. Koehler

Iowa State University

Several methods for analyzing proportions from split-plot and repeated measures experiments are illustrated and compared. One approach simply uses analysis of variance for the usual linear mixed model fit to split-plot and repeated measures experiments. Alternatively, logistic regression analysis is considered and a so-called robust estimate of the covariance matrix is used to adjust for possible correlations among responses. Finally, a quasi-likelihood approach to logistic regression analysis that requires more explicit specification of the covariance structure for the observed proportions is considered. These methods are illustrated with the analyses of data from a repeated measures study of acorn consumption by blue jays and a study of the effects of several environmental factors on nest predation for ground nesting birds.

KEY WORDS: logistic regression, robust covariance estimation, quasi-likelihood

1. Introduction   Studies involving repeated measures across time or measurements on sub-plots within whole plots frequently occur in agricultural and ecological research, and it is not unusual to have binary or multi-category responses. Two illustrations with binary responses are considered in their article. One study examines the effects of several environmental factors on nest predation among ground nesting birds. The binary response for a single nest is whether or not it was disturbed by a predator. In the other study, the binary response is whether or not a blue jay selects a particular acorn from a tray. There are many other applications. A botanist, for example, may repeatedly examine a particular site over a period of years to determine if a particular plant species exists or is dominant at the site. In a split-plot field experiment, the presence or absence of a particular disease, or the presence or absence of a certain insect species, may be recorded for plants in each sub-plot.

Commonly used statistical methods for counts and proportions, such as chi-square tests provided by PROC FREQ and logistic regression analysis provided by PROC LOGISTIC in SAS, do not account for correlations among responses arising in repeated measures studies and split-plot experiments. Provided the form of the logistic model is correctly specified, PROC LOGISTIC will generally produce consistent estimates of regression coefficients that have a limiting normal distribution, but PROC LOGISTIC will usually not provide an appropriate estimate of the covariance matrix or appropriate standard errors for the estimates of the regression parameters. We present a simple correction to the covariance matrix produced by PROC LOGISTIC, that is often referred to as a robust estimator for the covariance matrix. This approach does not require a specification of the form of the covariance matrix for repeated measurements taken from a single subject or responses obtained from different sub-plots within a single whole plot. We consider a second approach where the covariance structure is directly modeled and used in the estimation of the coefficients in the logistic regression model. Inferences from both methods are compared with ANOVA results from the linear mixed model commonly used for split-plot experiments and repeated measures studies.

2. Applications.

We consider two applications in animal ecology. The first is small study of blue jay consumption of acorns involving just one between bird factor and one within bird factor. In the second application we analyze data from a study of the effects of several environmental factors on nest predation rates for ground nesting birds. This is a larger and somewhat more complex experiment.

2.1 Blue Jay Study. When other sources of food are scarce, blue jays will eat acorns even though acorns contain a tannin that inhibits there ability digest protein. To investigate the hypothesis that blue jays offset the effect of the tannin by increasing protein uptake through selective foraging of acorns infested with weevil larvae, Dixon (1994) performed the following experiment.

Ten blue jays were used in the experiment. Five of the birds (USED birds) had been previously used in a similar experiment, and the other five (NEW birds) had not been used in any previous experiment. This is the between bird factor. The birds were kept in separate

cages and it could reasonably be assumed that each bird acted independently of any other bird. After not being fed for 24 hours, each bird was presented with a tray containing 8 weevil infested acorns and 8 uninfested acorns. The acorns were randomly arranged in the trays. The numbers of infested and uninfested acorns taken from the tray by each bird are shown in Table 1. Overall, 57 of the 80 uninfested acorns were taken and 32 of the 80 weevil infested acorns were taken, offering no support for the supposition that motivated the study.

2.2 Nest Predation Study.   This study was directed by Timothy Bergin and Louis Best, Department of Animal Ecology, Iowa State University, and Kathryn Freemark, National Wildlife Centre, Environment Canada, to study the effects of local habitat factors on nest predation along roadsides for ground nesting birds. Potential predators included raccoons (Procyon lotor), striped skunks (Mephitis mephitis), red fox (Vulpes vulpes), coyotes (Canis latrans), mink (Mustela vison), badgers (Taxidea taxa), grackles (Quisculus quiscula), crows (Corvus brachyrhynchos), bullsnakes (Pituophis melanolencus sayi), and rodents.

Transects consisting of two rows of nests, with 5 nests in each row, were set up in 136 locations along roadsides in 6 watersheds located in agricultural landscapes in South-Central Iowa. Roads are generally laid out as 1-square mile sections in each watershed. One transect was placed in each section selected for the study and the roadside was randomly selected from the four sides of the square section subject to the constraints: 1) transects in different sections could not be located on opposite sides of the same segment of road to maintain independence of results in different transects, and 2) no transect could not cross drainages, water courses, or roadways that might prevent predators from treating it as a single entity. A 200 meter buffer was maintained at the end of each road segment to avoid road intersections, and the transect was randomly positioned between the end buffers.

The transects were located in the drainage ditches along the road sides. Each transect contained two rows of artificial nests, with one row placed 2 meters from the road edge (along the fore-slope of the drainage ditch) and the other row located along the back-slope of the drainage ditch. Each row contained 5 nests spaced 20 meters apart and the rows were offset. Two Coturnix quail eggs were placed in each nest. For each transect, the nests were checked at the end of a 7 day exposure period during June, the main breeding season for common

farm land birds. If at least one egg was broken or removed from a nest, its fate was assigned to predation.

The data consist of the number predated nests and the total number of nests in both the fore-slope and back-slope of each transect along with information on the local environment for each transect. The road adjacent to the transect was classified at either paved or unpaved. The habitat adjacent to the other side of the transect was classified as either row crop or non-row crop. The border between the transect and the adjacent habitat was classified into three categories: herbaceous (less than 5% woody cover) without a fence, herbaceous with a fence, or wooded (at least 5% woody cover). Fences and woody cover provide perches for avian predators and shelter for raccoons and other predators. These are whole plot (between transects) factors. The fore-slope/back-slope factor is a sub-plot (within transect) factor.

For effective use of PROC LOGISTIC in SAS (SAS Institute, 1994), these factors were coded as binary regression variables:

$$\text{Road type:} \qquad X_1 = \begin{cases} 0 & \text{unpaved} \\ 1 & \text{paved} \end{cases}$$

$$\text{Adjacent habitat:} \qquad X_2 = \begin{cases} 0 & \text{non-row crop} \\ 1 & \text{row crop} \end{cases}$$

$$\text{Border habitat:} \qquad X_3 = \begin{cases} 0 & \text{other} \\ 1 & \text{herbaceous with no fence} \end{cases}$$

$$X_4 = \begin{cases} 0 & \text{other} \\ 1 & \text{woody} \end{cases}$$

$$\text{Foreslope/backslope:} \qquad X_5 = \begin{cases} 0 & \text{fore-slope} \\ 1 & \text{back-slope} \end{cases}$$

Two interaction terms, $X_1 X_3$ and $X_2 X_4$, are also included in the following analyses. Preliminary analyses showed that there were no interactions between the whole plot factors and the sub-plot factor and other interactions between whole plot factors were also insignificant. Consequently, more complex models will not be considered in this article.

The complete data set is too large to present here, but a summary of the observed nest predation rates is given in Table 2 for all combinations of factors occurring in the study. Note

that the combination of habitat factors corresponding to a woody border with an adjacent row group did not occur along a paved road in this study. Twenty-eight of the 1360 nests set up in this experiment were by weather or mowing and they were excluded from Table 2 and further analysis. Of the remaining 1332 nests, 307 nests (23%) were disturbed by predators.

3. Linear mixed models.

Simple analyses of the data sets described in the previous section are provided by a linear mixed model. Let P_{ij} denote the observed proportion of successes for the j-th observation taken on the i-th primary unit (bird or transect). Then a mixed model for the blue jay study is

$$P_{ij} = \mu + (\text{bird type}) + \eta_i + (\text{acorn type}) + (\text{bid by acorn interaction}) + \epsilon_{ij} \qquad (3.1)$$

where $\eta_i \sim$ i.i.d.$(0, \sigma_n^2)$ are random bird effects and $\epsilon_{ij} \sim$ i.i.d.$(0, \sigma_\epsilon^2)$. Similarly, a linear mixed model for the nest predation data is

$$P_{ij} = \begin{aligned} &\beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} \\ &+ \beta_5 X_{1i} X_{3i} + \beta_6 X_{2i} X_{4i} + \eta_i + \beta_7 X_{5ij} + \epsilon_{ij} \end{aligned} \qquad (3.2)$$

where $\eta_i \sim$ i.i.d.$(0, \sigma_\eta^2)$ are random transect effects and $\epsilon_{ij} \sim$ i.i.d.$(0, \sigma_\epsilon^2)$ are random within transect errors.

The mixed linear model approach has some attractive features, but at best it provides only a good approximation. ANOVA tables and F-tests are easily computed with standard statistical software, such as PROC GLM or PROC MIXED in SAS. The inclusion of η_i in the model induces a correlation of $\sigma_\eta^2 / (\sigma_\eta^2 + \sigma_\epsilon^2)$ between any two proportions (P_{ij}, P_{ik}) obtained from the same primary unit. This equal correlation assumption may be called into question in studies where there are more than two subunits in each primary unit. Moreover, since variances of sample proportions are generally functions of their expectations, the homogeneous variance assumption,

$$V(P_{ij}) = \sigma_\eta^2 + \sigma_\epsilon^2 \ ,$$

may be an issue of greater concern, especially when a substantial proportion of the observed P_{ij} values are zero (or one) or the model suggests a wide range of values for $\pi_{ij} = E(P_{ij})$. Finally, the mixed linear model could yield estimates of π_{ij} less than zero or greater than one.

These are not issues of great concern in either the blue jay data or the nest predation data because estimates of the π_{ij}'s are mostly between .2 and .8. Also, each P_{ij} is computed from a sample size of 8 acorns in the blue jay study and most P_{ij}'s are computed from a sample size of 5 nests in the nest predation study. Consequently, F-tests for the linear mixed model analysis provide reliable inferences for these two studies.

The ANOVA table for the blue jay study is shown in Table 3. This analysis shows a significant difference between acorn types, with a higher proportion of uninfested acorns taken. There appears to be no significant difference between NEW and USED birds, and no interaction between bird type and acorn type. An analysis of $\arcsin(\sqrt{P_{ij}})$ yields similar F-values and essentially the same inferences (see Table 7). The ANOVA table for the nest predation study is shown in Table 4. Comparisons of p-values with the same model applied to $\arcsin(\sqrt{P_{ij}})$ are provided by the last two columns of Table 8.

4. Logistic Regression.

Logistic regression analysis has become the most popular method of analyzing the effects of covariates on proportions in the human health sciences and its use in agricultural and ecological sciences is increasing. Standard implementation of logistic regression, such as PROC LOGISTIC in SAS, are based on an assumption that each observed binary outcome is an independent Bernoulli trial which is generally violated in split-plot experiments and repeated measures studies. Nevertheless, standard logistic regression will provide consistent estimates of regression parameters, but the computed covariance matrix for the parameter estimator will generally be inappropriate. In particular, standard errors of regression parameters tend to be too small when there is positive correlation among responses from repeated measures on the same subject or from sub-plots within the same whole plot. This follows from the work of Huber (1967) on properties of estimators obtained from maximizing the wrong likelihood function. White (1982) provides a more accessible account, and Liang and Zeger (1986) apply this approach to longitudinal studies. We will describe maximum likelihood estimation for the standard logistic regression model in this section and describe alternative robust estimators for covariance matrices and standard errors. Alternative estimators for regression coefficients are described in Section 5.

4.1 Independence Working Model.

Using the notation from Section 2, let π_{ij}

denote the conditional probability that a particular acorn of the j-th type (infested or unin-fested) is selected by the i-th bird. The logit link function,

$$\log\left[\pi_{ij}/(1-\pi_{ij})\right] = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2j} + \beta_3 X_{1i} X_{2j} = \mathbf{X}'_{ij}\,\boldsymbol{\beta}\ , \qquad (4.1)$$

links this probability to a linear combination of covariates,

$$\text{Bird type: } X_1 = \begin{cases} 0 & \text{New} \\ 1 & \text{Used} \end{cases} \qquad \text{Acorn type: } X_2 = \begin{cases} 0 & \text{non-row crop} \\ 1 & \text{row crop} \end{cases}$$

Then,

$$\pi_{ij}(\boldsymbol{\beta}) = \exp\left(\mathbf{X}'_{ij}\,\boldsymbol{\beta}\right) \Big/ \left[1 + \exp(\mathbf{X}'_{ij}\,\boldsymbol{\beta})\right]\ . \qquad (4.2)$$

Assuming that the result for each of the n_{ij} acorns of the j-th type presented to the i-th bird is an i.i.d. Bernoulli trial, then

$$Y_{ij} \sim \text{Binomial}\,[\text{n}_{ij},\,\pi_{ij}\,(\boldsymbol{\beta})] \qquad (4.3)$$

where Y_{ij} is the number of acorns of the j-th type taken by the i-th bird. Further assuming that each bird responds independently of any other bird, the joint log-likelihood function is

$$\begin{aligned} L(\boldsymbol{\beta}) &= \sum_{i=1}^{N}\sum_{j=1}^{m_i}\left\{\log\binom{n_{ij}}{Y_{ij}} + Y_{ij}\log\left[\pi_{ij}(\boldsymbol{\beta})\right] + (n_{ij}-Y_{ij})\log\left[1-\pi_{ij}(\boldsymbol{\beta})\right]\right\} \\ &= \sum_{i=1}^{N}\sum_{j=1}^{m_i}\left\{\log\binom{n_{ij}}{Y_{ij}} + Y_{ij}\,\mathbf{X}'_{ij}\,\boldsymbol{\beta} + n_{ij}[1+\exp(\mathbf{X}'_{ij}\,\boldsymbol{\beta})]^{-1}\right\}\ . \end{aligned} \qquad (4.4)$$

In the blue jay study $m_i = 2$ types of acorns for each bird. Maximum likelihood estimates for the elements of $\boldsymbol{\beta}$ are obtained by setting the score function

$$U(\boldsymbol{\beta}) = \frac{\partial\,L(\boldsymbol{\beta})}{\partial\,\boldsymbol{\beta}} = \sum_{i=1}^{N} X'_i\,W_i(\mathbf{P}_i - \boldsymbol{\pi}_i(\boldsymbol{\beta})) \qquad (4.5)$$

equal to the zero vector and solving the set of equations. Here

$$X_i = \begin{bmatrix} \mathbf{X}'_{i1} \\ \vdots \\ \mathbf{X}'_{i,m_i} \end{bmatrix} \qquad w_i = \begin{bmatrix} n_{i1} & & \\ & \ddots & \\ & & n_{i\,m_i} \end{bmatrix} \qquad \boldsymbol{\pi}_i(\boldsymbol{\beta}) = \begin{bmatrix} \pi_{i1}(\beta) \\ \vdots \\ \pi_{i\,m_i}(\boldsymbol{\beta}) \end{bmatrix}$$

and $\mathbf{P}_i = [Y_{i1}/n_{i1},\cdots,Y_{i\,m_i}/n_{i\,m_i}]'$ is the vector of observed proportions on the i-th primary unit (e.g., bird or transect). Generally, there is no closed form solution to these equations and

some numerical optimization algorithm must be used to obtain the values of the maximum likelihood estimates (m.l.e.).

The model underlying (4.4) and (4.5) is often referred to as an "Independence Working Model," and we will denote the m.l.e. by $\hat{\boldsymbol{\beta}}_{\text{IWM}}$. When the assumptions of the Independence Working Model (IWM) are satisfied, the limiting normal distribution for $\hat{\boldsymbol{\beta}}_{\text{IWM}}$ is given by

$$\sqrt{N}\left(\hat{\boldsymbol{\beta}}_{\text{IWM}} - \boldsymbol{\beta}\right) \xrightarrow{\text{dist'n}} N(\mathbf{0}, N\, C_1^{-1}) \tag{4.6}$$

where

$$C_1 = \sum_{i=1}^{N} (X_i'\, W_i\, V_i\, W_i\, X_i) \tag{4.7}$$

is the IWM Fisher Information Matrix and the diagonal matrix

$$V_i = \text{diag}\left\{\frac{\pi_{i1}(\boldsymbol{\beta})[1 - \pi_{i1}(\boldsymbol{\beta})]}{n_{i1}}, \cdots, \frac{\pi_{i,m_i}(\boldsymbol{\beta})[1 - \pi_{i,m_i}(\boldsymbol{\beta})]}{n_{i\,m_i}}\right\} \tag{4.8}$$

is the IWM covariance matrix for $\mathbf{P}_i$.

4.2 Robust Covariance Estimation. The estimate $\hat{C}_1^{-1}$, obtained by evaluating (4.7) and (4.8) at $\hat{\beta}_{\text{IWM}}$, yields a good approximation to the covariance matrix of $\hat{\beta}_{\text{IWM}}$ when the assumptions of the IWM model are satisfied. For the blue jay data in Table 1, the IWM model implies that the observed proportions of invested and uninfested nuts taken by a bird $\{P_{i1}, P_{i2}\}$ are uncorrelated. Otherwise, V_i in (4.8) is not a diagonal matrix. Moreover, it implies that $\{Y_{i1} = 8\, P_{i1}, i = 1, \ldots, 5\}$ are i.i.d. binomial random variables and $\{Y_{i1} = 8\, P_{i1}, i = 6, \ldots, 10\}$ are i.i.d. binomial random variables. Since the bird type factor is not significant and the overall average number of uninfested acorns taken per bird is 5.7, we would expect many of the observed counts to be 5, 6, or 7. Instead, most counts are more extreme, either at the upper bound of 8 or below 5, and the observed counts exhibit greater variation than binomial variances can accommodate. Some birds appear to have a stronger preference for uninfested acorns than other birds. Some birds take all or nearly all of the uninfested acorns and other birds take only a few. Hence, this extra-binomial variability in counts can also be viewed as positive within bird correlation. If a particular uninfested acorn is not taken it is likely the most other uninfested acorns and also not taken, but if it was taken then it is likely all of the other uninfested acorns were taken. Hence, the diagonal

elements of V_i in (4.8) are too small when there is positive within subject or within sub-plot correlation.

Although correlations among repeated measures, or correlations within whole plots, invalidate the use of $\hat{C}_1^{-1}$ as an estimate of the covariance matrix of $\hat{\beta}_{\text{IWM}}$, $\hat{\beta}_{\text{IWM}}$ still provides a consistent estimator for β in (4.1) with a limiting normal distribution. A consistent estimator for the covariance matrix of $\hat{\beta}_{\text{IWM}}$ is

$$\hat{C}^{-1}\,\hat{C}_2\,\hat{C}^{-1}\,, \tag{4.9}$$

where

$$\hat{C}_2 = \sum_{i=1}^{N} X_i'\,W_i\,[\mathbf{P}_i - \boldsymbol{\pi}_i(\hat{\boldsymbol{\beta}}_{\text{IWM}})]\,[P_i - \boldsymbol{\pi}_i(\hat{\boldsymbol{\beta}}_{\text{IWM}})]'\,W_i\,X_i \tag{4.10}$$

is an estimate of the covariance matrix of the score function in (4.5). Note that $\hat{C}_2$ is the formula for $\hat{C}_1$ with $\hat{V}_i$, the estimated IWM covariance matrix for $\mathbf{P}_i$, replaced by $[\mathbf{P}_i - \boldsymbol{\pi}_i(\hat{\boldsymbol{\beta}}_{\text{IWM}})]\,[\mathbf{P}_i - \boldsymbol{\pi}_i(\hat{\boldsymbol{\beta}}_{\text{IWM}})]'$, essentially a one-degree-of-freedom estimate of an arbitrary covariance matrix for $\mathbf{P}_i$. $\hat{C}_1^{-1}\,\hat{C}_2\,\hat{C}_1^{-1}$ is often called a "sandwich" estimator or a "robust" estimator. It allows for arbitrary correlation among Bernoulli outcomes within sub-plots and between sub-plots in the same whole plot. Values of $\hat{\boldsymbol{\beta}}_{\text{IWM}}$ and standard errors computed from both $\hat{C}_1^{-1}$ and $\hat{C}_1^{-1}\,\hat{C}_2\,\hat{C}_1$ are shown in Table 5 for the blue jay study and Table 6 for the nest predation study.

Experience has shown that $\hat{C}_1^{-1}\,\hat{C}_2\,\hat{C}_1^{-1}$ provides a surprisingly good estimator in large samples (e.g., $N > 100$ and $\boldsymbol{\beta}$ of low dimension), but its behavior in small samples, where it could be rather inefficient, is largely unexplored. Pendergast, et. al. (1996) further discussion of this approach and references to simulation studies.

4.3 Quasi-likelihood estimation. This approach requires the specification of a model for the covariance matrix of $\mathbf{P}_i$ based on the elements of $\boldsymbol{\pi}_i(\boldsymbol{\beta})$ and a few additional extra-variation or correlation parameters. For both the blue jay and nest predation studies we propose

$$\Sigma_i = V_i^{1/2} \begin{bmatrix} \theta_1 & 0 \\ 0 & \theta_2 \end{bmatrix}^{1/2} \begin{bmatrix} 1 & \gamma \\ \gamma & 1 \end{bmatrix} \begin{bmatrix} \theta_1 & 0 \\ 0 & \theta_2 \end{bmatrix}^{1/2} V_i^{1/2} \tag{4.11}$$

as the covariance matrix for $\mathbf{P}_i$, where V_i is defined in (4.8). Here γ denotes the correlation between P_{i1} and P_{i2}, and θ_1 and θ_2 are extra binomial variation parameters for P_{i1} and P_{i2},

respectively. Note that Σ_i reduces to V_i in (4.8) when $\theta_1 = \theta_2 = 1$ and $\gamma = 0$, requirements of the IWM.

Given a formula for Σ_i, quasi-likelihood estimation (McCullagh and Nelder, 1989, Chapter 8), or equivalently Gauss-Newton estimation (Morel and Koehler, 1995), could be used to obtain an improved estimator for $\boldsymbol{\beta}$. This is done by solving the equations obtained by setting the "score function,"

$$\mathbf{U}(\boldsymbol{\beta}) = \sum_{i=1}^{N} X_i' V_i W_i \Sigma_i^{-1} \left(\mathbf{P}_i - \boldsymbol{\pi}_i \right) , \tag{4.12}$$

equal to the zero vector. The solution, denoted by $\hat{\boldsymbol{\beta}}_Q$, must be obtained numerically. This can be done by starting with $\hat{\boldsymbol{\beta}}^{(0)} = \hat{\boldsymbol{\beta}}_{\text{IWM}}$ and computing

$$\hat{\boldsymbol{\beta}}^{(k)} = \hat{\boldsymbol{\beta}}^{(k-1)} + \left[\sum_i X_1' \hat{V}_i W_i \hat{\Sigma}_i^{-1} W_i \hat{V}_i X_i \right]^{-1} \sum_{i=1}^{N} X_i' \hat{V}_i W_i \hat{\Sigma}_i^{-1} \left(\mathbf{P}_i - \hat{\boldsymbol{\pi}}_i \right)$$

At each step, $\hat{\boldsymbol{\pi}}_i = \boldsymbol{\pi}_i(\hat{\boldsymbol{\beta}}^{(k-1)})$ and $\hat{V}_i$ and $\hat{\Sigma}_i$ are evaluated at $\hat{\boldsymbol{\beta}}^{(k-1)}$ and estimates of the extra variation/correlation parameters

$$\hat{\theta}_j = \frac{\sum\limits_{i=1}^{N} n_{ij} \left(P_{ij} - \hat{\pi}_{ij} \right)^2}{\sum\limits_{i=1}^{N} \hat{\pi}_{ij} \left(1 - \hat{\pi}_{ij} \right)} , \quad j = 1, 2 \tag{4.13}$$

and

$$\hat{\gamma} = \frac{\sum\limits_{i=1}^{N} n_{i1}^{-1/2} \left(P_{i1} - \hat{\pi}_{i1} \right) n_{i2}^{-1/2} \left(P_{i2} - \hat{\pi}_{i2} \right)}{\sqrt{\sum\limits_{i=1}^{N} \hat{\pi}_{i1}(1 - \hat{\pi}_{i1})\hat{\pi}_{i2}(1 - \hat{\pi}_{i2})}}. \tag{4.14}$$

Alternatively, a sum of squared Pearson residuals is often used to estimate $\hat{\theta}_j$ (Williams, 1982).

The limiting normal distribution for the final estimate $\hat{\boldsymbol{\beta}}_Q$ is given by

$$\sqrt{N} \left(\hat{\boldsymbol{\beta}}_Q - \boldsymbol{\beta} \right) \xrightarrow{\text{dist'n}} N(\mathbf{0}, V_Q) \quad \text{as} \quad N \to \infty$$

where

$$V_Q = \lim_{N \to \infty} N \left[\sum_i X_i' V_i W_i \Sigma_i^{-1} W_i V_i X_i \right]^{-1} . \tag{4.15}$$

Using $\hat{V}_i$ and $\hat{\Sigma}_i$ to denote V_i and Σ_i evaluated at $\hat{\boldsymbol{\beta}}_Q$ and the final values of $\hat{\theta}_1$, $\hat{\theta}_2$ and $\hat{\gamma}$, an estimate of the covariance matrix of $\hat{\boldsymbol{\beta}}_Q$ is

$$\left[\sum_i X_1' \, \hat{V}_i \, W_i \, \hat{\Sigma}_i^{-1} \, W_i \, \hat{V}_i \, X_i \right]^{-1} . \tag{4.16}$$

Values of $\hat{\boldsymbol{\beta}}_Q$ and standard errors obtained from (4.16) are shown in the last 2 columns of Tables 5 and 6 for the blue jay and nest predation data sets, respectively. For the blue jay study, $\hat{\theta}_1 = 2.56$, $\hat{\theta}_2 = 1.09$, and $\hat{\gamma} = 0.30$. For the nest predation study, $\hat{\theta}_1 = 1.62$, $\hat{\theta}_2 = 2.08$ and $\hat{\gamma} = 0.29$. These estimates indicate moderate positive correlations among binary outcomes both within sub-plots (acorn types) and between sub-plots (acorn types) in the same transect (bird).

5. Discussion

The values of $\hat{\boldsymbol{\beta}}_{IWM}$ are nearly the same as the corresponding values of $\hat{\boldsymbol{\beta}}_Q$ in both Tables 5 and 6. This is a typical result in such studies. The loss of efficiency in using $\hat{\boldsymbol{\beta}}_{IWM}$ instead of $\hat{\boldsymbol{\beta}}_Q$ is usually quite small. The IWM standard errors, however, tend to be too small in the presence of positive correlations among the binary responses. The robust estimator provides a good adjustment to the covariance matrix for $\hat{\boldsymbol{\beta}}_{IWM}$ and inferences derived from $\hat{\boldsymbol{\beta}}_{IWM}$ and robust standard errors are similar to inferences derived from $\hat{\boldsymbol{\beta}}_Q$.

Simple ANOVA methods need not always be abandoned, however, in favor of logistic regression or some other generalized linear model. Analysis of variance for standard linear mixed models provides essentially the same inferences as the quasi-likelihood approach to logistic regression for both the blue jay and nest predation studies. Tables 7 and 8 show that the use of the variance stabilizing transformation $\arcsin(\sqrt{P_{ij}})$ provides a closer match with the p-values from the quasi-likelihood approach to logistic regression, but even without the variance stabilizing transformation the split-plot ANOVA is not terribly misleading.

Differences in results from the three approaches: logistic regression with robust covariance estimation, the quasi-likelihood approach to logistic regression, and the linear mixed model ANOVA, are small relative to the ability of any of the models to approximate the true underlying biological relationships. Hence, it seems silly to quibble about which approach is more "correct" in either of these two studies. Of course, the similarity between

a linear model and a generalized linear model with a logistic link function will deteriorate as response probabilities approach either zero or one, and a logistic model would generally be preferred in situations where response probabilities and smaller than 0.1 or larger than 0.9.

References

Dixon, M. D. (1994). *Acorn Caching and Consumption of Weevil-infested Nuts: Possible Means of Tannin Circumvention for Blue Jays*, M.S. Thesis, Department of Wildlife and Fisheries Sciences, South Dakota State University.

Huber, P. J. (1967). The Behavior of Maximum Likelihood Estimates under Nonstandard Conditions, *Proceedings of the Fifth Berkeley Symposium*, 221-233.

Liang, K. Y. and Zeger, S. L. (1986). Longitudinal Data Analysis Using Generalized Linear Models, *Biometrika*, **73**, 13-22.

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

Morel, J. and Koehler, K. J. (1995). A One-Step Gauss-Newton Estimator for Categorical Data with Extraneous Variation, *Applied Statistics*, **44**, 187-200.

Pendergast, J. F., Gange, S. J., Newton, M. A., Lindstrom, M. J., Palta, M., and Fisher, M. R. (1996). A Survey of Methods for Analyzing Clustered Binary Response Data, *International Statistical Review*, **64**, 89-118.

White, H. (1982). Maximum Likelihood Estimation of Misspecified Models, *Econometrica*, **50**, 1-26.

Williams, D. A. (1982). Extra-binomial Variation in Logistic Linear Models, *Applied Statistics*, **31**, 144-148.

Table 1: Numbers of weevil-infested (W) and uninfested (U) acorns taken from trays presented to 10 different blue jays.

New Birds				Used Birds		
	Acorn Type				Acorn Type	
Bird	U	W		Bird	U	W
1	8	1		6	7	5
2	4	2		7	2	3
3	3	2		8	8	4
4	6	6		9	8	3
5	3	2		10	8	4

Table 2. Observed nest predation rates.

				Foreslope		Backslope	
				Number of	Observed predation	Number of	Observed predation
Habitat variables				nests	rate	nests	rate
X_1	X_2	X_3	X_4				
0	0	1	0	5	0.00	5	0.00
0	0	0	1	10	0.10	10	0.40
0	0	0	0	34	0.15	35	0.11
0	1	1	0	25	0.24	25	0.36
0	1	0	0	58	0.17	60	0.30
1	0	1	0	14	0.14	14	0.29
1	0	0	1	65	0.35	65	0.46
1	0	0	0	124	0.08	123	0.28
1	1	1	0	55	0.07	54	0.09
1	1	0	1	46	0.15	50	0.26
1	1	0	0	231	0.18	224	0.34

Table 3. Linear mixed model ANOVA for the blue jay data.

Source of variation	df	SS	MS	F	p-value
Bird type	1	.1758	.1758	2.42	.158
Birds within type	8	.5813	.0727		
Acorn type	1	.5695	.5695	14.02	.005
Bird type × acorn type	1	.0195	.0195	0.48	.507
Within bird error	8	.3250	.0406		

Table 4. Linear mixed model ANOVA (Type III sums of squares) for the nest predation data.

Source of variation	df	SS	MS	F	p-value
Road type (X_1)	1	0.1574	.1574	1.82	.1800
Adj. habitat (X_2)	1	0.2508	.2508	2.90	.0912
No Border Fence (X_3)	1	0.0071	.0071	0.08	.7750
Woody Border (X_4)	1	0.8298	.8298	9.58	.0024
$X_1 X_3$	1	0.1620	.1620	1.87	.1738
$X_2 X_4$	1	0.6172	.6172	7.13	.0086
Among transects	129	11.1754	0.0866		
Foreslope/Backslope	1	1.2719	1.2719	25.99	.0001
Within transects	135	6.6059	0.0489		

Table 5. Parameter estimates and standard errors for the blue jay data.

	Standard Errors				
Variable	$\hat{B}_{\text{IWM}}$	IWM	Robust	$\hat{B}_Q$	Standard error
Intercept	0.406	0.322	0.452	0.405	0.516
Bird type (X_1)	1.329	0.548	0.963	1.329	0.876
Acorn type (X_2)	-1.136	0.467	0.603	-1.136	0.530
$X_1 X_2$	-0.799	0.718	0.937	-0.799	0.867

Applied Statistics in Agriculture

Table 6. Parameter estimates and standard errors for the nest predation data.

	Standard errors				Standard
Variable	$\hat{B}_{IWM}$	IWM	Robust	$\hat{B}_Q$	error
Intercept	-2.220	0.232	0.301	-2.227	0.347
Road type (X_1)	0.278	0.195	0.290	0.270	0.302
Adj. habitat (X_2)	0.478	0.171	0.249	0.490	0.266
No Border Fence (X_3)	0.285	0.353	0.617	0.294	0.547
Woody Border (X_4)	1.107	0.225	0.361	1.138	0.348
$X_1 X_3$	-1.281	0.454	0.729	-1.277	0.704
$X_2 X_4$	-1.431	0.352	0.591	-1.417	0.545
Backslope (X_5)	0.785	0.137	0.155	0.784	0.156

Table 7. P-values for various methods of analysis of the blue jay data.

	Logistic regression			Linear	Linear mixed
	IWM	Robust	Quasi-likeihood	mixed	model using
Variable	(4.6)	(4.9)	(4.16)	model	$\arcsin(\sqrt{P})$
Bird type (X_1)	.015	.167	.129	.158	.123
Acorn type (X_2)	.015	.059	.032	.005	.008
Interaction $(X_1 X_2)$	.266	.394	.357	.507	.406

Table 8. P-values for various methods of analysis for the nest predation data.

	Logistic regression			Linear	Linear mixed
	IWM	Robust	Quasi-likelihood	mixed	model using
Variable	(4.6)	(4.9)	(4.16)	model	$\arcsin(\sqrt{P})$
Road type (X_1)	.154	.338	.371	.180	.273
Adj. habitat (X_2)	.005	.055	.065	.091	.071
No Border Fence (X_3)	.420	.645	.591	.775	.598
Woody Border (X_4)	.000	.002	.001	.002	.002
$X_1 X_3$	.005	.079	.070	.174	.091
$X_2 X_4$	.000	.015	.009	.009	.003
Backslope/Foreslope (X_5)	.000	.000	000	.000	.000