

Statistical Consulting Report for
Corey Shake (M.S. student)
Advisor: Dr. Chris Moorman
Fisheries and Wildlife Sciences (Dept. of FER)

April 2, 2009
By Fikret Isik
Statistical Liaison of CNR

I am studying nest survival of birds in patches of early-successional forest habitat. We collected data on nest survival over two years (2007 and 2008) in 12 different habitat patches. We are hypothesizing that multiple factors might influence nest survival. To evaluate which of these factors might have the greatest influence on nest survival and to determine the nature of the effect, we are using a specialized program that allows the user to build generalized linear models and compare them with Akaike's Information Criteria (AIC) model selection (called Program MARK). I understand the program well, and have no problem setting up and running the models. However, there is a statistical question/problem that I do not understand well.

We hypothesized that nest survival might vary between the two nesting seasons from which we collected data (hereafter referred to as "years"). We modeled the difference in nest survival (NS) between years by creating the following model:

$$\text{(Model 1) } NS = \beta_0 + \beta_1 x_1 + \varepsilon$$

where x_1 is coded in the design matrix as a dummy independent variable. For nests in 2007, $x_1 = 1$; and for nests in 2008, $x_1 = 0$. Thus, β_1 is an estimate of the differential expected nest survival in 2007 relative to 2008.

We also hypothesized that nest survival might be lower in patches where shrubs and saplings are taller and beginning to shade out understory vegetation that provides concealment for nests. Thus, we measured shrub/sapling height (in meters) for each patch in each year. We applied the shrub/sapling height of a given patch in a particular year as an independent variable to each nest in that patch in that year. As expected with tree growth, shrub/sapling height increased within each patch between 2007 and 2008 (average 1.6 m; range 0.7-2.2). We wanted to identify how shrub/sapling height influenced nest survival separately in each year while still considering the additional variation associated with years, so we considered the following model:

$$\text{(Model 2) } NS = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon$$

where $\beta_1 x_1$ is the same as in Model 1, and x_2 is the shrub/sapling height of nests in 2007 and x_3 is the shrub/sapling height of nest in 2008. For nests found in 2008, $x_2 = 0$; and

for nests found in 2007, $x_3 = 0$; so β_2 is an estimate of the incremental change in nest survival relative to a change in shrub/sapling height for nests in 2007 and β_3 an estimate of the incremental change in nest survival for each change in shrub/sapling height for nests in 2008.

My major concern here is multicollinearity: Is it legitimate to combine the year dummy variable with x_2 and x_3 ? I'm not sure how to assess multicollinearity between an essentially qualitative variable (year) and a quantitative variable that is measured only for nests in a given year (shrub/sapling height). Perhaps a simplified diagram of the design matrix for this model would be helpful:

| Intercept | x_1 | x_2 | x_3 |
|-----------|-------|-------|-------|
| 1 | 1 | 4.4 | 0 |
| 1 | 1 | 2.3 | 0 |
| 1 | 1 | 2.3 | 0 |
| 1 | 1 | 2.7 | 0 |
| 1 | 1 | 3.6 | 0 |
| 1 | 1 | 3.2 | 0 |
| 1 | 0 | 0 | 4.9 |
| 1 | 0 | 0 | 5.4 |
| 1 | 0 | 0 | 4.9 |
| 1 | 0 | 0 | 5.1 |
| 1 | 0 | 0 | 4.7 |
| 1 | 0 | 0 | 3.8 |

Do you think we have a problem with multicollinearity here? If so, do you have any suggestions about how we might avoid that problem while still getting at our research question, which is how does nest survival vary between years and in patches with different shrub/sapling heights?

COMMENTS:

As I understand, your response variable (nest survival) is a binary outcome (1=not survived, 0=survived).

If that is the case you are modeling the probability of survival (0 outcome). Most programs model 0 outcome by default. If that is not what you want you need define whether you are modeling 1 or 0 outcome in your model.

The predictor variables are Year (categorical) and Vegetation height (continuous) effects.

A simple logistic model would test the effects of predictor variables on the outcome as follows:

$$\log[P_i/(1-P_i)] = \text{logit}(p_i) = \alpha + \beta_1 \text{Year} + \beta_2 \text{VegHT} \quad (\text{model F1})$$

where P_i is the probability that the nest survives ($y=0$). β_1 is coefficient describing year effect and β_2 is the coefficient describing vegetation height effect.

The way you describe model 2 ($NS = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \epsilon$) does not look right and I do not think you can run such a model because you will not have data for X_2 in year 2008 or for X_3 in 2007. Most of logistic model algorithms use maximum likelihood methods to estimate parameters and in such a design matrix structure you described, the ML estimates are not estimable. In other words, you have complete separation of data. I am not sure if I understood your question about combining year dummy variable with X_2 and X_3 . My answer is 'you do not need to'.

I understand that you are concerned with the vegetation growth in 2008 compared to vegetation height in 2007.

You can account for the effects of vegetation heights in the model by adding an interaction term.

$$\log [P_i/(1-P_i)] = \alpha + \beta_1 \text{Year} + \beta_2 \text{VegHT} + \beta_3 \text{Year} * 2 \text{VegHT} \quad (\text{Model F2})$$

In logistic models with limited data the interaction terms may cause complete or quasi-complete separation of data which causes odd estimates of parameters. You need to be aware such problems. Does MARK software produce such messages?

I do not think you will have collinearity problem in the model. Collinearity may arise when you have many predictor variables and if they have a linear dependency. You have only 2 predictor variables. Collinearity will not bias your parameter estimates but it may produce large standard errors. A simple way to check the collinearity is to fit a model for each predictor variable separately first, e.g., $\log[P_i/(1-P_i)] = \alpha + \beta_1 \text{Year}$ and $\log[P_i/(1-P_i)] = \alpha + \beta_1 \text{VegHT}$. Then put them in the same model (Model F1). If Year effect is significant in simple logistic model but it is not in Model F1, then you can be suspicious. Same thing for VegHT only model.

I do not know if MARK program you are using has capability to check for collinearity for logistic regression (I suspect it does not). If it has, you can look at the Variance Inflation Factor or tolerance values for parameters. In SAS, Proc LOGISTIC does not have such capabilities but you can use linear regression procedure Proc REG to check collinearity.

One last thing; AIC fit statistic is useful when you are comparing different models. In your case the most important test statistic should be the chi-square statistics (Likelihood Ratio, Score or Wald). Such statistics tell you whether your model is significantly

different from the base model (intercept only). This is like ANOVA table in linear regression.