Appendix 6.A

The following text summarises the script files containing the R code for Chapter 6. The script files are available as individual .r files which can be opened in your favourite R code editor. The name of each file is prefaced by SCRIPT in **bold**, followed by a brief description. If data files are used the data files to go along with each script are prefaced by DATA and *in italics*. The names of separate files describing the metadata for each data file are prefaced by MD.

Scripts are described alphabetically and include: baccharis_seeds.r, binomial.r, normal_and_poisson_errors.r, normal_data_vs_normal_errors.r, pine_cones.r, poisson_r, tree_death.r

Data are also available including: baccharis_seeds.csv, pine_cones.csv and tree_death.csv

Buckley, Y.M. (in press) Generalised Linear Models, in Fox, G.A., Negrete-Yankelevich, S. and Sosa, V.J. Ecological Statistics (in press with Oxford University Press). 27 August 2014

SCRIPT: baccharis_seeds.r

DATA: baccharis_seeds.csv

MD: MD_baccharis_seeds.docx

This example contrasts an approach using transformed data with linear models & normal errors with a GLM approach using Poisson, Quasi-Poisson and Negative Binomial errors with log link functions.

Examples of plotting data and predicted models on response and linear predictor scales are given for multiple levels of a categorical variable.

This example demonstrates some of the "art" as well as the "science" of model development there isn't a simple script to follow, the answer will depend on what you want your model to do and who you are presenting your results to. This is the fun and frustration of ecological statistics!

Buckley, Y.M. (in press) Generalised Linear Models, in Fox, G.A., Negrete-Yankelevich, S. and Sosa, V.J. Ecological Statistics (in press with Oxford University Press). 27 August 2014

SCRIPT: binomial.r

The first part of the script shows how the variance of the binomial distribution changes with the mean and how this varies with the number of trials. The second part of the script shows the importance of the binomial denominator. If you have information on the denominator of a proportion and if this varies between samples it can have a large influence on the estimated proportion.

Buckley, Y.M. (in press) Generalised Linear Models, in Fox, G.A., Negrete-Yankelevich, S. and Sosa, V.J. Ecological Statistics (in press with Oxford University Press). 27 August 2014

SCRIPT: normal_and_poisson_errors.r

Here I simulate data and produce two figures (Fig.6.2 (a) & (b)), one showing data where the errors are normally distributed and where the variance is constant with regards to the mean and another where the errors are not necessarily normally distributed and are not constant, but the variance increases with the mean.

Thanks to Corey Merow for some example code (see Merow *et al.* 2014, Fig. 1).

C. Merow, J. P. Dahlgren, C. J. E. Metcalf, D. Z. Childs, M. E. K. Evans, E. Jongejans, S. Record, M. Rees, R. Salguero-Gomez, S. M. McMahon, Advancing population ecology with integral projection models: a practical guide. *Methods in Ecology and Evolution* **5**, 99-110 (2014)

Buckley, Y.M. (in press) Generalised Linear Models, in Fox, G.A., Negrete-Yankelevich, S. and Sosa, V.J. Ecological Statistics (in press with Oxford University Press). 27 August 2014

SCRIPT: normal_data_vs_normal_errors.r

Script for reproducing Fig. 6.1 Note that the random numbers are different each time so your figure will look different to the figure in the book chapter, run it a few times to see how different sets of random numbers look. Change the sample size to see how assessment of normality holds up at lower sample sizes (hint: can be very difficult to assess normality at low sample sizes).

Generates simulated y-values from the Normal distribution, with a mean equal to the fitted relationship and a constant standard deviation, i.e. the standard deviation does not change with the fitted values, it is constant. The errors of this relationship will be normally distributed, with constant variance and with a mean=0.

Buckley, Y.M. (in press) Generalised Linear Models, in Fox, G.A., Negrete-Yankelevich, S. and Sosa, V.J. Ecological Statistics (in press with Oxford University Press). 27 August 2014

SCRIPT: pine_cones.r

DATA: *pine_cones.csv*

MD: MD_pine_cones.docx

This script fits models to predict the number of cones based on plot (location), age and height, or basal area. See Coutts *et al.* (2012) for more detail on a larger dataset and its analysis. At the end of the script I give an example of how to fit and test an off-set in a GLM.

Height, age and basal area are all correlated so use only one in the analysis. Height was difficult to measure in the field, on some trees that were measured twice there were several meters difference between measures (both up and down), meaning we can have little confidence in any of the heights measured. Age was easy to measure (by counting annual growth nodes) and was measured in all plots so there is little missing data. Basal area was easy to measure but some of the data are missing. Overall it makes sense to use age as the covariate.

There are lots of 0's for cone counts, particularly at low ages. We could fit a zero-inflated model but we know why at least some of these 0's occur, they occur because the trees are too young. We would not expect a 2 year old tree to ever produce cones. 2009 was a particularly good coning year (mast year) and it appears as though trees younger than 13 did not produce cones that year. This also supports previous work in the same population that shows trees start to produce cones at around age 12 (Buckley et al. 2005). Exclude all trees that were 12 or younger because these 0's are not interesting and do not tell us anything except very young trees do not produce cones (which we already knew).

SIMPLE LINEAR REGRESSION

We might naively start with a classical linear regression of these data. I will use one year (2009) to avoid pseudo-replication due to multiple observations on the same individuals to analyse both years The variance increases with the mean fitted values and a histogram of residuals does not look normal. There is a "shot-gun" pattern in the residuals. Errors increase with the fitted values and poor performance in the Normal Q-Q plot indicates deviations from Normality.

TRANSFORMATION TO NORMALISE ERRORS

Log transforming count data is often recommended to improve model fit and normalise residuals, homogenise variance etc. As the log of zero cannot be taken your options are either to remove the zeros or add a small constant (1) to all values. Here I remove the zero counts as we could separately model the probability of producing cones process as a Binomial model with a binary (produce cones or not) response variable.

We use a model of only non-zero counts of cones*, i.e*. given a tree of a certain size produces cones, how many does it produce? This gives us some new problems, normality assumptions are still not appropriate and we appear to have over-corrected the dispersion problem with a reverse shotgun pattern now apparent. As an alternative approach to dropping the zeros we go back to the full 2009 data set and add a small amount, 1, to the observations before logging. The transformation has not

solved our assumption problems. Patterns are similar to the model where we removed zeros and log transformed.

GLM APPROACH

As an alternative to the transformation approach we could use a GLM. We start with a simple analysis by fitting a glm with a Poisson error distirbution (popular choice for count data). There is serious over dispersion in this model (residual deviance /degrees of freedom much greater than 1). Thus there is more variance in the error than is accounted for by the Poisson distirbution. It may be that the zeros in our data are over-inflating the variance, we could model these counts as a zero-inflated process or remove them and model the probability of producing cones as a separate process. Here I will remove the zeros and model the counts again as a Poisson process and see if this helps with correcting overdispersion.

Removing the zeros reduces the over dispersion from 190 to 173 but it is still a big problem. It could be our error distribution is wrong. We have some good evidence to suggest the error distribution might be negative binomial (Coutts et al. 2011), so we repeat the process with the negative binomial error distribution. The dispersion using negative binomial errors is now 48/40=1.2 which is close to 1. The AIC is much less than when we used the Poisson distribution indicating better model fit. The parameter estimates and conclusions change as now the intercept is no longer significantly different to zero whereas in the Poisson model with an equivalent data set it was positive and highly significantly different to zero.

If we'd really like to keep the zero cone counts in the data set we can try using the complete 2009 data set and see if the negative binomial model is still adequate. However, the model does not converge, so those zeros really are a problem. At this point you might be satisfied to model just the positive cone counts with the negative binomial model you might try to model the full 2009 data set with a zero-inflated model.

ALTERNATIVE APPROACH - using quasiPoisson

Instead of using a negative binomial model we might choose to use the quasi-Poisson distribution this allows the variance to increase faster than the mean by estimating a dispersion parameter. the dispersion parameter is not constrained to be equal to one (as for Poisson). The parameter estimates of the quasi Poisson model are identical to the Poisson model. Aas quasi models do not have a log-likelihood technically they don't have an AIC either. There are work-arounds for this if you search the R help lists. Also see the library `AICcmodavg`

The best model of cone counts for the 2009 positive cone counts (cones2009.pos) was a negative binomial model with a log link function.

FITTING AN OFF-SET

I might have a good *a priori* reason for assuming a certain value for the slope, for example there might be some ecological theory specifying a 1:1 relationship between cones & age, by using an offset I can test whether my model can be simplified in this way.

There is an implicit "x1" for the slope in the offset(age) part of the model, if I wanted to fit an offset with a slope=2 I would do "offset(age*2)". Note that if you compare the estimated slope for the offset with the no offset models you'll see exactly what's going on, the parameter estimate for the slope is re-calculated taking the offset slope into account. If the new parameter value for the slope was close to 0 and non-significant then you would have good evidence that the offset slope is a relationship which is supported. In the no offset model the p value on slope tests whether it is different to zero, in the offset model the slope parameter estimate is a difference from the offset slope so the p-value can be interpreted as testing to see if the estimated slope is different to the offset slope.

Buckley, Y.M. (in press) Generalised Linear Models, in Fox, G.A., Negrete-Yankelevich, S. and Sosa, V.J. Ecological Statistics (in press with Oxford University Press). 27 August 2014

SCRIPT: poisson.r

Reproduces Fig. 6.7

Shows the Poisson distribution for different means, the variance increases in direct proportion to the mean count (on x axis).

Buckley, Y.M. (in press) Generalised Linear Models, in Fox, G.A., Negrete-Yankelevich, S. and Sosa, V.J. Ecological Statistics (in press with Oxford University Press). 27 August 2014

SCRIPT: tree_death.r

DATA: tree_death.csv

MD: MD_tree_death.docx

These data are on mortality or survival of an Australian tree, *Eucalyptus melanophloia*. I show how to model survival/mortality binary data using a Binomial errors model. I also demonstrate a few different ways of visualising binary data and comparing binary data with model predictions. Data are from Dwyer, Fensham et al. 2010. Thanks to John Dwyer & Rod Fensham for providing the data.

Dwyer, J., R. Fensham, R. Fairfax, and Y. Buckley. 2010. Neighbourhood effects influence drought-induced mortality of savanna trees in Australia. Journal of Vegetation Science 21:573-585.

This script reproduces Fig. 6.4 & 6.5

The death model (m1) is a straightforward GLM with binomial errors and will be used throughout. The model is well behaved, not overdispersed and looks appropriate.

Fig. 6.4 demonstrates four different ways of visualising binary data. The raw observations are shown first and if there are a lot of data points it can be difficult to detect any pattern in the data as points are overlaid on eachother. By jittering the 0's and 1's by a small amount it can be easier to see what's going on. By averaging the data points in either evenly spaced categories or categories constructed to contain approximately the same number of data points the overall pattern of mortality in relation to size can be observed. Choosing an appropriate number of points to average or category size will determine how you view these data, you can play around with the number of categories to see how this changes your view. You could also use a box-plot here which would give more information about the spread of data points, not just the means.

We normally want to compare our model predictions with the underlying data to get a visual on the model fit. You can plot the modelled predictions on the linear predictor scale (so predictions will be linear) or on the scale of the response variable (proportions) and the predictions will appear non-linear, following the link function used (logit link here).