

# ***ENVIRONMENTAL SAMPLING AND ANALYSIS***

## ***WORKSHEET 6: COMPLEX LINEAR REGRESSION MODELS AND ANCOVA***

### ***QUESTION 1:***

Loyn (1987) was interested in what characteristics of habitat were related to the abundance and diversity of forest birds. He selected 56 forest patches in southeastern Victoria, Australia, and recorded the number of species and abundance of forest birds in each patch as two response variables. The predictor variables recorded for each patch included area (ha), the number of years since the patch was isolated by clearing (years), the distance to the nearest patch (km), the distance to the nearest larger patch (km), an index of stock grazing history from 1 (light) to 5 (heavy), and mean altitude (m).

- a) First, boxplot each of the variables. Any skewness or outliers?
  - b) Edit data file and log transform area, distance to the nearest patch and distance to the nearest larger patch.
  - c) Draw a scatterplot matrix (SPLOM) of all variables and also determine correlation matrix between all variables.
  - d) Fit a linear model relating bird abundance to all 6 predictor variables. Ignore all output except the tolerance values - any close to 0.1 indicating collinearity?
  - e) If collinearity is OK, refit the model and save the partial residuals. Plot residuals - any indication of unequal variances or outliers from fitted model? Examine the Cook's D statistics for each observation - any influential values?
  - f) Interpret multiple regression output. Which predictor variables are significantly related to forest bird abundance, holding the other predictors constant, i.e. which partial regression coefficients are significant?
  - g) Write out the full regression model, estimates of coefficients and their standard errors, and tests of each partial regression slope:
- 
- h) How much of the variation in forest bird abundance is explained by this set of 6 predictor variables?
  - i) Finally, draw a partial regression plot for  $\log_{10}$  area, the significant predictor. Remember, this plot shows the relationship between forest bird abundance and  $\log_{10}$  area, holding all other predictors constant. Note the strong positive relationship, indicating greater abundance in patches with larger areas, all other variables constant. Do the same partial regression plot for years isolated – what is your interpretation?

### ***QUESTION 2:***

Paruelo & Lauenroth (1996) analyzed the geographic distribution and the effects of climate variables on the relative abundance of a number of plant functional types (PFTs) including shrubs, forbs, succulents (*e.g.* cacti), C<sub>3</sub> grasses and C<sub>4</sub> grasses. They used data from 73 sites

across temperate central North America (see file PAREULO) and calculated the relative abundance of C<sub>3</sub> grasses (C3) at each site as a response variable. The predictor variables recorded for each site were LAT (latitude in centesimal degrees), LONG (longitude in centesimal degrees), MAP (mean annual precipitation in mm), MAT (mean annual temperature in °C), JJAMAP (proportion of MAP that fell in June, July and August) and DJFMAP (proportion of MAP that fell in December, January and February).

- First, boxplot each of the variables. Any skewness or outliers?
- Edit data file and square root transform C3.
- Draw a scatterplot matrix (SPLOM) of all variables and also determine correlation matrix between all variables.
- Fit a linear model relating  $\sqrt{C3}$  to all 6 predictor variables. Ignore all output except the tolerance values - any close to 0.1 indicating collinearity?

We obviously cannot easily incorporate all 6 predictors into the one model, because of the collinearity problem. Paruelo & Lauenroth (1996) separated the predictors into two groups for their analyses. One group included LAT and LONG and the other included MAP, MAT, JJAMAP and DJFMAP. We will focus on the relationship between the square root relative abundance of C<sub>3</sub> plants and latitude and longitude. This relationship will show the geographic pattern in abundance of C3 plants.

- Fit a linear model relating  $\sqrt{C3}$  to LAT and LONG and save the partial residuals. Plot residuals - any indication of unequal variances or outliers from model? Examine the Cook's D statistics for each observation - any influential values?
- Interpret multiple regression output. Which predictor variables are significantly related to  $\sqrt{C3}$  abundance, holding the other predictors constant, i.e. which partial regression coefficients are significant?
- Write out the full regression model:

$$\begin{array}{l} \sqrt{C3} \text{ abundance} \\ Y \end{array} = \text{intercept} + \text{slope}_1 \text{ LAT} + \text{slope}_2 \text{ LONG} \\ = \text{intercept} + \text{slope}_1 X_1 + \text{slope}_2 X_2$$

- How much of the variation in  $\sqrt{C3}$  abundance is explained by both predictor variables?
- Write the results out as though you were writing a research report. For example (cross-out which phrases do not apply and fill in gaps with your results):

A multiple linear regression of  $\sqrt{C3}$  abundance against latitude and longitude showed (choose correct option)

a significant positive / significant negative / no significant / partial regression slope for latitude

( $b = \underline{\hspace{1cm}}$ ,  $df = \underline{\hspace{1cm}}$ ,  $t = \underline{\hspace{1cm}}$ ,  $P = \underline{\hspace{1cm}}$ ) and

a significant positive / significant negative / no significant / partial regression slope for longitude

( $b = \underline{\hspace{1cm}}$ ,  $df = \underline{\hspace{1cm}}$ ,  $t = \underline{\hspace{1cm}}$ ,  $P = \underline{\hspace{1cm}}$ )

- Compare all three possible linear models (LAT, LONG, LAT & LONG). Based on adjusted  $r^2$  (highest) and  $MS_{\text{Residual}}$  (lowest), which is the best fitting model?

- k) Finally, draw a partial regression plot for LAT, the significant predictor. Remember, this plot shows the relationship between  $\sqrt{(C3)}$  abundance and LAT, holding LONG constant. Note the strong positive relationship, indicating greater abundance at higher latitudes (further north), for any given longitude. In contrast, the equivalent plot for LONG shows little relationship.

### **QUESTION 3:**

Partridge & Farquhar (1981) examined the effect of number and type of mating partners on longevity (response variable) of fruitflies. There was a single factor (partner type) with five treatments: one virgin female per day, eight virgin females per day, a control group with one newly inseminated female per day, a control group with eight newly inseminated females per day, a control group with no females. Also, the thorax length of each individual fly was recorded as a covariate. If thorax length explains some of the variation in longevity, then the test of the effect of partner type on longevity adjusted for thorax length will be more powerful.

- Write out the full linear model for this analysis, including the heterogeneity of slopes term.
- Plot log longevity against thorax length with OLS fitted linear regression model displayed for each treatment. Do the regression slopes appear very different, given the variability within each treatment?
- Test the null hypothesis that there was no difference between the log longevity – thorax length regression slopes between treatments. What are your conclusions?
- If the test in (c) was not significant, write out a modified linear model that assumes homogeneous slopes.
- Test the null hypothesis of no difference in adjusted mean log longevity between treatments. What are your conclusions? What is the estimate of the pooled regression slope?
- Plot a bar graph of adjusted means - what is the biological interpretation of these adjusted means?
- Finally, fit a model that ignores the covariate, i.e. a single factor ANOVA model testing log longevity against treatment. How does the  $MS_{\text{Residual}}$  compare to the fit of the ANCOVA model – has including the covariate thorax length given us a more powerful test of treatment?

### **QUESTION 4:**

In Chapter 5 of Q&K (2002), we describe the study by Peake & Quinn (1993) who looked at seasonal variation in the nature of species-area relationships. Their “islands” were clumps of mussels (*Brachidontes rostratus*) on a rocky shore at Phillip Island, southern Australia, and the species were the range of invertebrates that used these clumps as habitat during low (and possibly high) tides. We only used their data from one season in Chapter 5 but their full data set had four seasons (summer, autumn, winter and spring), with replicate mussel clumps of different sizes collected at each season and all invertebrates in each clump identified and counted and the area of each clump measured. The replication in each season varied between 20 and 25.

They wished to compare the relationships between number of species and clump area between the four seasons (dates). Species-area relationships are rarely linear (Q&K, Chapter 5) so Peake & Quinn (1993) linearised their species-area relationships by log transforming both variables. They then used analysis of covariance to compare slopes of linear regression models and adjusted

means between seasons. Log number of species was the response variable, season was the factor and clump log area was the covariate.

- (a) Write out the full linear model for this analysis, including the heterogeneity of slopes term.
- (b) Plot log species number against log area with OLS fitted linear regression model displayed for each date. Do the regression slopes appear very different, given the variability within each date?
- (c) Test the null hypothesis that there was no difference between the log species – log area regression slopes between dates. What are your conclusions?
- (d) If the test in (c) was not significant, write out a modified linear model that assumes homogeneous slopes.
- (e) Test the null hypothesis of no difference in adjusted mean log number of species between seasons. What are your conclusions?
- (f) What is the biological interpretation of these adjusted means?
- (g) Finally, fit a model that ignores the covariate, i.e. a single factor ANOVA model testing log species number against date. How does the  $MS_{\text{Residual}}$  compare to the fit of the ANCOVA model – has including the covariate given us a more powerful test of date?

#### ***QUESTION 5:***

This question uses the mussel clump data from Peake & Quinn (1993) again. Our predictor variable is still clump area but the response variable is now species number.

- a) Draw a scatterplot of species number against clump area with a Lowess smoother and bordered boxplots. Any evidence of nonlinearity?

We could transform clump area to try and achieve linearity but let's try fitting a polynomial model.

- b) Draw a scatterplot of species number against clump area with a quadratic polynomial smoother. Does this seem a reasonable fit?
- c) Fit a polynomial model relating species number to clump area and  $(\text{clump area})^2$ . Compare the fit of this full model to the reduced model (simple regression) with just clump area - does adding the quadratic term significantly improve the fit?