

# Quinn & Keough (2002) *Experimental Design and Data Analysis for Biologists*

## Chapter 6 Multiple and complex regression

### Question 1:

- What assumptions must be checked before interpreting the least squares fit of a multiple linear regression model?
- What is the interpretation of a partial regression coefficient in a multiple linear regression model?
- What is the interpretation of standardised partial regression coefficients – how are they different from the usual coefficients?
- Define collinearity and summarise its effects on the estimation of multiple regression models.

### Question 2:

Olesen & Jordano (2002) reviewed the available literature on plant-pollinator mutualistic networks to test how interactions between plants and pollinators vary with latitude, elevation and insularity. They examined 29 publications that summarised plant-pollinator networks and for each site/habitat combination in each paper they recorded the following measures: latitude (degrees N or S), altitude (m; an average was used if the site covered an altitudinal range), number of animal species ( $A$ ), number of plant species ( $P$ ), the size of the plant-pollinator network ( $M = A \times P$ ), number of recorded interactions ( $I$ ), connectance ( $C = 100 \times I/M$ ) and insularity (0.5 if site was island and 1.0 if site was mainland). They also derived two additional variables: mean number of interactions across animal species ( $L_m = I/A$ ) and mean number of interactions across plant species ( $L_n = I/P$ ). Olesen & Jordano (2002) were particularly interested in how various characteristics of these networks varied with latitude, altitude, and insularity.

- Draw a scatterplot matrix with boxplots to show the distribution of each variable ( $C$ ,  $I$ ,  $L_m$ ,  $L_n$  are the response variables;  $M$ , latitude, altitude are the predictor variables) and their relationships. Is there evidence of non-normality for any variable or nonlinearity in the relationships?

To improve normality, Olesen & Jordano (2002) transformed  $C$  to arcsin square root ( $100 \times C$ ). They first fitted simple regression models between each of the response variables and network size.

- Fit these four linear models and examine the residuals – any obvious problems? Any outliers or influential observations (use Cook's statistic) identified?

It should be clear from these exploratory analyses that transformations are likely to result in models that better fit the data so, like Olesen & Jordano (2002), we will transform  $I$ ,  $L_m$ ,  $L_n$  and  $M$  to natural logs.

- Refit the linear models and examine the residuals and diagnostics, to confirm that the models based on transformed variables are more appropriate. What statistical and biological conclusions can you draw from your analyses?

Olesen & Jordano (2002) next focused on the relationships between the four network characteristics and each of latitude, altitude, and insularity (a dummy variable). They wished

to examine these relationships correcting for network size so they also included  $\ln(M)$  in each of these regression models.

- d) Fit a multiple linear regression with  $\text{asn}(C)$  as the response variable and  $\ln(M)$  and latitude as predictors. Examine the residuals and diagnostics from this model. Also check for collinearity between these two predictors.

It appears that there are no obvious problems with any of the assumptions for this model.

- e) Now interpret the results of this regression analysis. What do the two regression coefficients represent? What conclusions would you draw from the hypothesis tests? How much of the variation in the response variable was explained by this model?
- f) If any of the predictors were statistically significant, use a partial regression plot to show the relationship between the response variable and that predictor, holding the other predictor constant.
- g) Now fit a model relating  $\text{asn}(C)$  to  $\ln(M)$  and altitude. Examine the residuals and diagnostics from this model. Also check for collinearity between these two predictors. Olesen & Jordano (2002) are somewhat confusing as to whether altitude was transformed (last paragraph of p.2417 suggests it was whereas it clearly isn't for the simple regression models presented in Fig. 2). The residual plots do not suggest a transformation is necessary. Interpret the results of this regression analysis. What conclusions would you draw? Again, represent significant relationships using partial regression plots.
- h) Finally, fit a model relating  $\text{asn}(C)$  to  $\ln(M)$  and insularity. Remember that insularity is a dummy variable. How do you interpret the regression coefficient for insularity? Plot the mean transformed connectivity, adjusted for  $\ln(M)$ , for each category of insularity (see Q&K sec 6.1.14 and Chapter 12 for explanation and interpretation of adjusted means).

Complete the remaining analyses for the other response variables. Compare your results to Table 2 of Olesen & Jordano (2002). Note that they present standardised regression coefficients and, for each response variable, the coefficients for  $M$  are from simple regression models and the coefficients for latitude, altitude and insularity are for multiple regression models including each of these variables separately and  $M$ . There may also be discrepancies between your analyses and their results. This may be due to which value they used when altitude was a range, software differences and rounding errors. They also used randomisation tests rather than traditional  $t$  or  $F$  tests.

### **Question 3:**

Dodson *et al.* (2000) collated data from numerous sources on primary productivity and species richness (of various taxa) for 33 well-studied lakes in North America. They were particularly interested in relationships between productivity ( $\text{g C}\cdot\text{m}^{-2}\cdot\text{yr}^{-1}$ ), lake size (ha) and species richness for each taxonomic group. We will focus on three of the groups for which data were complete: cladocerans, copepods and fish. Our aim is to estimate a linear model with species richness of each group as a response variable and lake size and primary productivity as predictor variables. We will use number of fish species first.

- a) Draw a scatterplot matrix with boxplots to show the distribution of each variable (no. fish species, lake size, primary productivity) and their relationships. Is there evidence of non-normality for any variable or nonlinearity in the relationships?
- b) Now fit a multiple linear regression model to these data, with number of fish species as the response variable and lake size and primary productivity as predictors. Examine the residuals from this model – any obvious problems? Any outliers or influential observations (use Cook's statistic) identified?
- c) Check for collinearity between primary productivity and lake size – any evidence of a correlation between these two predictors?

These assessments of the data and adequacy of the model probably convinced you that transformation of the variables might be appropriate. Dodson *et al.* used logs and we will do the same. Transform variables to  $\log_{10}$  (or  $\log_{10}+1$  if zeros are involved) as necessary and use transformed variables in subsequent analyses.

- d) Refit the multiple regression model using transformed variables. Interpret and summarise your conclusions, including indicating which of the two predictors was the strongest influence on species richness?
- e) Draw partial regression plots to summarise the relationship between species richness and each predictor, holding the other constant – any evidence of nonlinearity?

Follow the same procedure using number of species of cladocerans and of copepods, rather than fish, as the response variables. What conclusions can you draw from these analyses? Compare your results with those in Dodson *et al.* (2000: their Fig. 2).

Now, using the same dataset but just focussing on fish (log transformed), we will fit a slightly more complex model with an interaction term, i.e. a model that includes productivity, lake size and the interaction between productivity and lake size. Note that the predictors should probably be transformed as above.

- f) What null hypotheses are being tested by this model? In particular, what biological hypothesis is being tested by the interaction term?
- g) Fit the model based on log-transformed data. Is there any evidence of collinearity between the predictors, including the interaction term?
- h) Now refit the model based on centred predictors. Has collinearity been reduced? What conclusions would you draw from the model fit?

#### Question 4:

Polis *et al.* (1998) were interested in the factors that control the numbers of spiders on islands in the Gulf of California, Mexico. They sampled 17 islands and recorded the density of spiders (response variable) and two predictor variables: the ratio of island perimeter to island area (P/A) as a measure of relative input of resources from the ocean, and the density of scorpions, predators of the spiders. They fitted a multiple regression analysis relating the response variable (log transformed spider density) to the two predictors with the following results:

	Coefficient	Standard error	<i>t</i>	<i>P</i>
Intercept	1.609	1.799	0.894	0.135
P/A ratio	0.006	0.004	1.654	0.073
Scorpions	-0.746	0.060	-12.387	<0.001

The  $r^2$  from the regression was 0.73.

- (a) Besides normality and homogeneity of variance of the response variable, what other assumptions are important before we can assess each of the tests on regression slopes?
- (b) How would you interpret the regression slope for P/A ratio in this analysis?
- (c) Write out the complete regression equation.
- (d) What is the  $r^2$  value telling us?

### Question 5:

Germaine *et al.* (1998) were interested in the relationships among breeding birds, habitat, and residential development in Tuscon, Arizona, USA. They sampled 334 census plots and recorded from each plot the non-native bird species richness (no. species) as a response variable and four predictor variables: house density, % of plot that was paved/graded, % of plants that were urban-exotic and % of plants that were upland Sonoran vegetation (a vegetation classification type). They fitted a multiple regression analysis relating the response variable to the 4 predictors, with the following results:

	Coefficient	Stand. error	<i>t</i>	<i>P</i>
Intercept	0.545	0.126	4.317	<0.001
House density	0.196	0.044	4.425	<0.001
% paved/graded	0.026	0.005	5.596	<0.001
% urban exotic	0.009	0.003	3.138	0.002
% upland Sonoran	-0.004	0.002	-2.672	0.007

- Besides normality and homogeneity of variance of the response variable, and independence of observations, what other crucial assumption is necessary before we can assess each of the regression slopes?
- How would you interpret the regression slope for house density in this multiple regression analysis?
- What is the intercept measuring?

### Question 6:

Robinson *et al.* (2000) examined morphological divergence of pumpkinseed sunfish (*Lepomis gibbosus*) in lakes of the Adirondack State Park in New York State, USA. They observed differences in pumpkinseed morphology between lakes with and without the closely related bluegill sunfish (*Lepomis macrochirus*) and also divergence between pumpkinseeds in open-water and shallow-water habitats in lakes with only pumpkinseeds. They explored this latter pattern in more detail by sampling 22 lakes with only pumpkinseeds and calculating an index of body shape divergence between open- and shallow-water fish. They modelled this divergence index against three predictor variables (number of planktivore taxa, lake size and proportion of open water habitat). They also included the three two-way interactions in their multiple regression model. The results of the analysis were (see their Table 3):

Predictor	Coefficient	SE	Standardised coefficient	<i>t</i>	<i>P</i>
Constant	18.8	3.7	0	5.0	0.0001
Planktivores (P)	-3.2	1.4	-2.3	-2.2	0.043
Lake size (LS)	-0.036	0.019	-2.3	-1.9	0.073
Prop open water habitat (OW)	-11.6	4.2	-1.3	-2.8	0.014
P x LS	0.0043	0.0021	0.90	2.0	0.006
P x OW	1.88	1.7	1.2	1.1	0.27
LS x OW	0.039	0.022	2.3	1.8	0.095

$$F = 2.9, P = 0.043, r^2 = 0.54$$

- What are the assumptions underlying this multiple regression analysis?
- Which assumption is particularly difficult to meet because of the interaction terms? What are the options for dealing with this problem?

- c) What is the biological interpretation of the statistically significant P x LS interaction term? How does this affect interpretation of the main effects of number of planktivore taxa, lake size and proportion of open water habitat?
- d) Why was the three way interaction not included (see their paper)? If it had been included, how would a significant result have been interpreted?

### Question 7:

Welch *et al.* (1997) studied the intake rate of fruits by black and grizzly bears held in a specialised bear research facility in Washington State, USA. They were particularly interested in factors that might constrain the rate at which bears can consume different fruits, including huckleberry, serviceberry, and different sized grapes. The main constraining factor investigated was the density of berries (berries.m<sup>-3</sup>) and the response variables included intake rate (g.min<sup>-1</sup>), bite rate (bites.min<sup>-1</sup>) and bite size (berries.bite<sup>-1</sup>). For the first two response variables, the relationship was clearly not linear so a non-linear model commonly used in foraging studies was applied:

$$Y = \frac{\theta_1 \sqrt{X}}{(\theta_2 + \sqrt{X})}$$

where Y is the response variable (e.g. intake rate), X is berry density, and  $\theta_1$

and  $\theta_2$  are parameters. Referring to the paper by Welch *et al.* (1997), especially their Fig. 1, answer the following questions:

- a) What method was used to fit the non-linear models?
- b) How are the  $r^2$  values provided different from  $r^2$  values from simple linear regression?
- c) What is the biological interpretation of the parameters  $\theta_1$  and  $\theta_2$ ?
- d) If you wished to linearise these relationships, what transformation might have been appropriate?

### Question 8:

Lawrence & Ripple (2000) investigated revegetation of Mount St. Helens in Washington State, USA, based on data for 15 years after the massive eruption in 1980. Their data were based on 5000 randomly selected points from Landsat images. For each point, they recorded four response variables (e.g. number of years to reach 10% cover, maximum estimated vegetation cover, time-integrated cover over the study period etc.) and nine predictor variables [seven continuous variables: distance from crater, tephra thickness, distance from surviving forests, slope gradient, slope curvature, elevation, aspect, and two categorical variables: disturbance type (see their Table 2), blast exposure (directly exposed and unexposed)]. They used regression tree analysis to examine the relationship between each response variable and the nine predictor variables. Referring to their paper, especially their Fig. 2 and Table 3, answer the following questions:

- a) Which method of model-fitting (least squares or maximum likelihood) was used for the regression tree analyses?
- b) Selecting one of the three trees in Fig. 2, describe the interpretation at each split and the two values provided at each terminal node.
- c) Lawrence & Ripple (2000) also used a cross-validation to prune their regression tree. Describe the method they used and how the final trees were chosen.
- d) Table 3 of their paper provides  $r^2$  measures of fit for the original data plus an independent set of data. How were the independent data used to generate these measures of fit?
- e) What alternative analyses could have been used to describe the relationship between each response variable and the set of predictor variables?

### **Question 9:**

As described in the questions for Chapter 5, Ollinger *et al.* (2002) measured various plant and soil characteristics (related to C and N processing) for 30 plots in the White Mountain National Forest, New Hampshire, USA. Part of their analyses was to estimate the relationship between N mineralisation and soil C:N ratio and between nitrification and soil C:N ratio.

- a) Draw scatterplots for these two relationships and include a Lowess smoother. Is there any evidence for a nonlinear relationship?
- b) Fit a linear model relation N mineralisation and nitrification to soil C:N ratio and examine the residuals. Again, is a nonlinear relationship suggested?
- c) Now fit a nonlinear, exponential, model of the form  $Y = \alpha \cdot \exp(\beta X)$  to each data set. What are the parameter estimates (cf. Ollinger *et al.* p.343)? How much of the variation is explained by each model?

### **References:**

Dodson, S.I., Arnott, S.E. & Cottingham, K.L. (2000) The relationship in lake communities between primary productivity and species richness. *Ecology* **81**: 2662-2679.

Germaine, S.S., Rosenstock, S.S., Schweinsburg, R.E. & Robinson, W.S. (1998) Relationships among breeding birds, habitat, and residential development in greater Tuscon, Arizona. *Ecological Applications* **8**: 680-691.

Lawrence, R.L. & Ripple, W.J. (2000) Fifteen years of revegetation of Mount St. Helens: a landscape analysis. *Ecology* **81**: 2742-2752.

Olesen, J.M. & Jordano, P. (2002) Geographic patterns in plant-pollinator mutualistic networks. *Ecology* **83**: 2416-2424.

Ollinger, S.V., Smith, M.L., Martin, M.E., Hallett, R.A., Goodale, C.L. & Aber, J.D. (2002) Regional variation in foliar chemistry and N cycling among forests of diverse history and composition. *Ecology* **83**: 339-355.

Polis, G.A., Hurd, S.D., Jackson, C.T. & Sanchez-Pinero, F. (1998) Multifactor population limitation: variable spatial and temporal control of spiders on Gulf of California islands. *Ecology* **79**: 490-502.

Robinson, B.W., Wilson, D.S., Margosian, A.S. (2000) A plurastic analysis of character release in pumpkinseed sunfish (*Lepomis gibbosus*). *Ecology* **81**: 2799-2812.

Welch, C.A., Keay, J., Kendall, K.C. & Robbins, C.T. (1997) Constraints on frugivory by bears. *Ecology* **78**: 1105-1119.