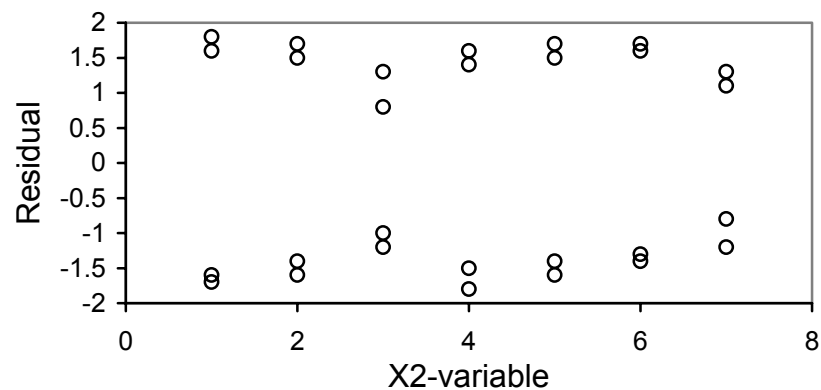
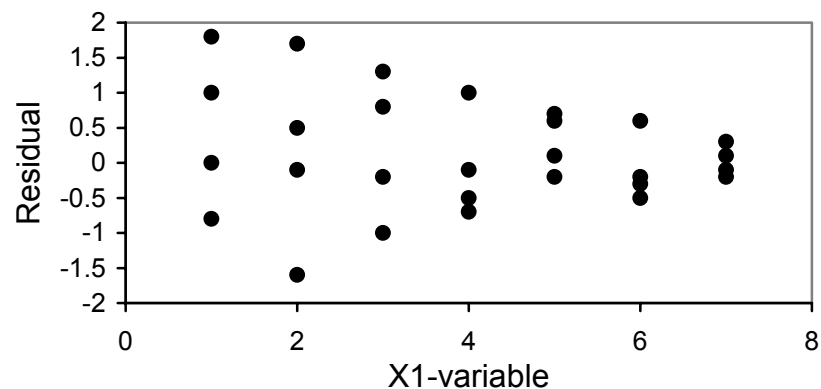


# Quinn & Keough (2002) *Experimental Design and Data Analysis for Biologists*

## Chapter 5 Correlation and regression

### Question 1:

- What assumptions must be checked before interpreting the least squares fit of a simple linear regression model?
- Explain what is represented by the intercept and slope in a simple linear regression model?
- In simple linear regression, how do we identify outliers?
- Interpret the following two residual plots from simple linear regression.



### Question 2:

Removal of unique combinations of frog toes is commonly used by amphibian researchers to identify individuals in studies of population ecology. Because of concern that this method (“toe clipping”) may harm the study organisms, researchers in three separate studies investigated relationships between the number of toes clipped and the probability of recapturing individuals in three different populations of the great barred frog. In these studies, frogs were captured over a period of one month, and a unique combination of digits were removed from previously unmarked individuals. The incidence of recapturing these marked individuals 12 months later was recorded. The resulting data were analyzed using linear

regression. A decline in the recapture rate with an increasing number of toes clipped would indicate a deleterious effect of toe clipping on the species. The data from the three studies are given below.

*Study 1*

No. toes clipped ( <i>X</i> variable)	No. frogs marked	No. recaptured at least once	Proportion recaptured ( <i>Y</i> variable)
2	9	3	0.333
3	20	7	0.350
4	23	7	0.304
5	4	3	0.750

*Study 2*

No. toes clipped ( <i>X</i> variable)	No. frogs marked	No. recaptured at least once	Proportion recaptured ( <i>Y</i> variable)
2	11	9	0.818
3	203	154	0.759
4	474	342	0.722
5	309	212	0.686
6	201	161	0.801
7	108	67	0.620

*Study 3*

No. toes clipped ( <i>X</i> variable)	No. frogs marked	No. recaptured at least once	Proportion recaptured ( <i>Y</i> variable)
2	8	6	0.750
3	22	17	0.773
4	136	98	0.721
5	202	139	0.688
6	21	14	0.667
7	3	2	0.667

The results of the linear regression analyses are given below. The *P* value for the test of the null hypothesis of zero slope is based on a two-tailed test.

*Study 1*

	Coefficient	S.E.	<i>P</i>
Intercept	0.013	0.288	
Slope	0.120	0.078	0.264

*Study 2*

	Coefficient	S.E.	<i>P</i>
Intercept	0.898	0.015	
Slope	-0.043	0.004	0.009

*Study 3*

	Coefficient	S.E.	<i>P</i>
Intercept	0.816	0.040	
Slope	-0.024	0.011	0.163

- What can you conclude about the effect of toe clipping on this species of frog? Discuss whether the results from the different studies are consistent.
- Discuss the different factors that would influence the relative power of the analyses used in these three studies?

- c) Which assumptions of linear regression analysis are likely to be violated in these three studies, and how might violations of these assumptions influence the results obtained?

### **Question 3:**

Schaefer & Mahoney (2001) were interested in functional explanations for horns and antlers on female ungulates, particularly caribou (*Rangifer tarandus*) in Canada. They surveyed 15 caribou herds across a 1000 km gradient in Newfoundland and Labrador and recorded the % of females with antlers, the population density of caribou (no. animals per km<sup>2</sup>) and various measures of snowfall (e.g. total annual snowfall, snow depth at the end of March). Our interest is whether the % of females with antlers is correlated with population density.

- a) Draw a scatterplot % of females with antlers against population density, with boxplots included for each variable. Is there evidence of non-normality for either variable or nonlinearity in the relationship?
- b) Schaefer & Mahoney (2001) considered both variables to have non-normal distributions and transformed population density to logs and % females with antlers to arcsin proportions. Apply these transformations. Is normality improved for each variable? Does the nature of the relationship seem to have changed? Does the use of a parametric correlation coefficient seem appropriate now?
- c) Test the null hypothesis that the correlation between asin (proportion of females with antlers) and log (population density) equals zero. What are your conclusions? What is the estimated correlation coefficient?
- d) If normality of the variables and linearity of the relationship can't be assumed, nonparametric correlation analysis might be appropriate. Use the untransformed data to test the null hypothesis of no monotonic relationship between the two variables, using Spearman's rank correlation. How do your conclusions compare to the test on transformed variables with Pearson's parametric correlation?
- e) Finally, what is one biological explanation for the relationship between % of females with antlers and population density?

### **Question 4:**

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 (g C.m<sup>-2</sup>.yr<sup>-1</sup>), lake size (ha) and species richness for each taxonomic group. We will use their data to examine bivariate relationships between variables and also to fit linear models where one variable is clearly a response and the other a predictor (see Question 6). In this question, we will test whether primary productivity is correlated with lake size.

- a) Draw a scatterplot of primary productivity against lake area, with boxplots included for each variable. Is there evidence of non-normality for either variable or nonlinearity in the relationship?
- b) Transformations seem appropriate (Dodson *et al.* used logs) so redraw the scatterplot and the boxplots based on transformed data (see their Fig. 1). Does the use of a parametric correlation coefficient seem appropriate now?
- c) Test the null hypothesis that the correlation between log (primary productivity) and log (area) equals zero. What are your conclusions? What is the estimated correlation coefficient?
- d) If normality of the variables and linearity of the relationship can't be assumed, nonparametric correlation analysis might be appropriate. Use the untransformed data

to test the null hypothesis of no monotonic relationship between the two variables, using Spearman's rank correlation. How do your conclusions compare to the test on log transformed variables with Pearson's parametric correlation?

### **Question 5:**

Ollinger *et al.* (2002) studied regional variation in the relationships between canopy chemistry, soil C:N ratios and soil N transformations in a forest in New Hampshire, USA. They sampled 30 plots in the forest, each plot with a different history of logging and a combination of one or more tree species (e.g. American beech, balsam fir, eastern hemlock etc.). For each plot, they measured the % lignin and % N from upper and mid-canopy foliage from trees of each species and combined these into single values for N and lignin. They also measured C:N ratios of replicate soil cores from each plot, as well as rates of N mineralization and nitrification and soil pH. Our interest is in the nature of the relationship between soil C:N ratios and foliar lignin:N ratios in the canopy. The latter can be measured across large spatial scales using various new remote sensing techniques (e.g. AVIRIS: Airborne Visible and InfraRed Imaging Spectrometer), whereas soil characteristics require time-consuming on-ground sampling, so predicting soil C:N from canopy lignin:N would be very useful.

- a) Draw a scatterplot of soil C:N ratio against foliar lignin:N ratio, with boxplots included for each variable. Also fit a Lowess smoothing function to the data – using default smoothing parameter (tension) should be fine. Is there evidence of non-normality for either variable or nonlinearity in the relationship?
- b) Now fit a linear regression model to these data, with soil C:N ratio as the response variable and foliar lignin:N ratio as the predictor. Examine the residuals from this model – any obvious problems? Any outliers or influential observations identified?

These assessments of the data and adequacy of the model probably convinced you that transformations are not necessary and the linear model is an adequate fit.

- c) Summarise the results of a Model 1 linear regression analysis of soil C:N as the response variable and canopy lignin:N as the predictor variable. Include the full regression model with confidence intervals on the parameter estimates, the measures of 'explained' variance and the test of the null hypothesis of zero slope.
- d) What biological conclusions would you draw from this analysis? Can you predict soil C:N from canopy lignin:N?

### **Question 6:**

A positive relationship between number of species and area sampled (the species-area relationship) has long been considered one of ecology's few "laws". Using the data from Dodson *et al.* (2000: - see question 4), we will examine the relationships between the number of species of fish and copepods and lake area.

- a) Draw a scatterplot of number of fish species against lake area, with boxplots included for each variable. Also fit a Lowess smoothing function to the data – using default smoothing parameter (tension) should be fine. Is there evidence of non-normality for either variable or nonlinearity in the relationship?
- b) Now fit a linear regression model to these data, with number of fish species as the response variable and lake area as the predictor. Examine the residuals from this model – any obvious problems? Any outliers or influential observations identified?

These assessments of the data and adequacy of the model probably convinced you that transformation of both variables might be appropriate. Dodson *et al.* used logs and we will do the same.

- c) Redo the scatterplot and Lowess smoother using transformed variables – do the boxplots appear more symmetrical and a linear relationship more plausible?
- d) Refit a linear regression model using transformed data and examine the residuals. Are they more reasonable than they were with untransformed variables?
- e) Write out the results from your linear regression analysis, including the full model with confidence intervals on the parameter estimates, the measures of ‘explained’ variance and the test of the null hypothesis of zero slope.
- f) What conclusions would you draw from the regression analysis using transformed variables?

Now follow the same procedure using number of species of copepods, rather than fish, as the response variable. What conclusions can you draw from the analysis? Compare your results with those in Dodson *et al.* (2000: their Fig. 2).

### **Question 7:**

It is often the case with bivariate relationships in biology that both variables are random but our research question requires us to determine more than just a simple correlation coefficient. While we can, as we did in questions 5 and 6, simply use a standard Model 1 OLS regression model and treat the predictor variable as fixed, we might prefer to fit a relationship that takes the random nature of both variables into account. Le Boeuf *et al.* (2000) examined the foraging behaviour of northern elephant seals (*Mirounga angustirostris*) that breed along the west coast of Mexico and the USA. They focused on the biannual migrations of seals from Año Nuevo, California, over three years (1995-1997). They attached platform satellite transmitter terminals (PTTs) to 27 male seals and recorded, for each seal, the distance (km) to its focal foraging area (FFA), the time (d) it took to travel to and from the FFA and the duration (d) spent at the FFA. Five seals had problems with their PTTs so there were 22 seals with complete data. We will focus on two relationships – first, the time taken to travel to the FFA and distance to FFA and second, the duration spent at the FFA and distance to FFA (see their Fig 3).

- a) Draw scatterplots of duration to FFA against distance from rookery, and duration on FFA against distance from rookery, with boxplots included for each variable. Also fit a Lowess smoothing function to the data – using default smoothing parameter (tension) should be fine. Is there evidence of non-normality for either variable or nonlinearity in the relationship?
- b) Now fit linear regression models to these two data sets, with duration to FFA as the response variable and distance from rookery as the predictor in the first and duration on FFA as the response variable and distance from rookery as the predictor in the second. Examine the residuals from each model – any obvious problems? Any outliers or influential observations identified?

These assessments of the data and adequacy of the model should have convinced you that transformations are not necessary and the linear models are adequate fits.

- c) Summarise the results of the regression analyses including both models with confidence intervals on the parameter estimates, the measures of ‘explained’ variance and the tests of the null hypothesis of zero slope. Do your results agree with those presented in Fig. 3 of Le Boeuf *et al.* (2000)?

These regression models ignore the random nature of the predictor variable (distance to FFA) so a Model II regression might be more appropriate for estimating the true value of the slope of the linear regression model. As discussed in Q&K2002, reduced major axis (RMA) regression is usually the most suitable approach.

- d) Estimate the RMA regression model for both data sets, either using the equations in Q&K2002 (Section 5.3.14) or using appropriate software (e.g. Prof. P. Legendre's software at <http://www.fas.umontreal.ca/biol/legendre/indexEnglish.html>).
- e) How do the Model I and Model II estimates of the regression slopes differ for the two relationships? Are these differences related to the correlation coefficients (or  $r^2$  values) for the two data sets?

### **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.

Le Boeuf, B.J., Crocker, D.E., Costa, D.P., Blackwell, S.B., Webb, P.M. & Houser, D.S. (2000) Foraging ecology of northern elephant seals. *Ecological Monographs* **70**: 353-382.

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.

Schaefer, J.A. & Mahoney, S.P. (2001) Antlers on female caribou: biogeography of the bones of contention. *Ecology* **82**: 3556-3560.