Worksheet 3 - Regression and linear models

Linear regression references

- Quinn & Keough (2002) - Chpt 5

Question 1 - Simple linear regression

Here is an example from Fowler, Cohen and Parvis (1998). An agriculturalist was interested in the effects of fertilizer load on the yield of grass. Grass seed was sown uniformly over an area and different quantities of commercial fertilizer were applied to each of ten 1 m^2 randomly located plots. Two months later the grass from each plot was harvested, dried and weighed. The data are in the file fertilizer.csv.

Format of fertilizer.csv data files

FERTILIZER	YIELD
25	84
50	80
75	90
100	154
125	148
...	...

FERTILIZER	Mass of fertilizer ($g.m^{-2}$) - Predictor variable
YIELD	Yield of grass ($g.m^{-2}$) - Response variable

Open the fertilizer data file.

Q1-1. List the following

a. The biological hypothesis of interest

b. The biological null hypothesis of interest

c. The statistical null hypothesis of interest

Q1-2. Test the assumptions of simple linear regression using a **scatterplot** of YIELD against FERTILIZER. Add boxplots for each variable to the margins and fit a lowess smoother through the data. Is there any evidence of violations of the simple linear regression assumptions? (Y or N) ☐

If there is no evidence that the assumptions of simple linear regression have been violated, **fit the linear model** YIELD = intercept + (SLOPE * FERTILIZER). At this stage ignore any output.

Q1-3. **Examine the regression diagnostics** (particularly the **residual plot**). Does the residual plot indicate any potential problems with the data? (Y or N) ☐

Q1-4. If there is no evidence that any of the assumptions have been violated, **examine the regression output**. Identify and interpret the following;

- **a.** sample y-intercept ☐
- **b.** sample slope ☐
- **c.** t value for main H_0 ☐
- **d.** P-value for main H_0 ☐
- **e.** R^2 value ☐

Q1-5. What conclusions (statistical and biological) would you draw from the analysis?

Q1-6. Significant simple linear regression outcomes are usually accompanied by a scatterpoint that summarizes the relationship between the two population. Construct a **scatterplot** **without a smoother or marginal boxplots**.

Question 2 - Simple linear regression

Christensen et al. (1996) studied the relationships between coarse woody debris (CWD) and, shoreline vegetation and lake development in a sample of 16 lakes. They defined CWD as debris greater than 5cm in diameter and recorded, for a number of plots on each lake, the basal area ($m^2.km^{-1}$) of CWD in the nearshore water, and the density (no.km^{-1}) of riparian trees along the shore. The data are in the file christ.csv and the relevant variables are the response variable, CWDBASAL (coarse woody debris basal area, m^2.km-1), and the predictor variable, RIPDENS (riparian tree density, trees.km^{-1}).

Format of christ.csv data files

LAKE	RIPDENS	CWDBASAL
Bay	1270	121
Bergner	1210	41
Crampton	1800	183
Long	1875	130
Roach	1300	127
...	...	...

LAKE	Name of the North American freshwater lake from which the observations were collected
RIPDENS	Density of riparian trees (trees.km^{-1}) Predictor variable
CWDBASAL	Course woody debris basal area (m^{2}.km^{-1}) Response variable

Open the christ data file.

Q2-1. List the following

a. The biological hypothesis of interest

b. The biological null hypothesis of interest

c. The statistical null hypothesis of interest

Q2-2. In the table below, list the assumptions of simple linear regression along with how violations of each assumption are diagnosed and/or the risks of violations are minimized.

Assumption	Diagnostic/Risk Minimization
I.	
II.	
III.	
IV.	

Q2-3. **Draw a scatterplot** of CWDBASAL against RIPDENS. This should include boxplots for each variable to the margins and a fitted lowess smoother through the data HINT.

a. Is there any evidence of **nonlinearity**? (Y or N) ☐

b. Is there any evidence of **nonnormality**? (Y or N) ☐

c. Is there any evidence of **unequal variance**? (Y or N) ☐

Q2-4. The main intention of the researchers is to investigate whether there is a linear relationship between the density of riparian vegetation and the size of the logs. They have no of using the model equation for further predictions, not are they particularly interested in the magnitude of the relationship (slope). Is **model I or II regression** appropriate in these circumstances?. Explain?

If there is no evidence that the assumptions of simple linear regression have been violated, **fit the linear model** CWDBASAL = (SLOPE * RIPDENS) + intercept HINT. At this stage ignore any output.

Q2-5. **Examine the regression diagnostics** (particularly the **residual plot**) HINT. Does the residual plot indicate any potential problems with the data? (Y or N) ☐

Q2-6. At this point, we have no evidence to suggest that the hypothesis tests will not be reliable. Examine the regression output and identify and interpret the followingCalculate the following:

a. sample y-intercept ☐

b. sample slope ☐

c. t value for main H_0 ☐

d. P-value for main H_0 ☐

e. R^2 value ☐

Q2-7. What conclusions (statistical and biological) would you draw from the analysis?

Question 3 - Simple linear regression

Here is a modified example from Quinn and Keough (2002). Peake & Quinn (1993) investigated the relationship between the number of individuals of invertebrates living in amongst clumps of mussels on a rocky intertidal shore and the area of those mussel clumps.

Format of peakquinn.csv data files

AREA	INDIV
516.00	18
469.06	60
462.25	57
938.60	100
1357.15	48
...	...

AREA	Area of mussel clump mm^2 - Predictor variable
INDIV	Number of individuals found within clump - Response variable

Open the peakquinn data file.

The relationship between two continuous variables can be analyzed by simple linear regression. As with question 2, note that the levels of the predictor variable were measured, not fixed, and thus parameter estimates should be based on model II RMA regression. Note however, that the hypothesis test for slope is uneffected by whether the predictor variable is fixed or measured.

Before performing the analysis we need to check the assumptions. To evaluate the assumptions of linearity, normality and homogeneity of variance, construct a **scatterplot** of INDIV against AREA (INDIV on y-axis, AREA on x-axis) including a lowess smoother and boxplots on the axes.

Q3-1. Consider the assumptions and suitability of the data for simple linear regression:

a. In this case, the researchers are interested in investigating whether there is a relationship between the number of invertebrate individuals and mussel clump area as well as generating a predictive model. However, they are not interested in the specific magnitude of the relationship (slope) and have no intension of comparing their slope to any other non-zero values. Is **model I or II regression** appropriate in these circumstances?. Explain?

b. Is there any evidence that the other assumptions are likely to be violated? ☐

To get an appreciation of what a residual plot would look like when there is some evidence that the assumption of homogeneity of variance assumption has been violated, perform the **simple linear regression (by fitting a linear model)** purely for the purpose of **examining the regression diagnostics** (particularly the **residual plot**)

Q3-2. How would you describe the residual plot?

Q3-3. What could be done to the data to address the problems highlighted by the scatterplot, boxplots and residuals?

Q3-4. Describe how the scatterplot, axial boxplots and residual plot might appear following successful data transformation.

Transform both variables to logs (base 10), replot the scatterplot using the transformed data, refit the linear model (again using transformed data) and examine the residual plot.HINT

Q3-5. Would you consider the transformation as successful? (Y or N) ☐

Q3-6. If you are satisfied that the assumptions of the analysis are likely to have been met, perform the linear regression analysis (**fit the linear model**) HINT, examine the output.

a. Test the null hypothesis that the population slope of the regression line between log number of individuals and log clump area is zero - use either the t-test or ANOVA F-test regression output. What are your conclusions (statistical and biological)? HINT

b. If the relationship is significant construct the regression equation relating the number of individuals in the a clump to the clump size. Remember that parameter estimates should be based on RMA regression not OLS!

DV	= intercept +	slope	x	IV
$\mathrm{Log_{10}Individuals}$	☐	☐		$\mathrm{Log_{10}Area}$

Q3-7. Write the results out as though you were writing a research paper/thesis. For example (select the phrase that applies and fill in gaps with your results):
A linear regression of log number of individuals against log clump area showed (choose correct option)

(choose correct option) ☐ (b = ☐, t = ☐, df = ☐, P = ☐)

Q3-8. How much of the variation in log individual number is explained by the linear relationship with log clump area? That is , what is the R^2 value? ☐

Q3-9. What number of individuals would you **predict** for a new clump with an area of 8000 mm^2? HINT

☐

Q3-10. Given that in this case both response and predictor variables were measured (the levels of the predictor variable were not specifically set by the researchers), it might be worth presenting the less biased model parameters (y-intercept and slope) from RMA model II regression. Perform the **RMA model II regression** and examine the slope and intercept.

a. b (slope): ☐

b. c (y-intercept): ☐

Q3-11. Significant simple linear regression outcomes are usually accompanied by a scatterpoint that summarizes the relationship between the two population. Construct a **scatterplot** **without a smoother or marginal boxplots**. **Consider whether or not transformed or untransformed data should be used in this graph**.

Question 4 - Model II RMA regression

Nagy, Girard & Brown (1999) investigated the allometric scaling relationships for mammals (79 species), reptiles (55 species) and birds (95 species). The observed relationships between body size and metabolic rates of organisms have attracted a great deal of discussion amongst scientists from a wide range of disciplines recently. Whilst some have sort to explore explanations for the apparently 'universal' patterns, Nagy *et al.* (1999) were interested in determining whether scaling relationships differed between taxonomic, dietary and habitat groupings.

Format of nagy.csv data file

Species	**Common**	**Mass**	**FMR**	**Taxon**	**Habitat**	**Diet**	**Class**
Pipistrellus pipistrellus	Pipistrelle	7.3	29.3	Ch	ND	I	Mammal
Plecotus auritus	Brown long-eared bat	8.5	27.6	Ch	ND	I	Mammal
Myotis lucifugus	Little brown bat	9.0	29.9	Ch	ND	I	Mammal
Gerbillus henleyi	Northern pygmy gerbil	9.3	26.5	Ro	D	G	Mammal
Tarsipes rostratus	Honey possum	9.9	34.4	Tr	ND	N	Mammal
..	..	..	..	..	..	..	..

Open the nagy data file.

For this example, we will explore the relationships between field metabolic rate (FMR) and body mass (Mass) in grams for the entire data set and then separately for each of the three classes (mammals, reptiles and aves).

Unlike the previous examples in which both predictor and response variables could be considered 'random' (measured not set), parameter estimates should be based on model II RMA regression. However, unlike previous examples, in this example, the primary focus is not hypothesis testing about whether there is a relationship or not. Nor is prediction a consideration. Instead, the researchers are interested in establishing (and comparing) the allometric scaling factors (slopes) of the metabolic rate - body mass relationships. Hence in this case, model II regression is indeed appropriate.

Q4-1. Before performing the analysis we need to check the assumptions. To evaluate the assumptions of linearity, normality and homogeneity of variance, construct a **scatterplot** of FMR against Mass including a lowess smoother and boxplots on the axes.

a. Is there any evidence of non-normality? ☐

b. Is there any evidence of non-linearity? ☐

Q4-2. Typically, allometric scaling data are treated by a log-log transformation. That is, both predictor and response variables are $\log_{10}$ transformed. This is achieved graphically on a scatterplot by plotting the data on log transformed axes. Produce such a scatterplot (HINT). Does this improve linearity? ☐

Q4-3. Fit the linear models relating log-log transformed FMR and Mass using both the Ordinary Least Squares and Reduced Major Axis methods separately for each of the classes (mammals, reptiles and aves). Indicate the following;

	OLS		**RMA**		
Class	**Slope**	**Intercept**	**Slope**	**Intercept**	**R^2**
Mammals (HINT and HINT)	☐	☐	☐	☐	☐
Reptiles (HINT and HINT)	☐	☐	☐	☐	☐
Aves (HINT and HINT)	☐	☐	☐	☐	☐

Q4-4. Produce a scatterplot that depicts the relationship between FMR and Mass just for the mammals (HINT)

a. Fit the OLS regression line to this plot (HINT)

b. Fit the RMA regression line (in red) to this plot (HINT)

Q4-5. Compare and contrast the OLS and RMA parameter estimates. Explain which estimates are most appropriate and why the in this case the two methods produce very similar estimates.

Q4-6. To see how the degree of correlation can impact on the difference between OLS and RMA estimates, fit the relationship between FMR and Mass for the entire data set.

a. Complete the following table

	OLS		**RMA**		
Class	**Slope**	**Intercept**	**Slope**	**Intercept**	**R^2**
Entire data set (HINT and HINT)	☐	☐	☐	☐	☐

b. Produce a scatterplot that depicts the relationship between FMR and Mass just for the mammals (HINT)

c. Fit the OLS regression line to this plot (HINT)

d. Fit the RMA regression line (in red) to this plot (HINT)

e. Compare and contrast the OLS and RMA parameter estimates. Explain which estimates are most appropriate and why the in this case the two methods produce not so similar estimates.

Question 5 - Power analysis and regression

A feeding ecologist wished to investigate the energetic and nutritional consequences of lactation on captive Tasmanian pademelons (*Thylogale billardierii* - a small wallaby). The researcher was primarily interested in compensatory alterations in food intake and chewing parameters (such as chew rate). Such measures are extremely difficult to obtain accurately and require intense investigation, thereby restricting the sample sizes. Prior to commensing the investigation, the researcher wisely decided to perform a quick power analysis so as to gauge the estimated sample size necessary to detect a linear trend in chewing rate (chews. min^{-1}) with increasing joey mass (g). Previous research (Rose *et al*, 2005) into changes in milk composition and growth and joey growth in the species had demonstrated a 4-fold increase in the energy content of milk throughout lactation.

Fig. 5 from Rose *et al*, 2005

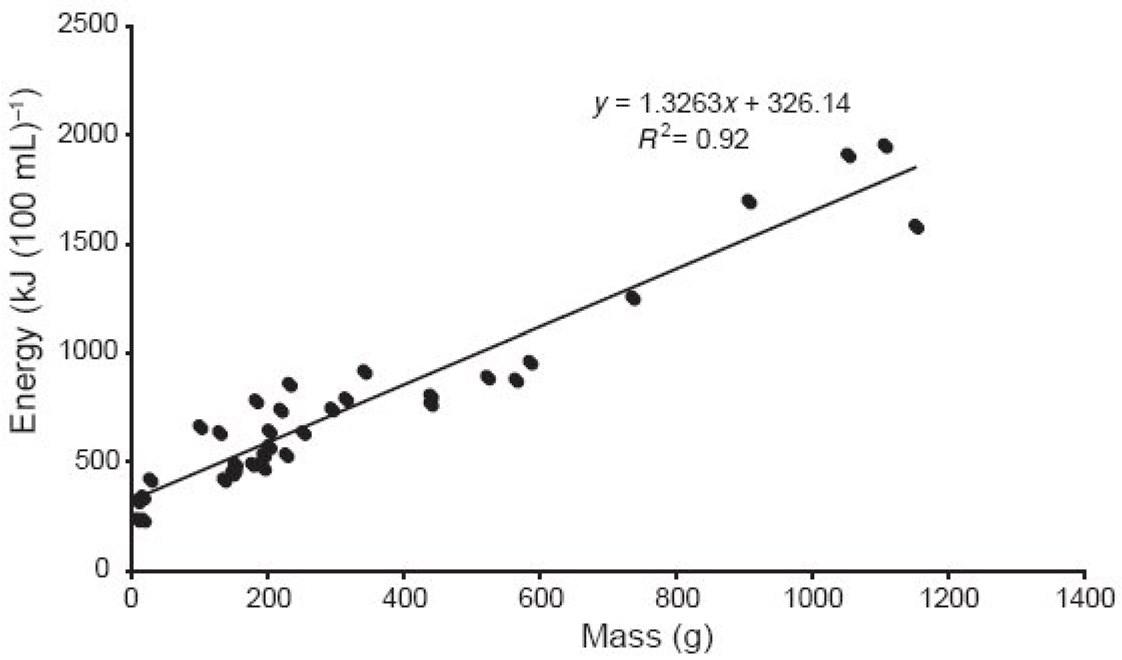

Fig. 5. The change in mass of pouch young plotted against energy content of Tasmanian pademelon milk during lactation (the regression equation and correlation coefficient, R^2, are included).

Q5-1. Using the information from the previous research, and assuming that feeding parameters are expected to be primarily influenced by energetic demands, **estimate the sample size required to have an 80% change of detecting** a linear relationship between chew rate and pouch young mass in captive Tasmanian pademelons. HINT

Q5-2. Alternatively, given the difficulty of obtaining accurate chew rate data over 24 hour periods, the researchers may have deemed that it was not possible to collect more than six observations. Given this restriction, **estimate the probability of detecting** a linear relationship between chew rate and pouch young mass in captive Tasmanian pademelons. HINT

Q5-3. As with power analyses for t-tests, it is often useful to be able to visualize the relationship between sample size and power over a range of sample sizes (or power outcomes). Given the degree of correlation, estimate the relationship between power and sample size for:

a. **a range of sample sizes (between 4 and 10)**. Note that regression and correlation with fewer than 4 observations are of limited value. HINT

b. **a range of power (between 0.6 and 0.99)**. Note that power values that yeild sample sizes less than 4 will invoke errors. HINT

Welcome to the end of Worksheet 3!