Worksheet 3 - Regression and linear models

Linear regression references

- Fowler *et al.* (1998) -Chpts 14 & 15
- Holmes *et al.* (2006) - Chpt 6
- Quinn & Keough (2002) - Chpt 5
- Rowntree (1981) - Chpts 12

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. Examine the suitability of the data for simple linear regression.

a. 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) ☐

b. Verify that **model I regression** is appropriate. Is model I regression appropriate, explain?

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**.

Q1-7. Within a research report, such a figure should either be embedded within the results section, or else on a separate page at the end of the report. **Save the graph as a picture** and **import the picture into Microsoft Word**. Alternatively, the graph can be **cut and pasted into Word**.

Question 2 - Linear relationships

Let's use the example from Quinn and Keough (2002) to become familiar with fitting linear models and interpreting the output. 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, $trees.km^{-1}$).

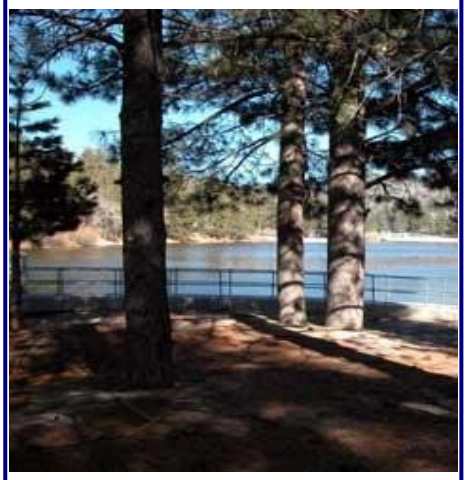
Format of christ.csv data files

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`. Add **boxplots** for each variable to the margins and fit a **lowess smoother** through the data.

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 = intercept + (SLOPE * RIPDENS)`. At this stage ignore any output.

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

Q2-6. 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 ☐

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

Q2-8. 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**.

Q2-9. Within a research report, such a figure should either be embedded within the results section, or

else on a separate page at the end of the report. **Save the graph as a picture** and **import the picture into Microsoft Word**. Alternatively, the graph can be **cut and pasted into Word**.

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 was seen in question 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 INDIV against AREA (INDIV on y-axis, AREA on x-axis) including a **lowess smoother** and **boxplots** on the axes.

Q3-1. Is there any evidence that any of the assumptions or conditions of simple linear regression are likely to be violated?

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 **appropriate 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.

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

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**), examine the output, and use the information to construct the regression equation relating the number of individuals in the a clump to the clump size (note that as the estimates are based on model I OLS regression, the estimates may be heavily biased):

DV	=	intercept	+	slope	x	IV
Log_{10}Individuals						Log_{10}Area

Q3-6. 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)?

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?

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**.

Q3-12. Within a research report, such a figure should either be embedded within the results section, or else on a separate page at the end of the report. **Save the graph as a picture** and **import the picture into Microsoft Word**. Alternatively, the graph can be **cut and pasted into Word**.

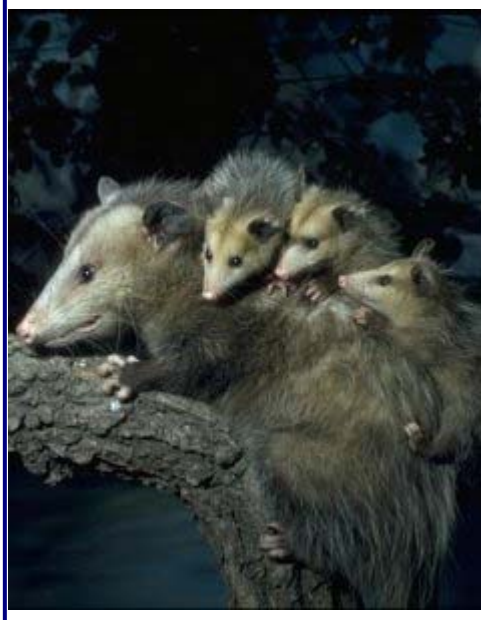
Question 4 - Polynomial regression

Rademaker and Cerqueira (2006), compiled data from the literature on the reproductive traits of opossoms (*Didelphis*) so as to investigate latitudinal trends in reproductive output. In particular, they were interested in whether there were any patterns in mean litter size across a longitudinal gradient from 44^{o}N to 34^{o}S. Analyses of data compiled from second hand sources are called metaanalyses and are very usefull at revealing overal trends across a range of studies.

Format of rademaker.csv data files

SPECIES	LATITUDE	L/Y	MAOP	MLS	REFERENCE
D.v.	44	2	16.8	8.4	Tyndale-Biscoe and Mackenzie (1976)
D.v.	41	1.5	14.1	9.4	Hossler et al. (1994)
D.v.	41.5	?	?	8.6	Reynolds (1952)
D.v.	41	2	18	9	Wiseman and Hendrickson (1950)
D.v.	40	2	15.8	7.9	Sanderson (1961)
...	...	...	...	...	...

SPECIES	*Didelphid* species (*D.al.=Didelphis albiventris*, *D.au.=Didelphis aurita*, *D.m.=Didelphis marsupialis*, *D.v.=Didelphis virginiana* - Descriptor variable
LATITUDE	Lattitude (degees) of study site - Predictor variable
L/Y	Mean number of litter per year - Response variable
MAOP	Mean annual offspring production - Response variable
MLS	Mean litter size - Response variable
REFERENCE	Original source of data



Open the rademaker data file.

The main variables of interest in this data set are MLS (mean litter size) and LATITUDE. The other variables were included so as to enable you to see how meta data might be collected and collated from a number of other sources.

The relationship between two continuous variables can be analyzed by simple linear regression, as was seen in

question 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 MLS against LATITUDE including a **lowess smoother** and **boxplots** on the axes.

Q4-1. Is there any evidence that any of the 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 linearity 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**)

Q4-2. How would you describe the residual plot?

For this sort of trend that is clearly non-linear (yet the boxplots suggest normal data), transformations are of no use. Therefore, rather than attempt to model the data on a simple linear relationship (straight line), it is better to attempt to model the data on a curvilinear linear relationship (curved line). **Note** it is important to make the distinction between **line (relationship) linearity and model linearity**

Q4-3. If the assumptions are otherwise met, perform the second order polynomial regression analysis (**fit the quadratic polynomial model**), examine the output, and use the information to construct the regression equation relating the number of mean litter size to latitude:

DV	=	intercept	+	$slope_1$	x	IV^2	+	$slope_2$	x	IV
Mean litter size	=	☐	+	☐	x	$latitude^2$	+	☐	x	latitude

Q4-4. In polynomial regression, there are two hypotheses of interest. Firstly, as there are two slopes, we are now interested in **whether the individual slopes are equal to one another and to zero** (that is, does the overall model explain our data better than just a horizontal line (no trend). Secondly, we are interested in whether the **second order polynomial model fits the data any better than a simple first order (straight-line) model**. Test these null hypotheses.

a. The slopes (choose correct option) ☐ significantly different from one another and to 0 (F = ☐, df = ☐, ☐, P = ☐)

b. The second order polynomial regression model (choose correct option) ☐ fit the data significantly better than a first order (straight line) regression model (F = ☐, df = ☐, ☐, P = ☐)

Q4-5. What are your conclusions (statistical and biological)?

Q4-6. Such an outcome might also be accompanied by a scatterpoint that illustrates the relationship between the mean litter size and latitude. Construct a **scatterplot** **without a smoother or marginal boxplots**. **Include a quadratic (second order) trendline on this graph**.

Question 5 - Nonlinear Regression

Recall from question 3 above the data set of Peake and Quinn (1993) that demonstrated the relationship between island (individual mussel clumps) size and the number of individuals supported on the islands. Peake and Quinn (1993) also investigated the relationship between the size of mussel clumps (m^2) and the number of other invertebrate species supported.

Format of peake.csv data file

AREA	SPECIES	INDIV
516.0	3	18
469.06	7	60
462.25	6	57
...	...	...

AREA Area of the mussel clump (m^2)- predictor variable

SPECIES Number of other invertebrate species found in the mussel clumps - response variable

INDIV Number of other invertebrate individuals found in the mussel clumps - ignore this response variable

Open the peake data file.

Q5-1. For this question we will focus on an examination of the relationship between the number of species occupying a mussel clump and the size of the mussel clump. Construct a **scatterplot** to investigate the assumptions of simple linear regression

a. Focussing on the assumptions of Normality, Homogeneity of Variance and Linearity, is there evidence of violations of these assumptions (y or n)? ☐

Q5-2. Although we could try a logarithmic transformation of the AREA variable (as we did in question 3), species/area curves are known to be non-linear. We might expect that small clumps would only support a small number of species. However, as area increases, there should be a dramatic increase in the number of species supported. Finally, further increases in area should see the number of species approach an asymptote. Species-area data are often modeled with non-linear functions.

a. Try fitting a second order **polynomial trendline** through these data. E.g. Species = α*Area + β*Area^2 + c (note that this is just an extension of y=mx+c)

b. Try fitting a **power trendline** through these data. E.g. Species = $\alpha(\text{Area})^{\beta}$ where the number of Species is proportional to the Area to the power of some coefficient (β).

c. Which model is the most appropriate for these data and why?

Q5-3. Fit the above power function to the data and complete the following table.

Parameter	Estimate	t-value	P-value

α				
β				

Q5-4. What conclusions (statistical and biological) would you draw from the analysis?

Q5-5. Create a plot of species number against mussel clump area.

a. **Fit the nonlinear trend line to this plot**

Welcome to the end of Worksheet 3