Worksheet 4 - Multiple and non-linear regression models

Multiple and non-linear regression references

- Quinn & Keough (2002) - Chpt 6

Question 1 - Multiple Linear Regression

Paruelo & Lauenroth (1996) analyzed the geographic distribution and the effects of climate variables on the relative abundance of a number of plant functional types (PFT's) including shrubs, forbs, succulents (e.g. cacti), C3 grasses and C4 grasses. They used data from 73 sites across temperate central North America (see pareulo.syd) and calculated the relative abundance of C3 grasses at each site as a response variable

Format of paruelo.csv data file

C3	LAT	LONG	MAP	JJAMAP	DJFMAP
..	..	..	..	..	..

C3	Relative abundance of C3 grasses at each site - response variable
LAT	Latitude in centesimal degrees - predictor variable
LONG	Longitude in centesimal degrees - predictor variable
MAP	Mean annual precipitation (mm) - predictor variable
MAT	Mean annual temperature (^{0}C) - predictor variable
JJAMAP	Proportion of MAP that fell in June, July and August - predictor variable
DJFMAP	Proportion of MAP that fell in December, January and Febrary - predictor variable

Open the paruelo data file. HINT.

Q1-1. In the table below, list the assumptions of multiple 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.	
V.	

Q1-2. Construct a **scatterplot matrix** to investigate these assumptions(HINT)

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

b. Try applying a temporary square root transformation (HINT). Does this improve some of these specific assumptions (y or n)? ☐

c. Is there any evidence that the assumption of Collinearity is likely to be violated (y or n)? ☐

d. Collinearity occurs when one or more of the predictor variables are correlated, therefore this assumption can also be examined by calculating the pairwise correlation coefficients between all the predictor variables (HINT). Which predictor variables are highly correlated?

Q1-3. (Multi)collinearity can also be diagnosed by **tolerance and variance inflation factor (VIF) measures**.

a. Calculate the VIF values for each of the predictor variables (HINT).

b. Calculate the tolerance values for each of the predictor variables (HINT).

Predictor	**Tolerane**	**VIF**
LAT	☐	☐
LONG	☐	☐
MAP	☐	☐
MAT	☐	☐
JJAMAP	☐	☐
$\log_{10}$(DJFMAP)	☐	☐

c. Is there any evidence of (multi)collinearity, and if so, which variables are responsible for violations of this assumption?

We obviously cannot easily incorporate all 6 predictors into the one model, because of the collinearity problem. Paruelo and 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 C3 plants and latitude and longitude. This relationship will investigate the geographic pattern in abundance of C3 plants.

Q1-4. Just like Paruelo and Lauenroth (1996), we will fit the multiplicative model for LAT and LONG.

- **a.** Write out the full multiplicative model
- **b.** Check the assumptions of this linear model. In particular, check collinearity. HINT
- **c.** Obviously, this model will violate collinearity. It is highly likely that LAT and LONG will be related to the LAT:LONG interaction term. It turns out that if we centre the variables, then the individual terms will no longer be correlated to the interaction. Centre the LAT and LONG variables (HINT) and (HINT)

Q1-5. Fit a linear multiplicative model on the centrered LAT and LONG (HINT)

- **a.** Examine the diagnostic plots (HINT) specially the residual plot to confirm no further evidence of violations of the analysis assumptions
- **b.** Complete the following table (HINT).

Coefficient	Estimate	t-value	P-value
Intercept	☐	☐	☐
cLAT	☐	☐	☐
cLONG	☐	☐	☐
cLAT:cLONG	☐	☐	☐

Q1-6. Examine the partial regression plots of LAT and LONG (HINT).

Q1-7. There is clearly an interaction between LAT and LONG. This indicates that the degree to which latitude effects the relative abundance of C3 plants depends on longitude. To investigate this further, we will examine the simple effects of latitude at a specific range of longitudes. The levels of longitude that we will use are the mean longitude value as well as the mean plus or minus 1 SD and plus or minus 2 SD.

a. Calculate the five levels of longitude on which the simple effects of latitude will be investigated. (HINT, HINT, etc).

b. Investigate the simple effects of latitude on the relative abundance of C3 plants for each of these levels of longitude. (HINT), (HINT), etc).

Question 2 - Multiple Linear Regression

Loyn (1987) modeled the abundance of forest birds with six predictor variables (patch area, distance to nearest patch, distance to nearest larger patch, grazing intensity, altitude and years since the patch had been isolated).

Format of loyn.csv data file

ABUND	DIST	LDIST	AREA	GRAZE	ALT	YR.ISOL
..	..	..	..	..	..	..

ABUND	Abundance of forest birds in patch- response variable
DIST	Distance to nearest patch - predictor variable
LDIST	Distance to nearest larger patch - predictor variable
AREA	Size of the patch - predictor variable
GRAZE	Grazing intensity (1 to 5, representing light to heavy) - predictor variable
ALT	Altitude - predictor variable
YR.ISOL	Number of years since the patch was isolated - predictor variable

Open the loyn data file. HINT.

Q2-1. In the table below, list the assumptions of multiple 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.	
V.	

Q2-2. Construct a **scatterplot matrix** to investigate these assumptions(HINT)

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

b. Try applying a temporary $\log_{10}$ transformation to the skewed variables(HINT). Does this improve some of these specific assumptions (y or n)? ☐

c. Is there any evidence that the assumption of Collinearity is likely to be violated (y or n)? ☐

Predictor	**Tolerance**	**VIF**
$\log_{10}$(DIST)	☐	☐
$\log_{10}$(LDIST)	☐	☐
$\log_{10}$(AREA)	☐	☐
GRAZE	☐	☐
ALT	☐	☐
YR.ISOL	☐	☐

d. Collinearity occurs when one or more of the predictor variables are correlated, therefore this assumption can also be examined by calculating the pairwise correlation coefficients between all the predictor variables (use transformed versions for those that required it) (HINT). Which predictor variables are highly correlated?

Since none of the predictor variables are highly correlated to one another, we can include all in the linear model fitting.

Q2-3. **Fit an additive linear model** relating ABUND to each of the predictor variables, but no interactions (HINT)

a. Examine the diagnostic plots (HINT) specially the residual plot to confirm no further evidence of violations of the analysis assumptions

b. Were any of the partial regression slopes significantly different from 0? Which one(s)?

Coefficient	**Estimate**	**t-value**	**P-value**
Intercept	☐	☐	☐
$\log_{10}$(DIST)	☐	☐	☐
$\log_{10}$(LDIST)	☐	☐	☐
$\log_{10}$(AREA)	☐	☐	☐
GRAZE	☐	☐	☐
ALT	☐	☐	☐
YR.ISOL	☐	☐	☐

Q2-4. We would now like to be able to find the **'best' regression model**. Calculate the adjusted r^2 (HINT), AIC (HINT) and BIC (HINT) for the full regression containing all six predictor variables.

Model	**Adj. r^2**	**AIC**	**BIC**
Full	☐	☐	☐

Q2-5. Compare all possible models and select the 'best' model based on AIC and BIC (HINT). Note that this requires loading the biology package!.

Selection based on:	**Model (e.g. ABUND ~ log10(DIST) + ALT)**	**Adj. r^2**	**AIC**
Adj. r^2		☐	
AIC		☐	
BIC		☐	

Question 3 - Hierachical partitioning

An alternative model selection procedure is called hierarchical partitioning. **Hierarchical partitioning** essentially determines the contributions of each predictor variable as both an individual predictor as well as a joint predictor in explaining the variation in the response variable. We will use hierarchical partitioning for model selection on the Loyn (1987) data set.

Open the loyn data file. HINT.

Q3-1. Perform the **hierarchical partitioning**

Q3-2. There are no formal hypothesis tests to determine what constitutes a significant contribution, and therefore which predictor variables to retain in the model. However, there are two ways in which significance of predictors can be inferred:

a. Retain all those predictors whose total contribution is greater than an appropriate critical correlation coefficient. We do this by first **Convert the r^2 values into standardized, normal z-scores (via correlation coefficients)** and then comparing these scores to a critical z-score of 1.65 (for 0.05). Perform these calculations for the individual and total contributions and indicate which of the variables contribute the most to the explained variance in forest bird abundance.

b. The second method is to use a randomization procedure to generate a distribution of partitioned r^2 values. The hier.part package comes with one such routine, that performs a randomization procedure for the independent contributions only. Run the **randomization procedure** and indicate which of the variables contribute the most to the explained variance in forest bird abundance.

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°N to 34°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. (HINT)

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** (HINT). Include a quadratic (second order) trendline on this graph (HINT).

Question 5 - Nonlinear Regression

Peake and Quinn (1993) 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. HINT.

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(HINT)

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, 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 = α(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?

Species = &alpha(Area)$^{\text{\&beta}}$
whereby the number of Species is proportional to the Area to the power of some coefficient (&beta). Fit the above power function to the data (HINT).

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

Parameter	Estimate	t-value	P-value
&alpha	☐	☐	☐
&beta	☐	☐	☐

Q5-4. Examine the residual plot (HINT).

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 (HINT).

a. Fit the nonlinear trend line to this plot (HINT)

Welcome to the end of Worksheet 4!