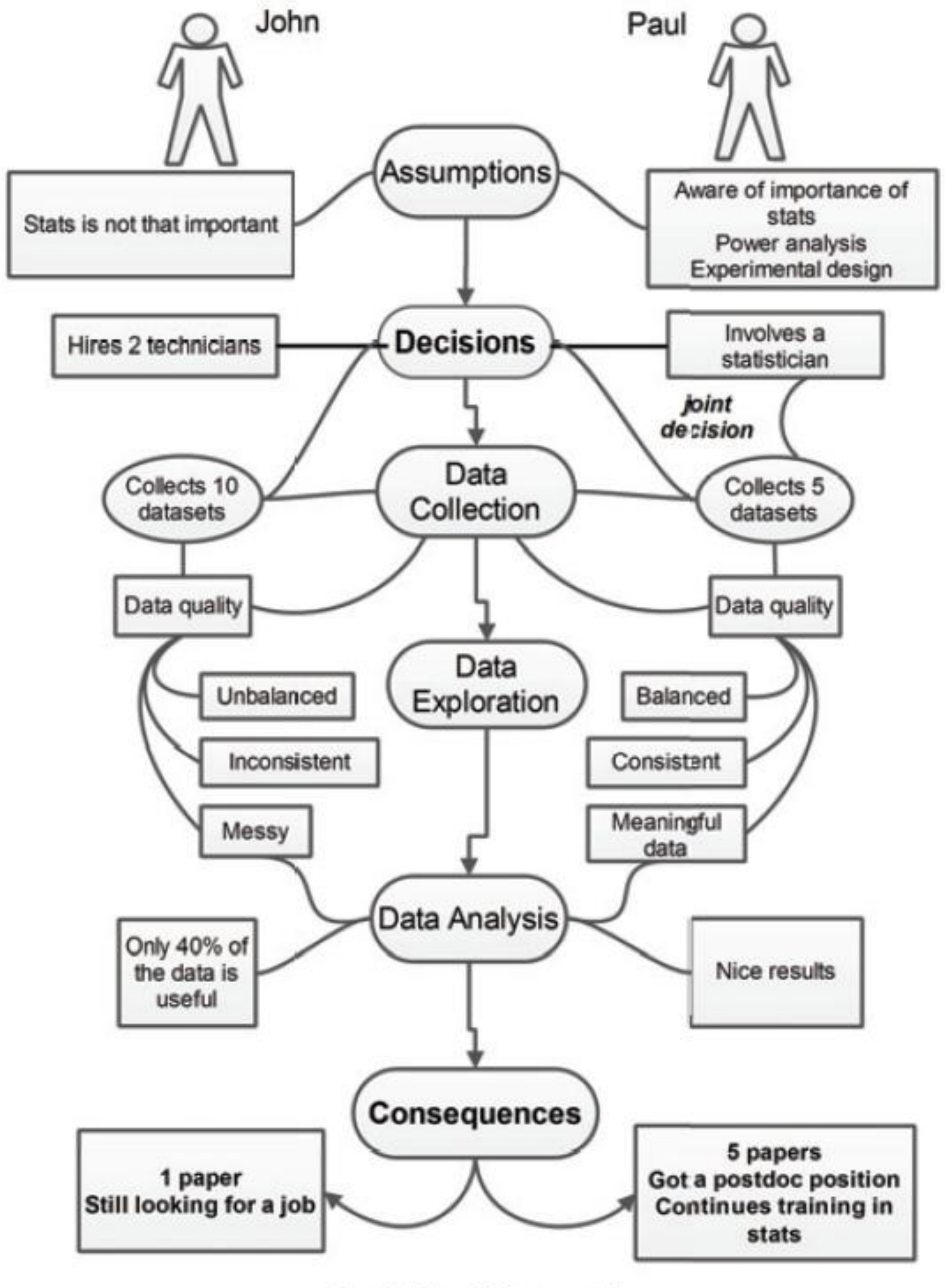


Spot the difference!

Bio 599: Week 2
Data Exploration

Take your name tent

Reminders/Updates

1. Thursday:
    - Read: “Zuur 2010 - A protocol for data exploration to avoid common statistical problems”
    - Workshop #2: Data exploration - due the following MONDAY.
2. Submit your ppt ‘paper discussion’ to Canvas
    - end of day Thursday.
    - See the assignment (and PPT) for details/expectations.
3. Office hours: Monday 10:30-11:30 and Thursday 4:00-5:00
4. Complete pre-coarse survey on canvas
5. Assignment #1 DUE SEPT 12

Exploratory Data Analysis

- What is it?
- What do you do?

Exploratory Data Analysis

- Before making inferences from data, it is essential to examine all your variables.

Why?

- to catch mistakes
- to see patterns in the data
- to find violations of statistical assumptions

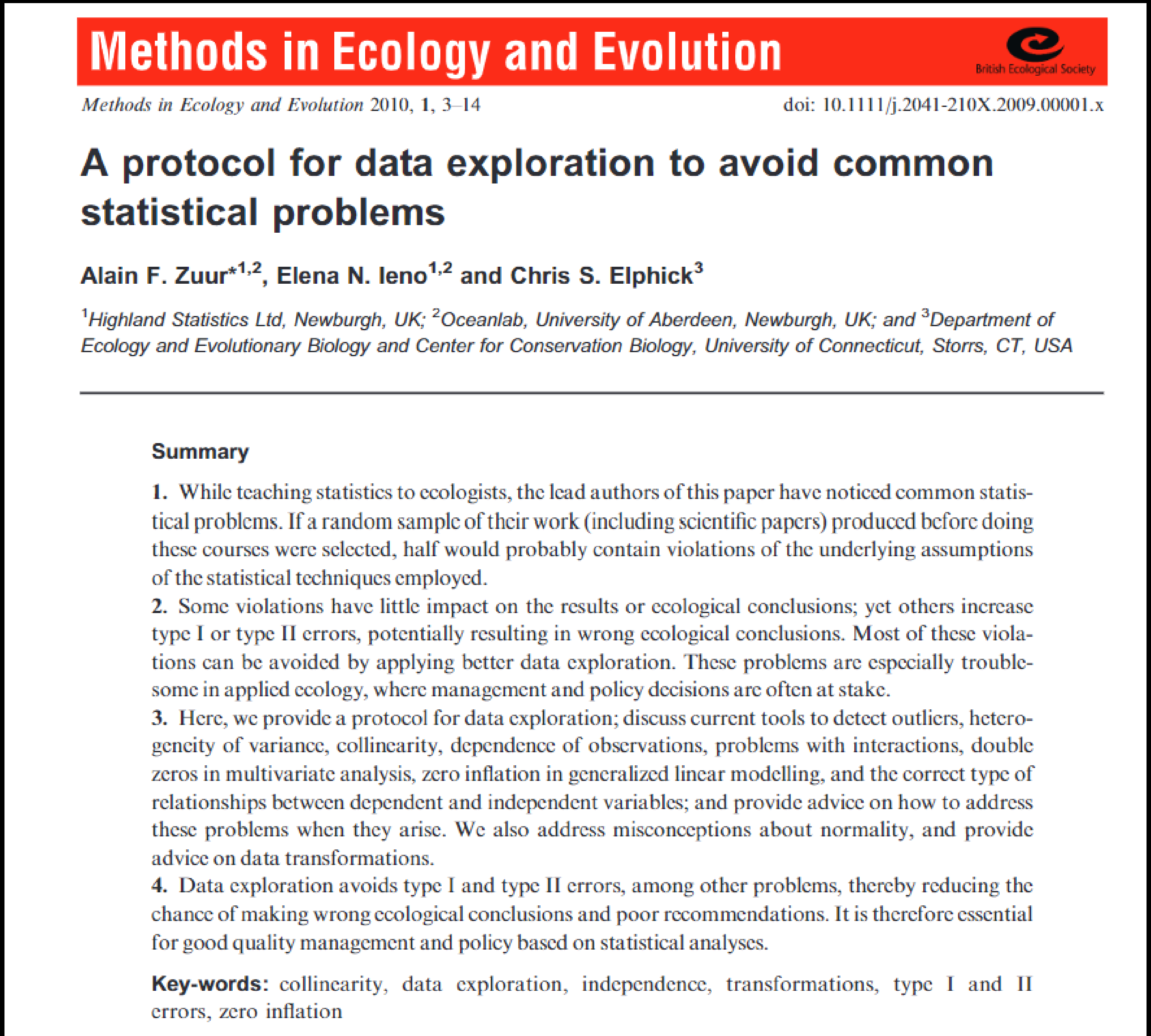
...and because if you don’t, you will have trouble later

Most violations can be avoided by applying better data exploration!

Exploratory Data Analysis

- Do NOT use data exploration to generate hypotheses!
- That is:
  - Data dredging
  - Data phishing
  - Data snooping

Exploratory Data Analysis



Methods in Ecology and Evolution

British Ecological Society

Methods in Ecology and Evolution 2010, **1**, 3–14 doi: 10.1111/j.2041-210X.2009.00001.x

A protocol for data exploration to avoid common statistical problems

Alain F. Zuur*[1,2], Elena N. Ieno[1,2] and Chris S. Elphick[3]

[1]Highland Statistics Ltd, Newburgh, UK; [2]Oceanlab, University of Aberdeen, Newburgh, UK; and [3]Department of Ecology and Evolutionary Biology and Center for Conservation Biology, University of Connecticut, Storrs, CT, USA

Summary

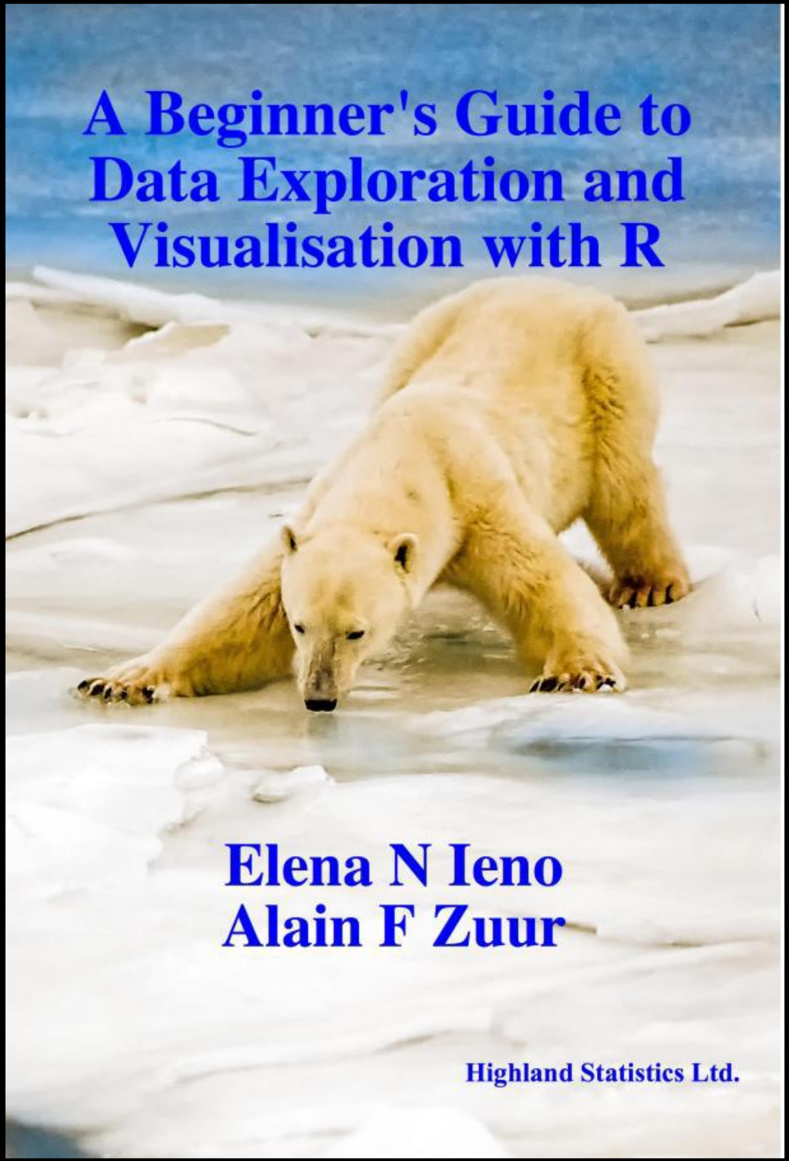
1. While teaching statistics to ecologists, the lead authors of this paper have noticed common statistical problems. If a random sample of their work (including scientific papers) produced before doing these courses were selected, half would probably contain violations of the underlying assumptions of the statistical techniques employed.

2. Some violations have little impact on the results or ecological conclusions; yet others increase type I or type II errors, potentially resulting in wrong ecological conclusions. Most of these violations can be avoided by applying better data exploration. These problems are especially troublesome in applied ecology, where management and policy decisions are often at stake.

3. Here, we provide a protocol for data exploration; discuss current tools to detect outliers, heterogeneity of variance, collinearity, dependence of observations, problems with interactions, double zeros in multivariate analysis, zero inflation in generalized linear modelling, and the correct type of relationships between dependent and independent variables; and provide advice on how to address these problems when they arise. We also address misconceptions about normality, and provide advice on data transformations.

4. Data exploration avoids type I and type II errors, among other problems, thereby reducing the chance of making wrong ecological conclusions and poor recommendations. It is therefore essential for good quality management and policy based on statistical analyses.

Key-words: collinearity, data exploration, independence, transformations, type I and II errors, zero inflation



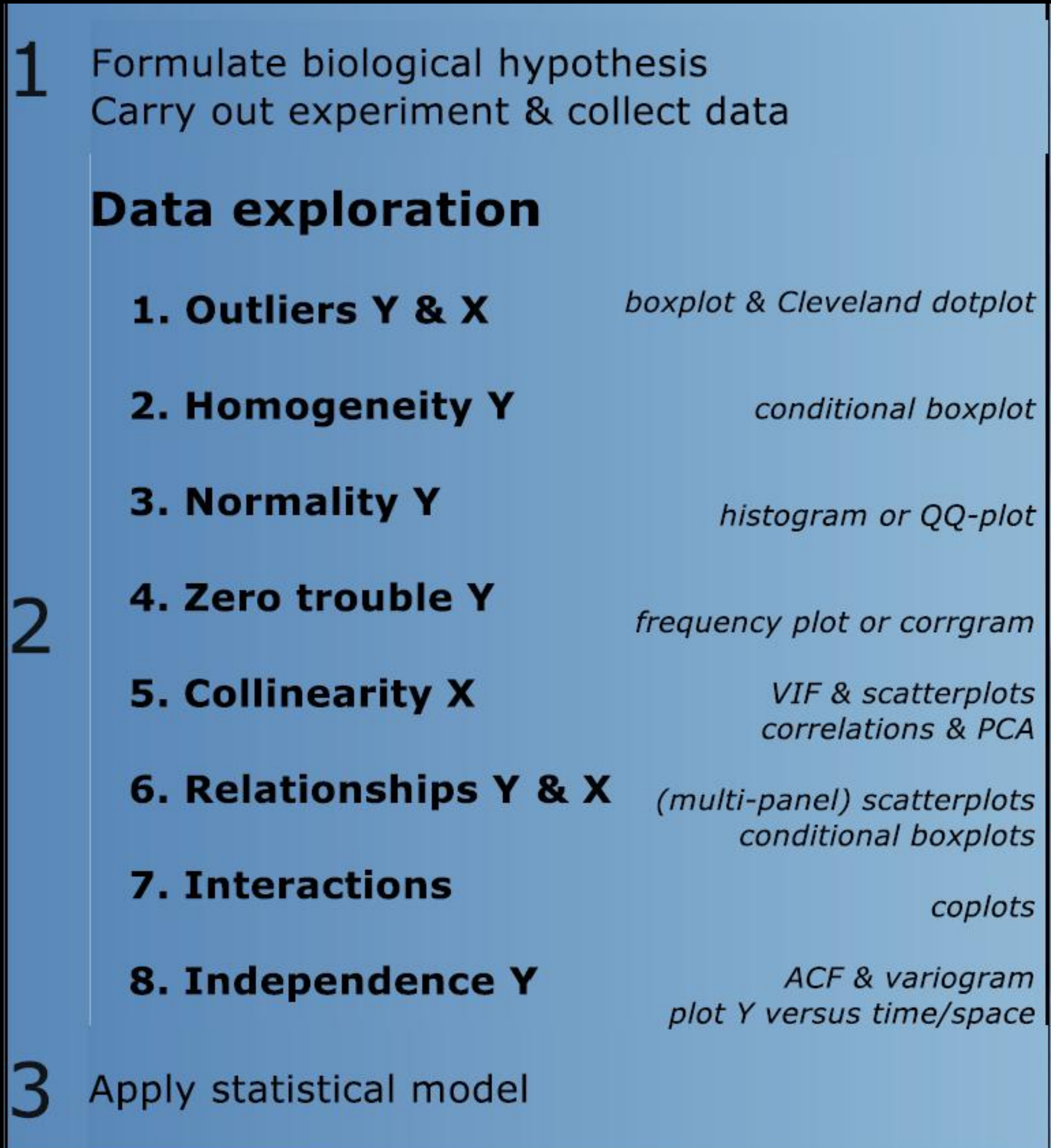
A Beginner's Guide to Data Exploration and Visualisation with R

Elena N Ieno
Alain F Zuur

Highland Statistics Ltd.

Exploratory Data Analysis

1. Outliers Y & X
2. Homogeneity Y
3. Normality Y
4. Zero trouble Y
5. Collinearity X
6. Relationships Y & X
7. Interactions
8. Independence Y



1 Formulate biological hypothesis
Carry out experiment & collect data

2 **Data exploration**

1. Outliers Y & X	*boxplot & Cleveland dotplot*
2. Homogeneity Y	*conditional boxplot*
3. Normality Y	*histogram or QQ-plot*
4. Zero trouble Y	*frequency plot or corrgram*
5. Collinearity X	*VIF & scatterplots* *correlations & PCA*
6. Relationships Y & X	*(multi-panel) scatterplots* *conditional boxplots*
7. Interactions	*coplots*
8. Independence Y	*ACF & variogram* *plot Y versus time/space*

3 Apply statistical model

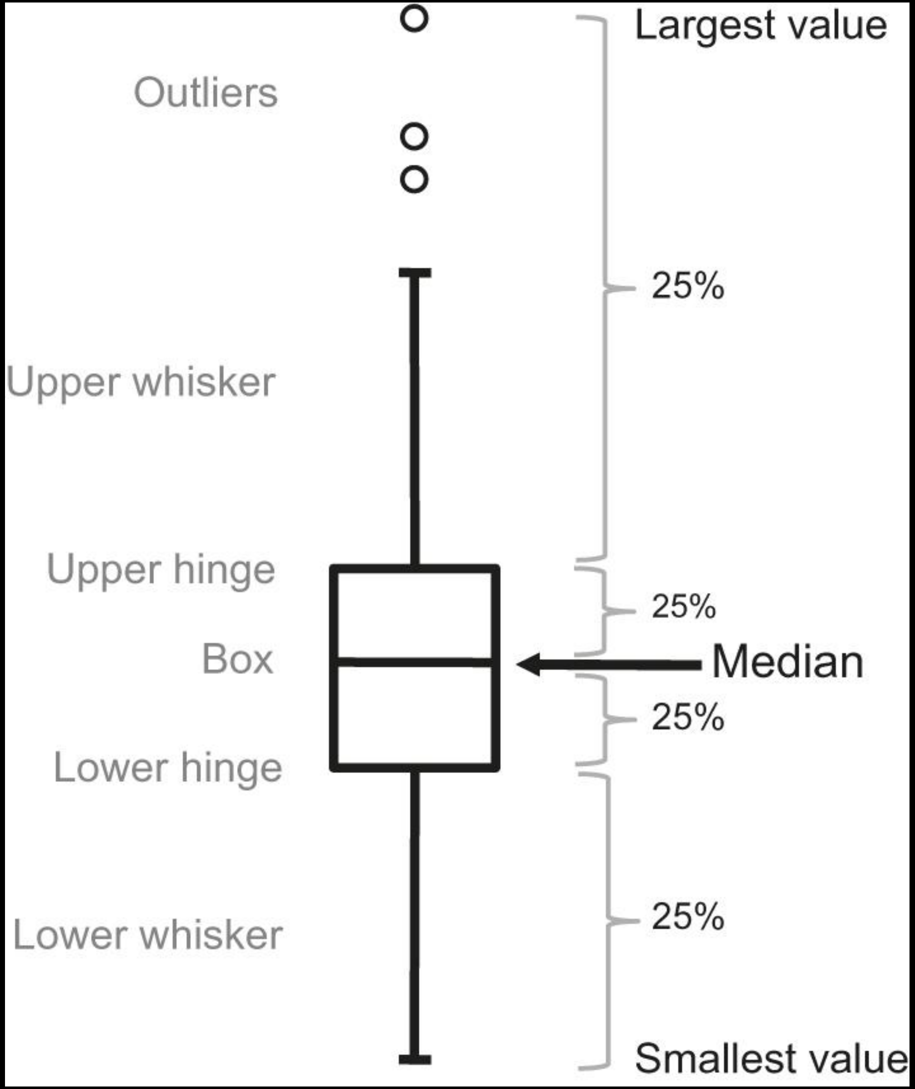
Outliers

- What are they?
- How do we detect them?
- What do we do with them?

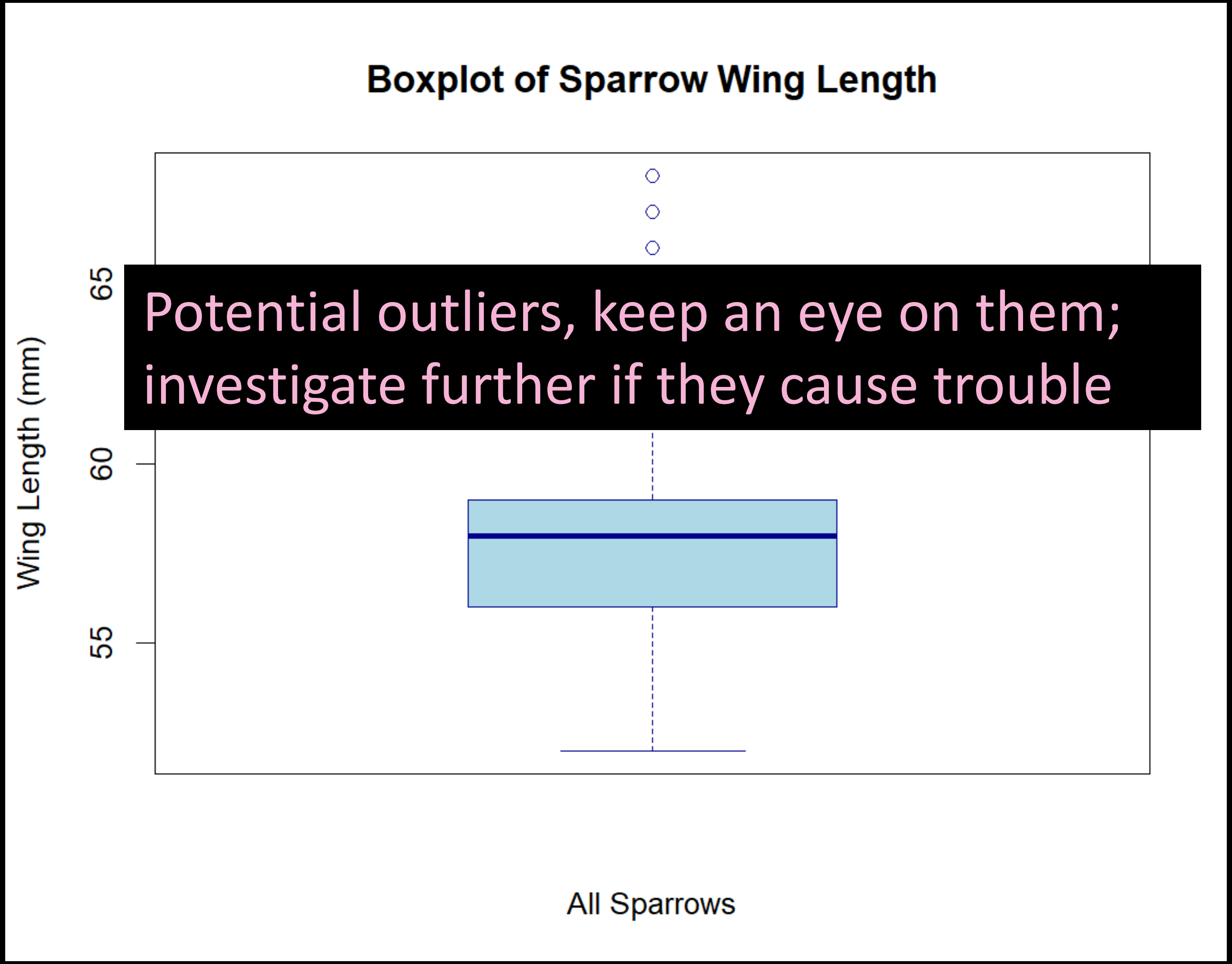
Outliers – what is an outlier?

- An outlier is an observation that has a rather large, or rather small, value compared to the bulk of observations for a variable of interest.
- Types of outliers:
  1. Outliers in one-dimensional space
  2. Outliers in two-dimensional space
  3. Influential observations in a regression-type analysis
  4. Nature of the data

Outliers - boxplots



Outliers - boxplots



Potential outliers, keep an eye on them; investigate further if they cause trouble

Figure 2.1. Boxplot for wing length of sparrows. Note that there are at least 6 observations that are considerably larger than the majority of the observations.

Outliers – tests

Z-score: an observation where the Z-score exceeds 3 may be viewed as an outlier.

- Z = (obs-mean)/std.dev

Grubbs test: checks the presence of a **single** outlier in a set of observations (assumes normality)

- *G* is the deviation between the mean and the maximum or minimum value, divided by the standard deviation
- Use the grubbs.test function in the outliers package

Studentised deviate test (ESD): checks if there are numerous outliers in your data.

- gesdTest in the PMCMRplus package

Outliers in two-dimensional space

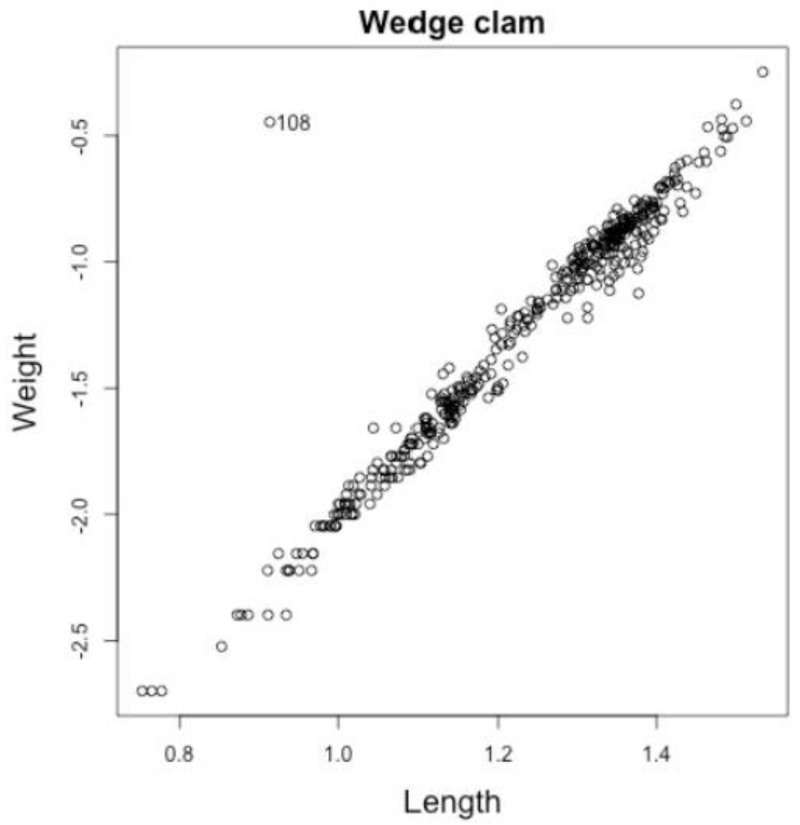
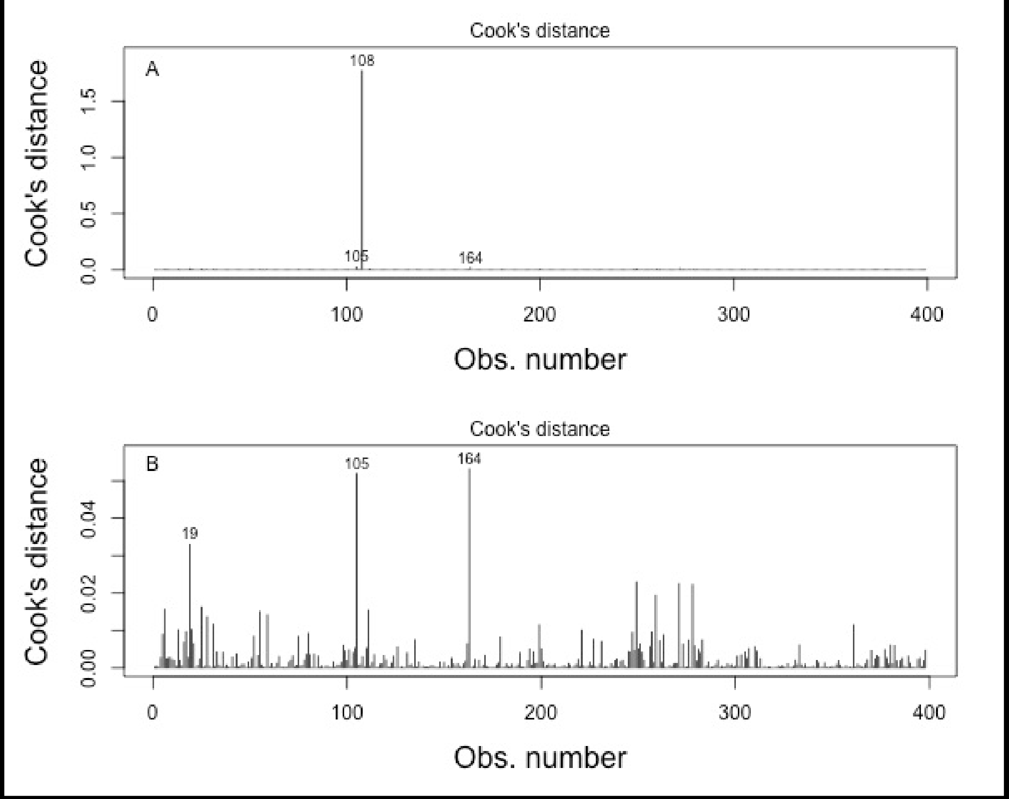


Figure 2.12. Scatterplot for the wedge clam data. Observation 108 is an outlier in the *x*-*y* space because it does not comply with the relationship.

Outliers - Influential observations

- Cook’s distance plot – assess the influence of each individual observation on the fitted values in regression type models
- Cook’s distance – identifies points which have a large influence on the fit. Value >1 is influential
- Does not necessarily mean you should remove point!



Discussion – how to deal with potential outliers in YOUR data

- Think about your data set (or the data set you will collect)
- What are some potential sources of outliers?
- How will you deal with outliers?

How do we deal with potential outliers?

1. Remove Outliers
2. Do not do anything and apply a statistical technique. This may work, or it may result in disaster
3. Present the results of the models with and without the outliers
4. Apply a transformation (distinction here between outliers in response variables and explanatory variables)

Homogeneity of Variance

- What is it and why do we care?
- How do we detect heterogeneity?
- What shall we do if we have heterogeneity?

Homogeneity of Variance

- Homogeneity (or homoscedasticity) occurs when the spread of all possible values of the population is the same for every value of the covariate
- The variance of the residuals (difference between observed and predicted values) is constant across all levels of the independent variables.

Homogeneity

Ignoring the problem may result in:

1. Increases Type 1 error
2. Regression parameters with biased standard errors
3. F statistics is no longer F-distributed
4. T statistics is no longer t-distributed
5. P-values are unreliable

Homogeneity - Boxplots

- Different centers = GOOD (Taxon effect)
- Different spread = BAD (heterogeneity)

- Need to use residuals to assess whether we really have heterogeneity

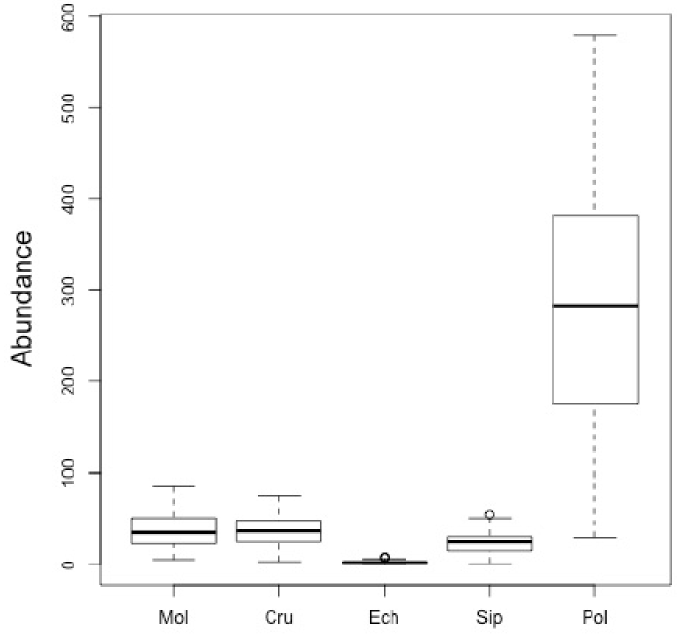
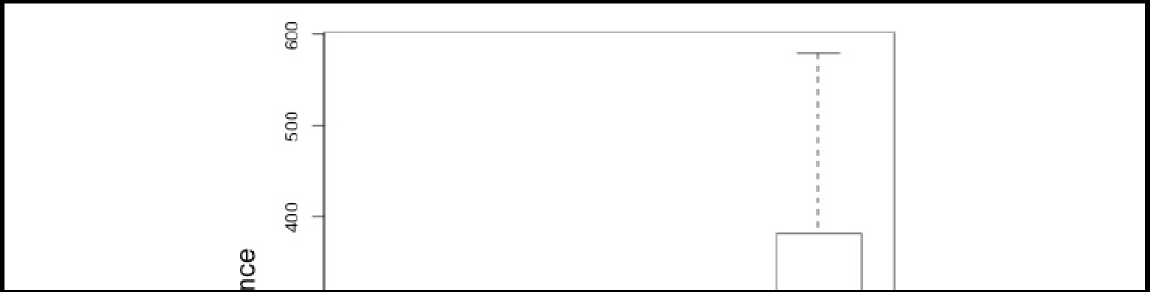


Figure 3.8. Conditional boxplot showing the conditional distributions of the variable Abundance given each value of the variable TaxonID. According to Fox (2008), in a linear regression model heterogeneity seriously degrades the least-square estimators if the ratio between the largest and smallest variance is 4 (conservative) or more.

Homogeneity - Boxplots

- Different centers = GOOD (Taxon effect)
- Different spread = BAD (heterogeneity)

- Need to use residuals to assess whether we really have heterogeneity



In multiple regression, some of the variation could be explained by a different covariate!

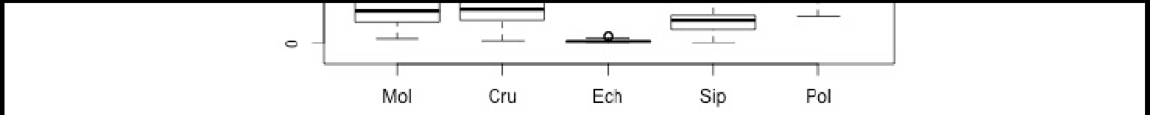


Figure 3.8. Conditional boxplot showing the conditional distributions of the variable `Abundance` given each value of the variable `TaxonID`. According to Fox (2008), in a linear regression model heterogeneity seriously degrades the least-square estimators if the ratio between the largest and smallest variance is 4 (conservative) or more.

Homogeneity - Scatterplots

- Useful for continuous explanatory variables

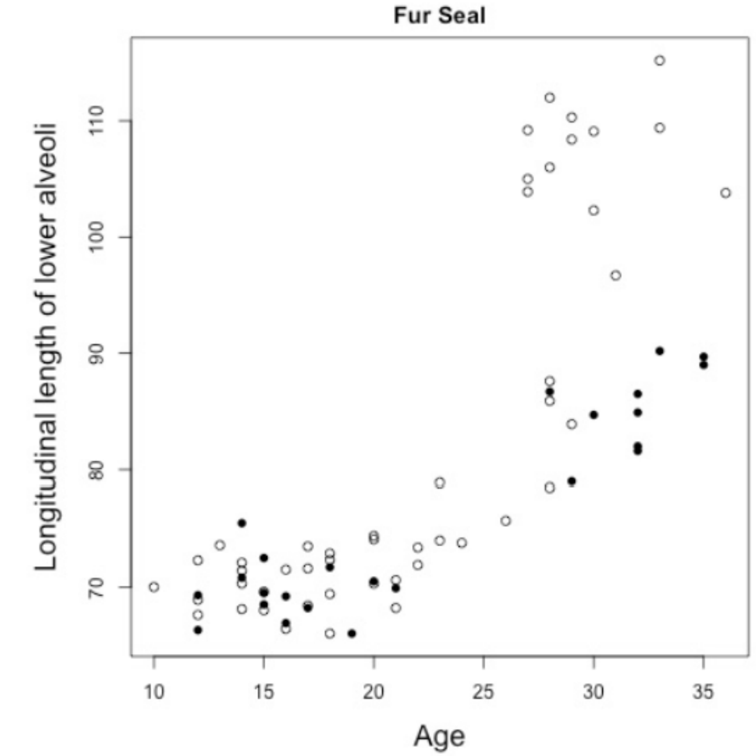


Figure 3.9. Scatterplot of longitudinal length of lower alveoli canine and age conditional on sex for the southern fur seal (*Arctocephalus australis*). Open circles indicate males and filled circles indicate females. Note the variation in length for the male adult specimens as age increases.

Homogeneity - tests

- Bartlett test – null hypothesis assumes that the variances in each of the groups (or samples) are the same. Requires normality of data.
- F-ratio test – decides if the variances in two independent samples are equal. Requires normality of data.
- Leven’s test – tests the assumption of equal population variances. Better for non-normal data.

Homogeneity – solutions?

- Transformation of the response variable to stabilize the variance
- Apply a statical technique that does not require homogeneity (GLS)

Normality?

- What is it?
- How do we detect it?

Normality

- Linear regression is often associated with normality of the response variable and each covariate

“Data have been checked for normality before doing the analysis”

Normality

- The normality assumption in regression means that the response variable is normally distributed at each value of the covariate

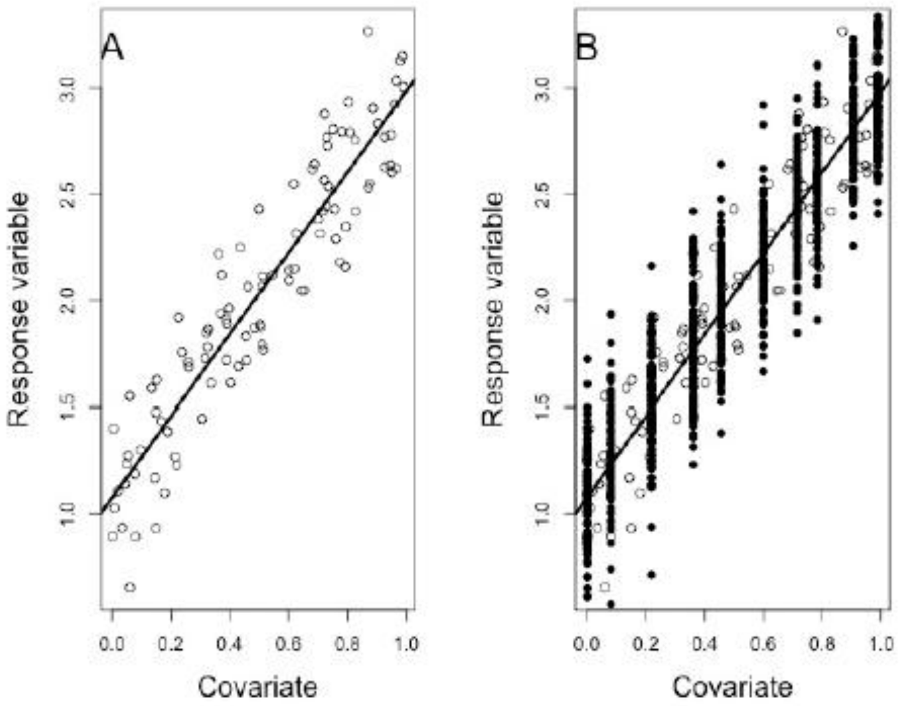
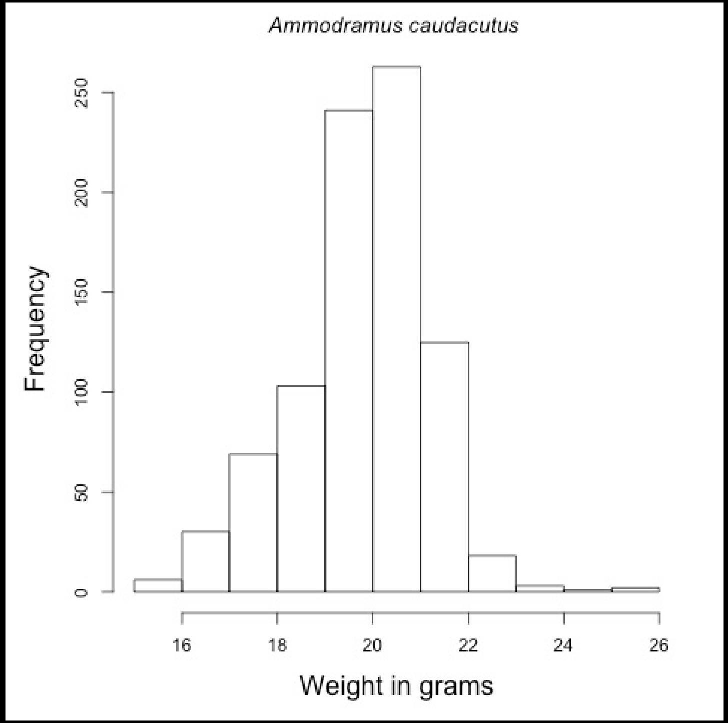
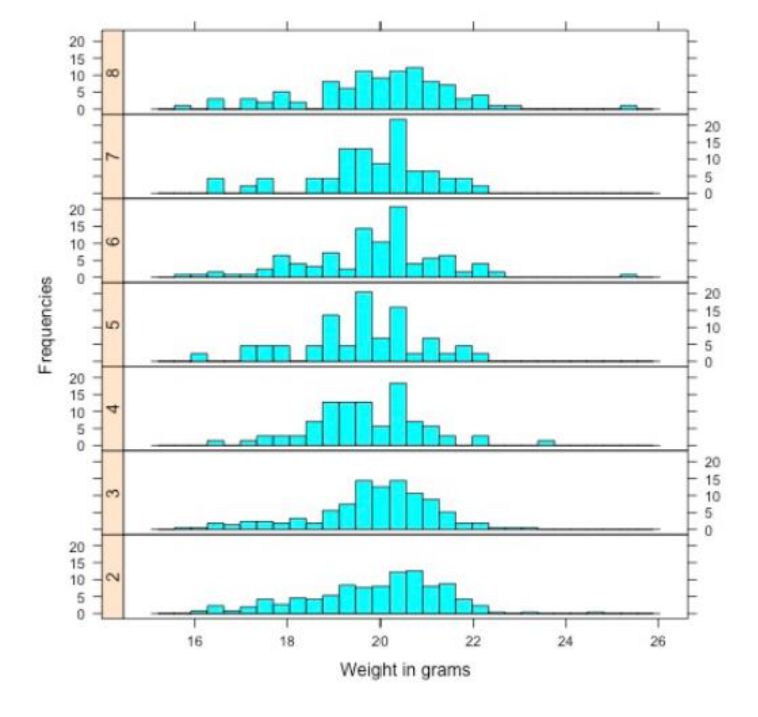


Figure 3.1. A: Simulated dataset visualising the normality assumption. B: Normality at each covariate value.

Normality - histograms

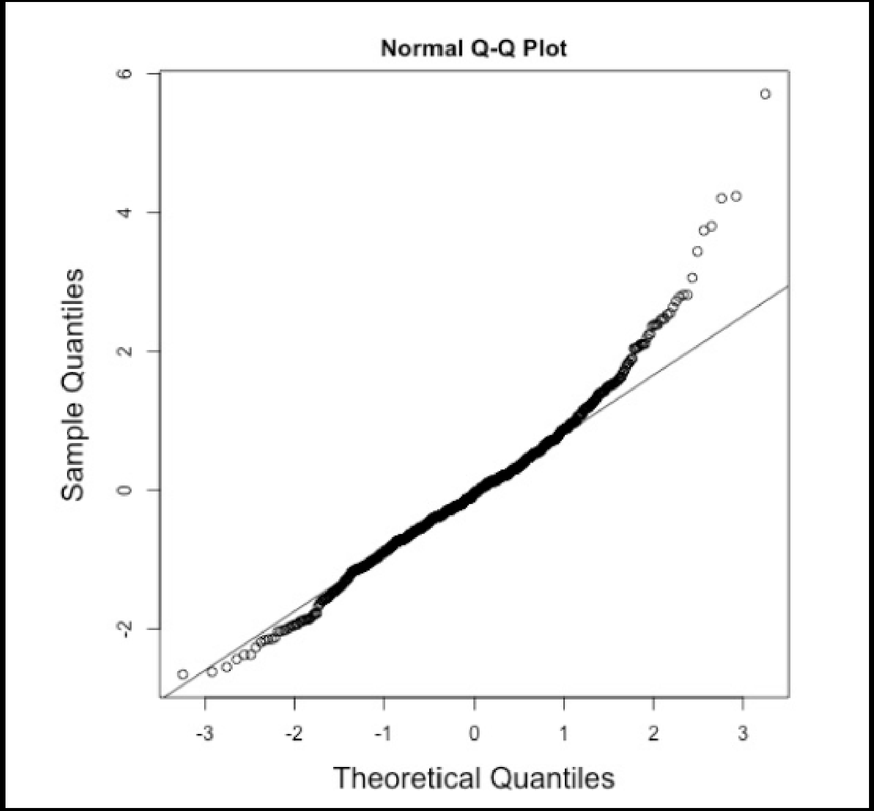
- when you finalize your analysis and do model validation, the residuals should look normally distributed
- Normal distribution is needed for t-test or ANOVA – normality of the data within each group is expected before performing the test





Normality - Quantile-quantile plots

- Q–Q plots plot the quantiles of two distributions versus each other.
- Points on an approximate straight line indicate that the two distributions are similar



Normality - tests

1. Shapiro-Wilks test (shapiro.test)
2. Kolmogorov-Smirnov test (ks.test from pnorm package)
3. D’Agostino test (sgostino.test from moments package)

All of these test depend on sample size. Use graphs to assess normality instead!

Do you need to check normality of covariates in regression? NO

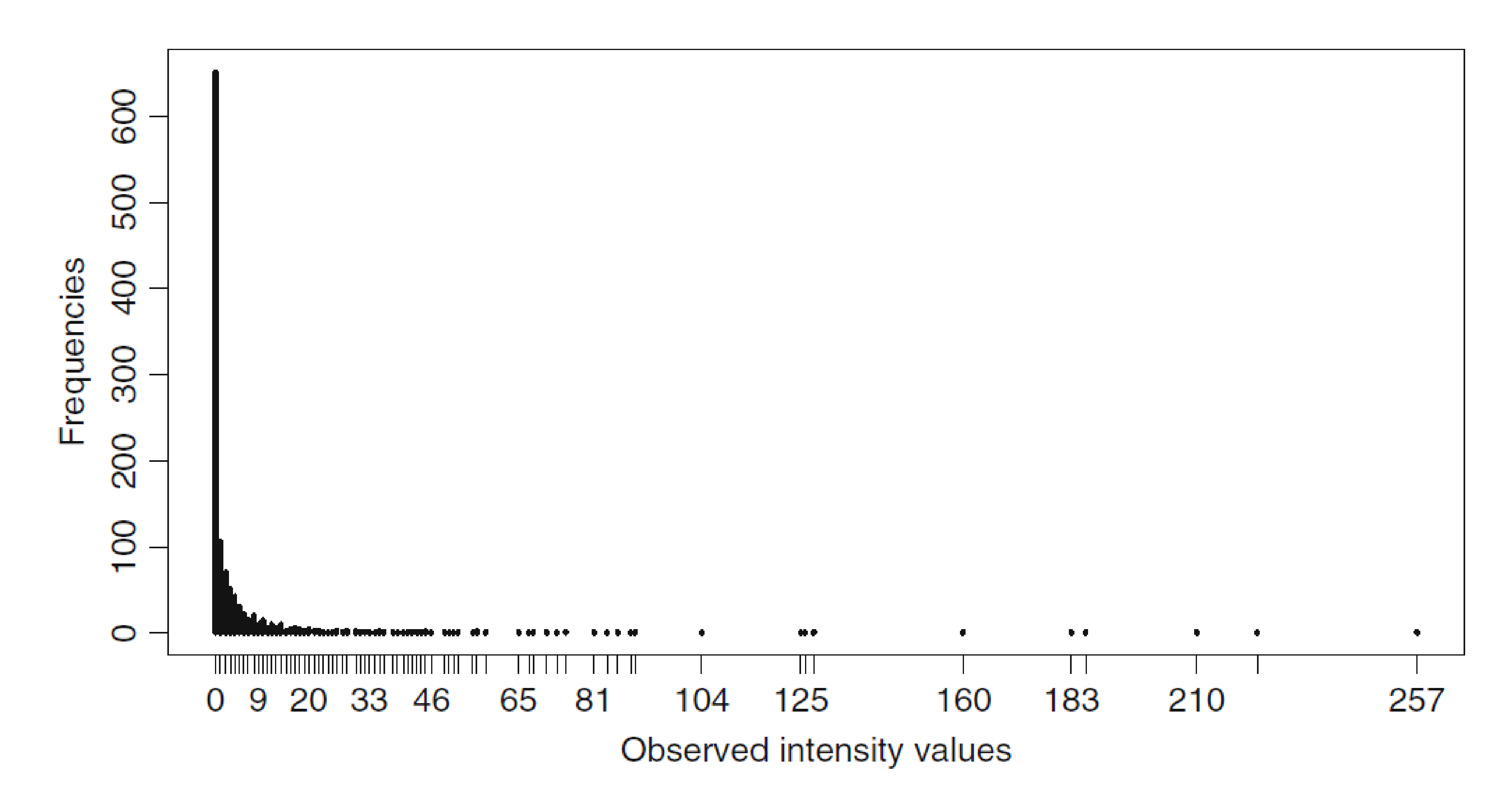
Normality – solutions?

- Linear regression does assume normality BUT is reasonably robust against violation of the assumption (Fitzmaurice et al. 2004)
- Data transformation (but be careful!)

Zeros

- Frequency plot
- Lots of zeros or excess zeros?

Zeros - frequency plots



Zeros – lots or excess?

- Excess = Zero-inflation
  - response variable contains more zeros than expected based on the distribution.
  - Having a lot of zeros doesn't necessarily mean that you need a zero-inflated model.
  - Need to check if count model is over- or under-fitting zeros in the outcome (performance or Dharma package)

Collinearity

- What is it and why is it bad?
- How do we detect it?
- What do we do?

Collinearity

- Correlation between two explanatory variables
- Multicollinearity = correlation between multiple explanatory variables

- If collinearity is ignored:
  - Coefficient estimates are unreliable: standard errors are inflated and p-values get larger making it more difficult to detect an effect
  - Cannot choose which variables are significant

Collinearity

- The sample correlation coefficient falls between -1 and 1
- Pearson, Kendall, Spearman
- Outliers may change the value and sign of the sample correlation coefficient, and therefore the collinearity between covariates. Deal with outlier first!
- Correlations between explanatory variables larger than 0.80 are said to be critical (this is a rule of thumb), but special care also should be taken when intermediate correlation values are found between 0.50 and 0.70
- Interactions = collinearity. Need to center the main terms before calculating interaction term.

Collinearity - pairsplot

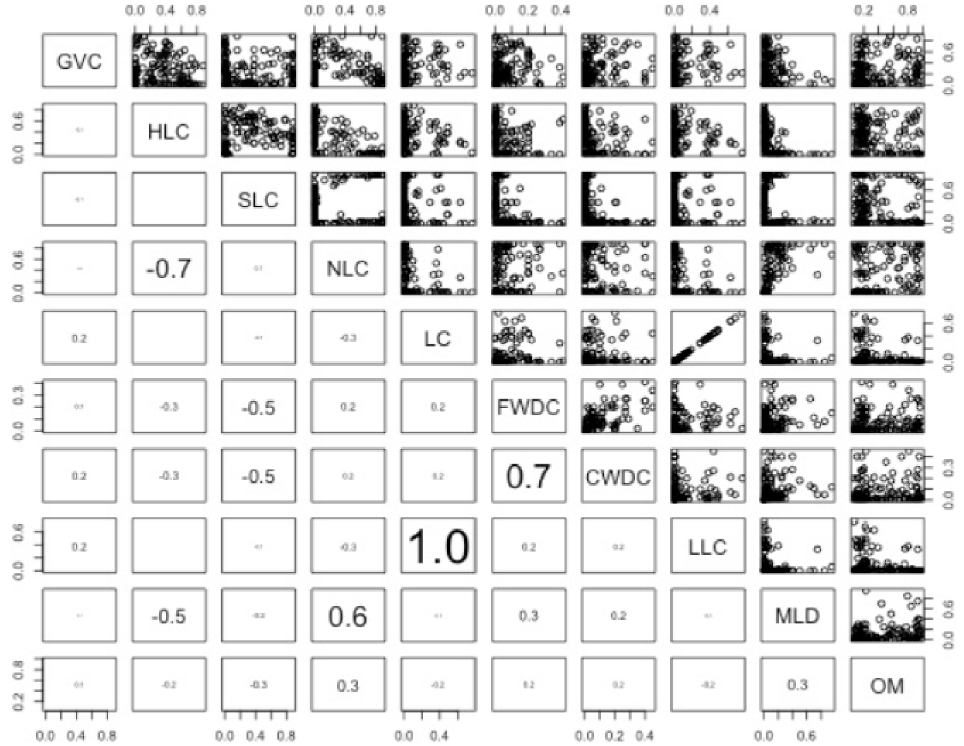


Figure 5.1. Pairplot of selected explanatory variables from the Spider data. The lower panel contains estimated pairwise Pearson correlations and the font size is proportional to the absolute value of the estimated correlation coefficient. The diagonal shows the abbreviations of variables.

Collinearity - boxplots

- run a linear regression model
- If significant effect and, if the model explains more than 10% or 15% of the variation (*R*2), then collinearity is present in your data.

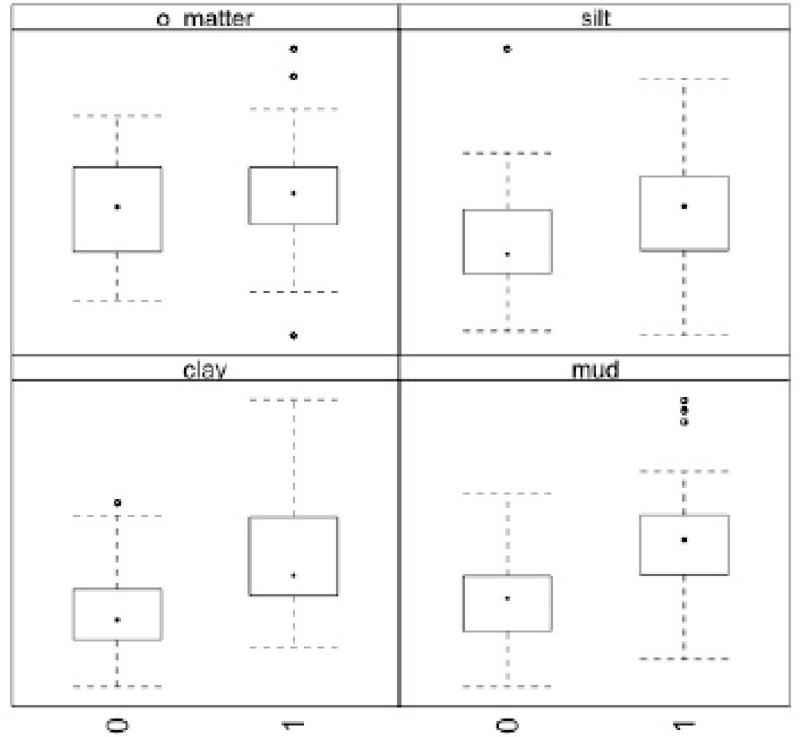


Figure 5.3. Multipanel boxplots showing the relationship between the continuous explanatory variables and the categorical variable CT (0 = no fishing, 1 = fishing).

Collinearity - VIFs

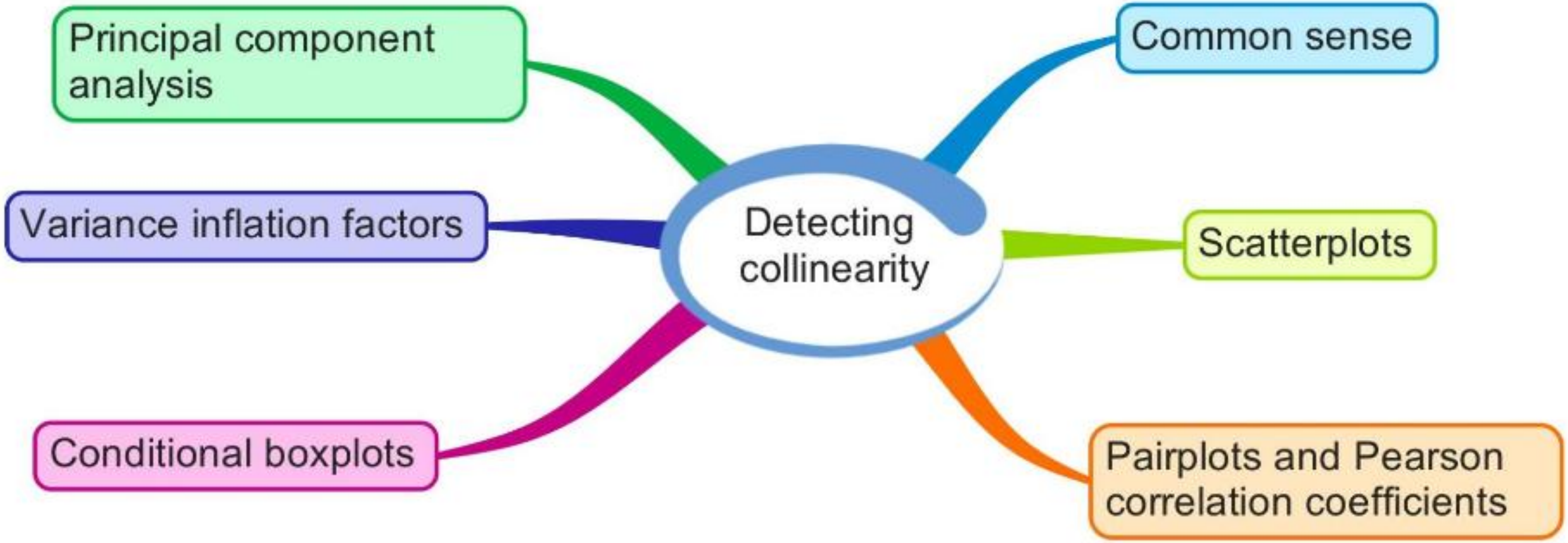
- Variance Inflation factors (VIFs) estimate how much the variance of a regression coefficient is inflated due to multicollinearity in the model.
- VIFs are calculated by taking a predictor and regressing it against every other predictor in the model. This gives you the R-squared value, which can then be plugged into the VIF formula. “i” is the predictor you’re looking at:
- VIF = $1/(1-R^2_i)$
- VIFs range from 1 upwards.
- VIF tells you (in decimal form) what percentage of the variance (i.e., the standard error squared) is inflated for each coefficient.
- VIF> 10 highly correlated - Zuur et al. 2013 recommends 5 or even 3.
- Interactions = collinearity = high VIF

Collinearity – what to do?

Do NOT need to include all recorded explanatory variables in a model

1. Drop covariates
   - Decide which covariate is cheaper
   - Decide which covariate is easier to measure
   - Remember your hypothesis!!
2. Convert them into an index (e.g. PCA)
3. When collinearity is present you will not be able to say which covariate is driving the system – even after a variable is dropped
4. Change experimental design

Our strategy



Detecting collinearity

Principal component analysis

Variance inflation factors

Conditional boxplots

Common sense

Scatterplots

Pairplots and Pearson correlation coefficients

Relationships

- Plotting the response variable versus each of the covariates

Relationships - scatterplots

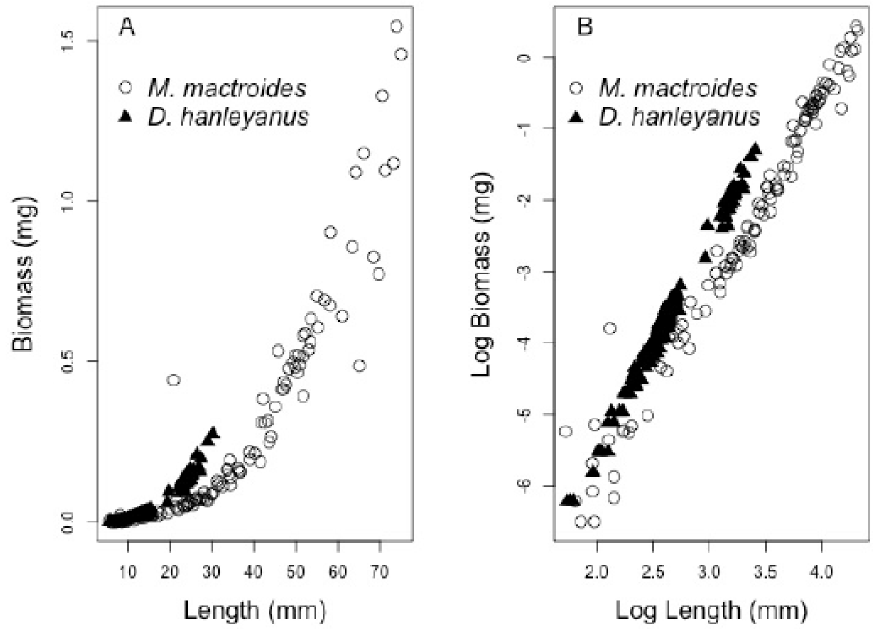


Figure 4.1. A: Scatterplot of adult clam biomass (dry weight) against body length for the yellow clam (*Mesodesma mactroides*) and the wedge clam (*Donax hanleyanus*). B: Same as A, but now both variables are log-transformed.

Scatterplots show that we need mathematical tools that can cope with nonlinear patterns and heterogeneity. Alternatively, you can log-transform the data and apply linear regression.

Relationships - pairplots

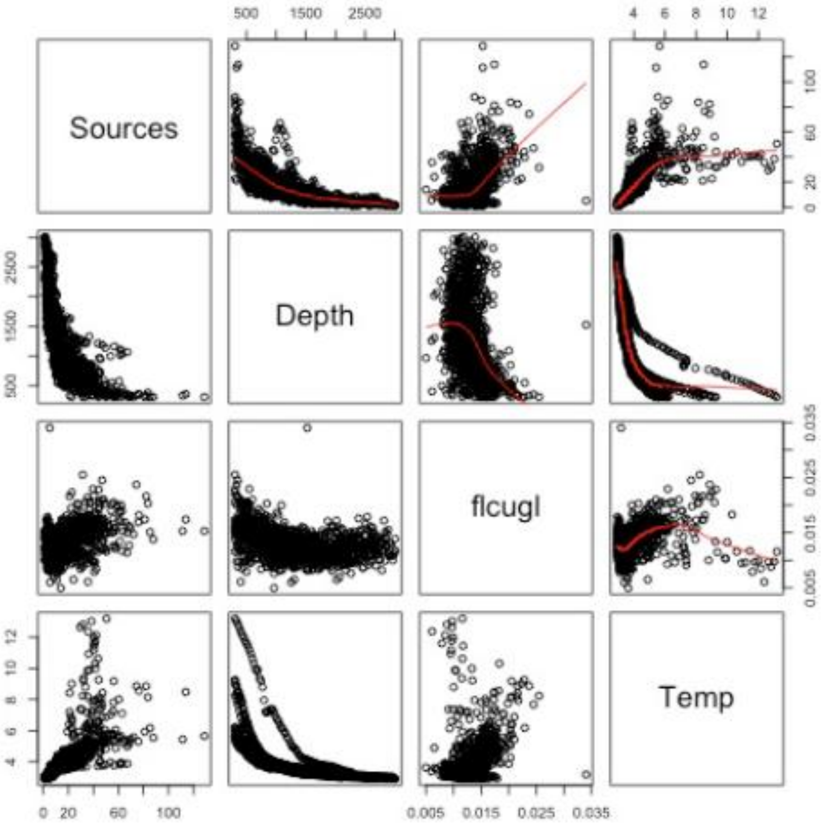


Figure 4.6. Multiple pairwise scatterplots (pairplot) showing the relationship between the abundance (number of sources m^{-3}) of pelagic bioluminescent organisms and environmental variables.

Relationships

- Note that the absence of a clear pattern does not mean that there are no relationships
- A model with multiple explanatory variables can still be a good fit!

Interactions

Interactions: if the *y*–*x* relationship changes depending on *z*

Three main types of interactions:

1. Between a continuous variable and a factor
2. Between two continuous variables
3. Between two factors

Use graphs to determine if we can apply models with interaction terms

Interactions

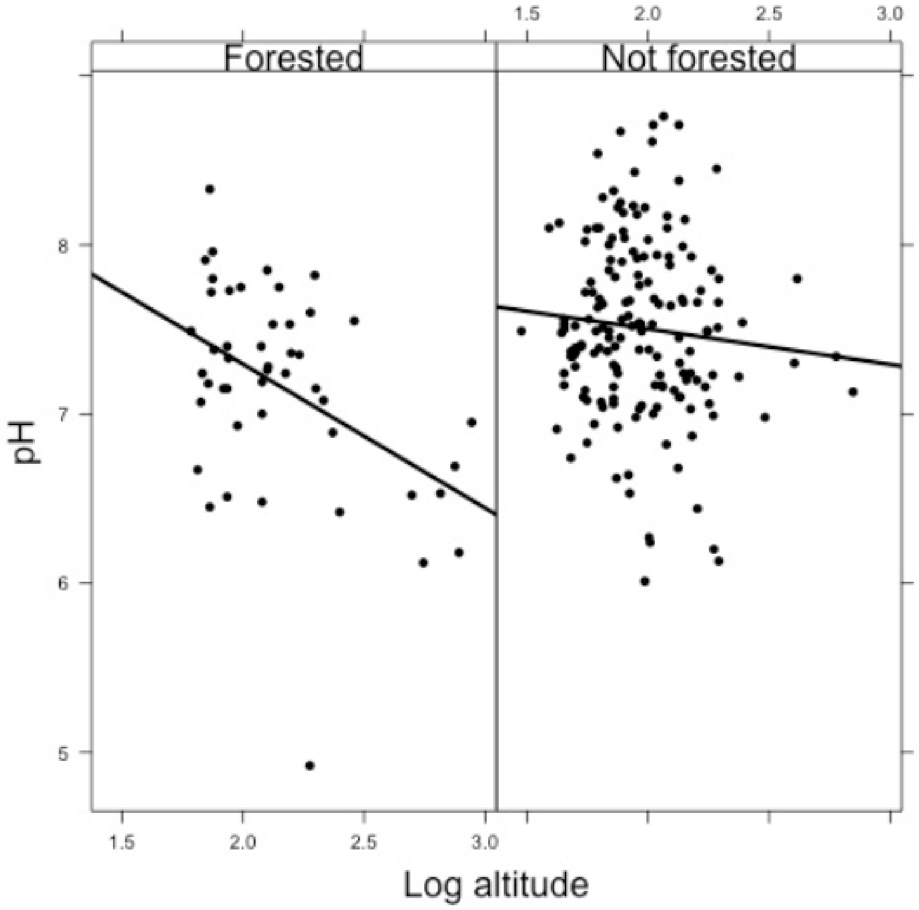


Figure 4.9. Multipanel scatterplot of pH versus log altitude conditional on the categorical covariate fForested.

Interactions

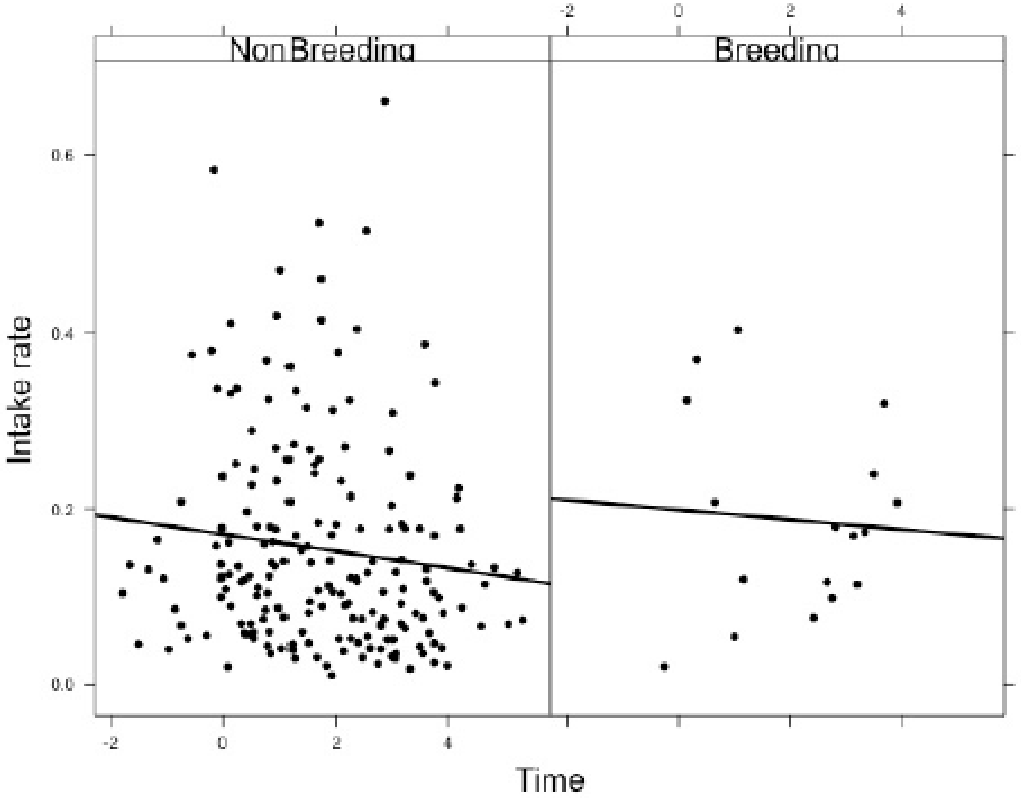
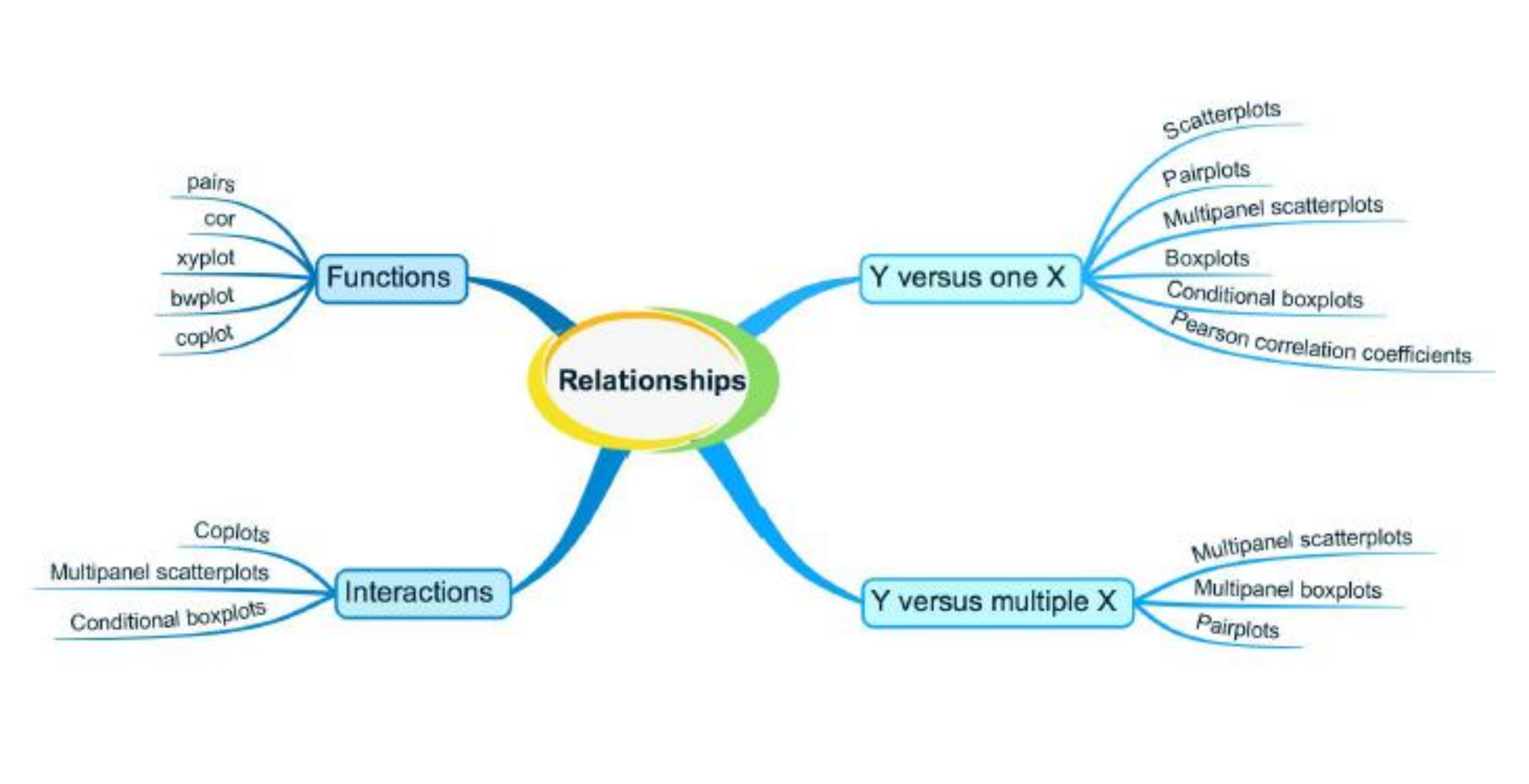


Figure 4.11. Multipanel scatterplot of intake (AFDW per second feeding) versus time (time since low tide in hours) conditional on breeding plumage for the *Limosa* data.

Relationships & Interactions

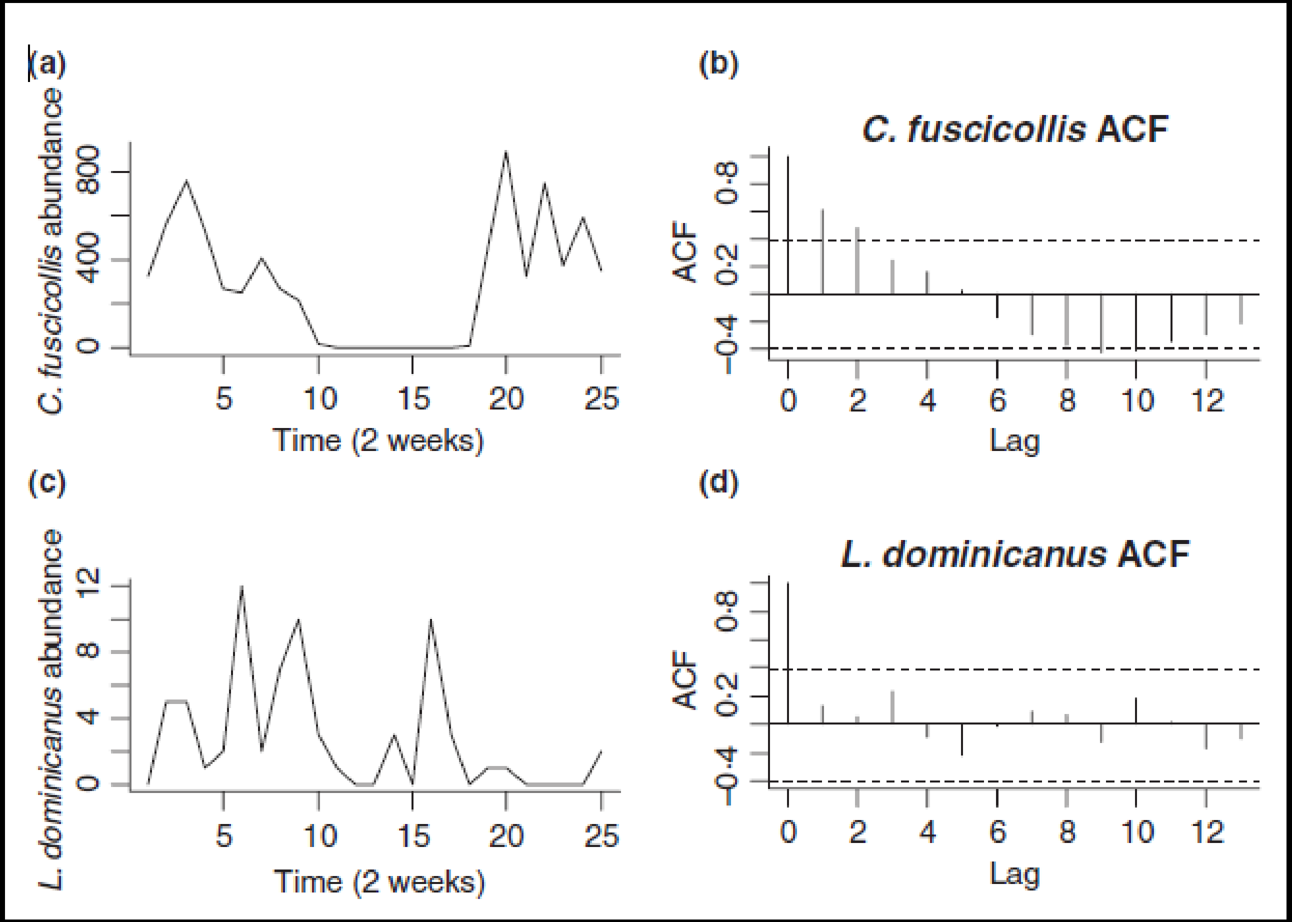


Independence

- Information from any one observation should not provide information on another observation - after the effects of other variables have been accounted for.
- Temporal and spatial dependence: observations closer in space and time are often more related to one another than observations farther away in space and time

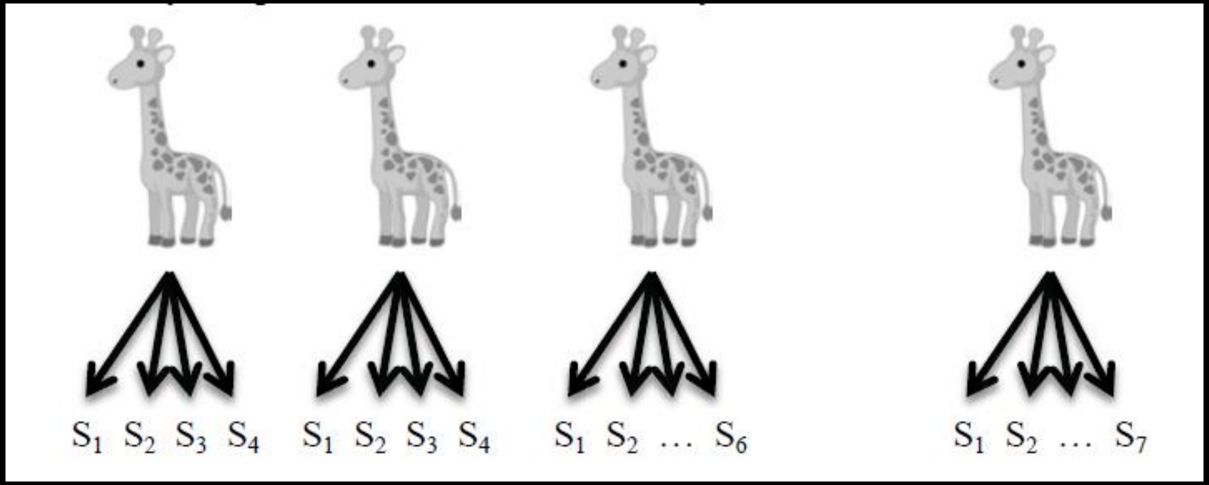
Independence

- Plot the response variable vs. time or spatial coordinates
- ACFs and variograms



Discussion - Independence

- Think about your data set (or the data set you will collect)
- What is your sampling design?
- Is there independence?
- Visualize the structure of your data

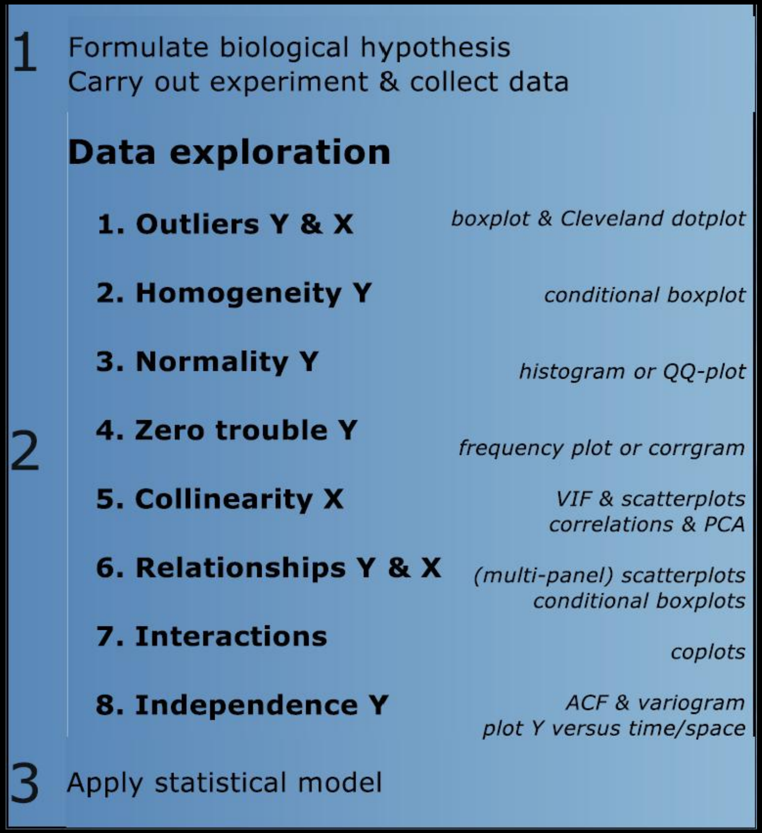


Data transformations

- Transformations - good:
  - Reduce the effect of outliers.
  - Improve linearity between variables.
  - Make the data and error structure closer to the normal distribution.
  - Stabilize the variance
- Transformations - bad:
  - Complicate interpretation (back transformation of estimated parameters)
  - Change the nature of the data (interaction terms change from sig to not)

Summary

- Not every data set requires each step!
- Order also depends on data set
- For some analyses, assumptions can ONLY be verified after the analysis
- Treat list as a series of questions



1 Formulate biological hypothesis
Carry out experiment & collect data

2 **Data exploration**

1. Outliers Y & X	*boxplot & Cleveland dotplot*
2. Homogeneity Y	*conditional boxplot*
3. Normality Y	*histogram or QQ-plot*
4. Zero trouble Y	*frequency plot or corrgram*
5. Collinearity X	*VIF & scatterplots* *correlations & PCA*
6. Relationships Y & X	*(multi-panel) scatterplots* *conditional boxplots*
7. Interactions	*coplots*
8. Independence Y	*ACF & variogram* *plot Y versus time/space*

3 Apply statistical model

Reminders/Updates

1. Thursday:
   - Read: “Zuur 2010 - A protocol for data exploration to avoid common statistical problems”
   - Workshop #2: Data exploration - due the following MONDAY.
2. Submit your ppt ‘paper discussion’ to Canvas
   - end of day Thursday.
   - See the assignment (and PPT) for details/expectations.
3. Office hours: Monday 10:30-11:30 and Thursday 4:00-5:00
4. Complete pre-coarse survey on canvas
5. Assignment #1 DUE SEPT 12