THE UNIVERSITY OF WESTERN AUSTRALIA
Achieve International Excellence

School of Agriculture and Environment & School of Biological Science

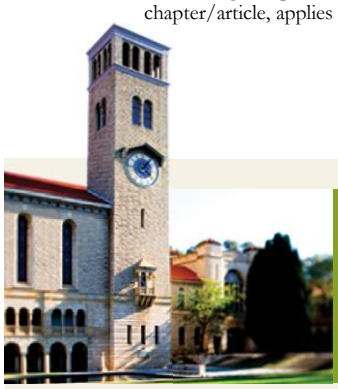
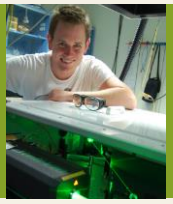
Data Management and Analysis in the Natural Sciences (SCIE4402)

Unit Guide
Semester 2, 2024

Crawley campus and Albany campus

Unit Coordinators:
Michael Renton and James Fogarty

Last revised: 22 July 2024


Introduction
This unit consists of two parts spread across five intensive contact days. Part 1 (Days 1-3) is the same for all students, but for Part 2 (Days 4&5) students select the options most relevant to their research project needs. Part 1 starts by considering the kinds of questions asked in the natural and agricultural sciences, and the types of data needed to address them. Within this context, Part 1 covers development of clear hypotheses or questions, design of experiments or surveys to address these questions, and introduces the statistical software program R as a powerful tool for managing, presenting, and analysing biological, environmental, and economic data. There is a broad overview of some of the analysis methods likely to be of most use to natural scientists, with a more focused coverage of basic approaches such as linear regression and analysis of variance. The emphasis is on the applicability and limitations of different methods in different situations, and the actual application of the methods using the R software.

For Part 2 there are multiple options each day and each option addresses the needs of different research areas within natural and agricultural science. These may include approaches for multivariate species abundance data, species count data, germination data, growth data, time series data, survey data, and more complex experimental designs and use of linear mixed models.

Broad learning outcomes
On completion of this unit you should be able to:

- Develop clear questions or hypotheses regarding natural or agricultural systems
- Design experiments or surveys to collect data to address these questions
- Understand the analysis methods likely to be of most use to natural scientists
- Understand regression and ANOVA approaches as well as more sophisticated approaches relevant to your research area
- Appreciate the applicability and limitations of different analysis methods in different situations
- Use computer software to manage, present, and analyse data appropriate to your research area.

Unit-specific prerequisites: Any Level 1 STAT unit or STAT2210 Biometrics 1 or SCIE1104 Science, Society and Data Analysis or SCIE4401 Data Use in the Natural Sciences or equivalent or approval of unit coordinator. You are also assumed to have some experience with the R software package. If you don't have one of the specified prerequisites then you should discuss with the co-ordinator as soon as possible, and also do the recommended preparation provided on the LMS as soon as possible. It is possible to do the unit without one of the specified prerequisites and recent experience with R, but you must be proactive in your preparation so that you can see whether you will be able to do the unit, or whether it would be better to withdraw before the census date and take the alternative bridging unit instead!

Software Requirements: The core statistics software used during the course is R, which can be freely downloaded from the R website, and installed on any computer. Additionally, you will need to be familiar with Word (or equivalent word processing software) to present your work, and Excel (or equivalent) for some data management activities.

Teaching Staff
Michael Renton and James Fogarty designed the course and are the main teaching staff. They may get help as required through the semester.
James Fogarty: James.Fogarty@uwa.edu.au
Michael Renton: Michael.Renton@uwa.edu.au

Unit co-ordinators: James Fogarty and Michael Renton

Consultation hours: Students are ***strongly*** urged to make optimal use of staff availability in scheduled lab and discussion times, and online LMS forums. Email should only be used for personal/private enquiries that are not relevant to other students. LMS forums will usually be answered more promptly

than email, both lecturers can answer as relevant, and all students can benefit from the response. If you email regarding a private matter, and don't receive a response within a couple of days, then post a short reminder to the LMS forum asking us to check for your email. For personal/private enquiries related to specific course assessments, email the relevant lecturer. For more general personal/private enquiries, email both.

Unit structure and schedule

The unit includes six scheduled contact days. The first day includes just a single introductory lecture/seminar. The other five days are the five intensive contact days, which include a Q&A discussion session and computer labs. You might like to cross-check against the online UWA timetable and let us know if there are any discrepancies.

Contact Day	**Date**
Intro Lecture	25th July **
Day 1	8th Aug
Day 2	22nd Aug
Day 3	12th Sept
Day 4	26th Sept
Day 5	10th Oct

* 9-10am intro lecture and discussions are scheduled for the AGRI: [G013] Agriculture Lecture Theatre; 10-12 and 1-3 labs are scheduled for BOBIA1: [G11] East Undergrad Computer Lab. Venues are subject to change; please check LMS for updates. All classes are also available for online attendance (see below).

** There are no scheduled labs on this first day, just the intro lecture – but you are strongly encouraged to use the labs time for revision if needed (see LMS for suggestions).

Introductory Lecture

The course approach, structure, components, requirements, and assessment will be explained, and students will have a chance to ask questions about the first assessment quiz. Make sure you attend or if you miss it for some reason watch online as soon as possible!

Lectures: The rest of the lecture content (apart from the introductory lecture) is delivered via online recordings that are available via the LMS/ LCS. It is essential that you watch the lecture recordings prior to the Q&A and computer lab sessions.

Q & A sessions: There is a Q&A session at the start of each contact day; this provides an opportunity to discuss the topics covered in online lectures and/or the completed assessments, ask questions, and seek clarification on any aspects you may still be having trouble with.

Computer labs: These sessions are an opportunity to work through structured example questions at a time when you can ask questions of a lab demonstrator. All computer lab example questions and data sets can be downloaded from the LMS. After the intro lecture all students will be allowed to self-allocate to either the morning or afternoon lab slot for priority in-person attendance. If/when numbers allow you are welcome to attend both the morning and afternoon labs in person (recommended) or online. If you find yourself struggling with course material then you should aim to attend all scheduled computer labs (all four hours on each scheduled day) if at all possible in order to get maximum help. If you have clashes with other courses, then aim to attend as much as possible and plan to catch up in your own time as needed.

Online attendance: The intro lecture, all labs, and all discussion seminars will be available for online attendance via Teams.

Part 1 Topics
For part 1, all students cover the same material.

Day 1: R basics, graphing, hypothesis testing, and classical tests
We will learn the basics of the R software package and start to apply it to data exploration, manipulation, and plotting. Following the development of data exploration skills we then consider some of the simpler classical statistical tests, including proportion, binomial, chi-squared and t-tests. We will also revise some of the basic statistical concepts such as populations, samples, and hypothesis testing.

Day 2: ANOVA, ANCOVA, linear regression: same family, different name
One-way ANOVA, simple linear regression, factorial ANOVA, multiple regression, ANCOVA: they may sound different but they are all linear models. We will revise the differences and similarities between these methods and learn how they can all be addressed in similar ways within R. We will also touch on nesting, blocking, pair-wise comparisons and contrasts, and data transformations.

Day 3: When the model assumptions do not hold
When using the ANOVA method or linear regression models to investigate research questions it is necessary to make a number of assumptions. For any given real world data set it is often the case that one or more of the key model assumptions do not hold. The focus of the material on Day 3 is how to test whether the model assumptions hold; and practical strategies that can be used when the model assumptions do not hold. We also introduce R packages, which are user contributed R functions that make it possible to implement relatively complex tests and adjustments for assumption violations with ease.

Part 2 Topics
For part 2 (Days 4&5), there are multiple streams on each day. Students select the stream that seems most relevant to their research area or interests on each of the days. Each choice is designed to stand alone, so choices for Day 5 do not depend on choices for Day 4.

Day 4 Stream A: Extensions to linear models for ecological data
Ecology often involves counting the numbers of different species or recording how many individuals lived or died, or germinated or didn't. Generalised linear models (GLMs) are an extension to linear models that allows us to deal with this kind of count, binomial, or survival data relatively easily. This module will cover the theory and practice of using GLMs to analyse ecological data. Topics covered include binomial and Poisson GLMs, testing for and dealing with over-dispersion, and model selection and simplification based on AIC. We also introduce mixed-effects linear models and how they can be used to analyse more complex experimental designs that include blocking and nesting.

Day 4 Stream B: Extensions to linear models with applications in Natural Resource Management, Environmental Management, and Applied Economics
The day starts by introducing the methods that can be used to model choices, where the context is selecting between two alternatives. The material in the day then builds to cover the methods that can be used to value natural resources such as native parks and other recreational areas. These methods involve analysis of survey data, where we collect information on things like how far someone travelled to visit a park or recreation area, or how they selected a specific recreation area from multiple possible recreation sites. Another type of survey response we can use to estimate natural resources values is the answer to questions such as 'how much would you be willing to pay to improve the facilities at a national park?' We also model this type of survey data. In these models a key issue is that the dependent variable is either yes/no, which we code as zero/one, or a count measure. Such data require use of generalised linear models.

Day 5 Stream A: Multivariate techniques for ecological data
Ecology often involves big multivariate data sets, such as counts or other abundance measures of lots of

different species at lots of different sites, or multiple measurements on different individuals. Multivariate techniques provide means to visualise and analyse these kinds of complex data sets, both when there are many explanatory variables, or many variables to be explained. Topics covered include Principle Component Analysis, non-metric Multi-dimensional scaling, ANOSIM, distance-based redundancy analysis (PERMANOVA/DISTLM) and multiple regression.

Day 5 Stream B: Mixed effects, variance components, panel data models, and experimental designs approaches

Where there are measurements on multiple items through time, say the growth of ten trees through time, regression models can be used that take advantage of both the cross sectional nature of the data -- ten values at one point in time -- and the time series nature of the data -- measurements on the same tree through time. These models, and associated estimation issues are covered. The same techniques used for repeat measures through time regression can also be applied in the experimental design context, and so we also review classical approaches to experimental design and then look at modern estimation techniques to these problems. Some more complicated ANOVA models are also considered.

Day 5 Stream C: Introduction to time series data

There are many approaches to modelling time series data. The primary focuses of this module is to introduce some of the workhorse models used to model time series data. The first set of models considered are extensions of the standard linear regression modelling framework, and include the partial adjustment model and the distributed lab model. Both these models give rise to estimates of the short run effect or immediate period effect and a long run or equilibrium effect. The final traditional model considered is the autoregressive distributed lag model. This model can be used to model both stationary (non-trending) and non-stationary (trending) data. Time series data are also used in forecast models. Forecast models typically use only the historical sample data to *forecast* future values. These models typically do not have explanatory variables, and so are philosophically different to traditional regression models. Some useful forecasting methods will be introduced, with a focus on applying the different methods rather than the mathematics behind these models. Using forecast methods involves the researcher making subjective choices, and some automated methods that help the user get started are also introduced.

Assessments
The assessment tasks focus on applied skills and general understanding rather than detailed understanding of theoretical concepts or technical details. All assignments and quizzes are to be completed/submitted in the format specified by the relevant instructor via LMS. Some assessments are submitted online through an LMS 'quiz', where you enter multiple choice or numerical answers to specific question parts, while others are submitted as written reports.

The unit uses a regular ongoing assessment approach, to help ensure that students are staying on track. The assessments include an initial quiz (worth 5%), five assignments (worth 16% each, or 80% in total), and a final written reflective report with two parts (7.5% each part, or 15% in total). The initial quiz is an introductory/revision quiz, generally due two days *before* the first intensive contact day. The due date for assignments is generally two days before the next intensive contact day (usually 12 days after the relevant intensive contact day). The due date for the written reflective report will generally be in the first week of the final exam period. Assessments will generally be due at the end of the day. You can confirm these dates and times for submissions for each individual online assessment task as it appears on the relevant section of the LMS. If there is any disagreement between the LMS and this guide, then post a query to the LMS forum ASAP.

Assessment	**Due Date**	**Weighting**
Intro/Revision Quiz	6th Aug	5%
Assignment 1	20th Aug	16%
Assignment 2	10th Sept	16%
Assignment 3	24th Sept	16%
Assignment 4	8th Oct	16%
Assignment 5	27th Oct	16%
Written reflective report (Part A & Part B)	1st Nov	15% (7.5% each)

Scaling and moderation: As there are different streams within the unit that have different assessment tasks, marks for individual assessment items may be adjusted to standardise across streams. All processes used are compliant with the UWA assessment policy and this means that on rare occasions adjustment to marks can take place at the Board of Examiners meeting. It is important to realise that on rare occasions the marks showing on the LMS during semester may not be your final mark for that assessment item.

Late submission of assessments: The university policy for late assignment is to apply a late penalty deduction of 5 percentage points per day, for the first seven days, including weekends and public holidays. For example, if an assignment is submitted one day late and the assignment receives a mark of 73 percent before the late penalty, the mark after the late penalty will be 68 percent. Assignments submitted more than seven days late, without an application for mitigating circumstances, receive a mark of zero. Usually it will be possible to submit online assessments late, but you then need to contact the relevant lecturer and request that the assessment is marked and any relevant late penalty to be applied, before any mark will be awarded. If you have a compelling reason for an extension then you should submit an application for special consideration to the student office with evidence (e.g. a medical certificate or letter from supervisor explaining extended field work for your research project); if this is approved then no late penalty will apply.

Missed Assessment: If you miss an online quiz or online assignment a mark of zero will be recorded unless formal special consideration has been approved, in which case assessment weights may be adjusted to take this into account.

Final Exam: There is no final exam for this unit. Yippee!

Plagiarism: All forms of cheating, plagiarism, and copying are condemned by the University as unacceptable behaviour. The Faculty's policy is to ensure that no student profits from such behaviour. Generally a failure will be recorded for the subject in which the cheating has occurred. Serious cases shall be referred to the University's Board of Discipline. All students should note that cases of copying automatically trigger a reporting and action process. Students should note specially that the online assessments used in this unit are designed to detect plagiarism and copying, and that plagiarism has been detected and penalised in the past.

References, resources and reading materials

As this is an honours level unit we expect you to take ownership of the learning experience. In addition to the important resources we provide on the LMS, there are many statistics textbooks available through the UWA library as ebooks and/or hard copies. We particularly recommend textbooks that teach both statistics and R at the same time. Searching the internet for 'R intro stats books' brings up several good books that do exactly this, and that are also available in the UWA library.

We hesitate to recommend any particular books, as different styles will appeal to different people, and to people with different backgrounds and experience, and new books are coming out all the time. However, a couple of good old books with a bit of ecological flavour that have stood the test of time (reasonably well?) are these two by Michael Crawley. One is focused more on statistics and one more on R.

Statistics: An Introduction using R by **Michael Crawley**
Like it says – a very good introduction to statistic using R.

The R Book by **Michael Crawley**
A full practical introduction to R and many statistical methods within R – includes chapters on count data and proportion data (GLMs), mixed effect models, and some multivariate techniques among many other things.