Advanced Biometrics, Bangor University
Course Materials 2008

Instructor: Dr. Valerie LeMay , Professor in Measurements/Biometrics
EMAIL: Valerie.LeMay@ubc.ca Home University: Faculty of Forestry, University of British Columbia, Vancouver, BC, Canada

Course Objectives and Overview:

The objectives of this course are:

1. To be able to use simple linear and multiple linear regression to fit models using sample data;
2. To be able to design and analyze simple lab and field experiments;
3. To be able to interpret results of model fitting and experimental analysis; and
4. To be aware of other analysis methods and more complex experimental designs not explicitly covered in this course.

In order to meet these objectives, background theory and examples will be used. Analyses will be done using a mixture of:

1. "Hand" calculations, using calculators and/or spreadsheet packages such as EXCEL.
2. R code. This is a free software package with many built in functions for statistical analyses and graphs.

Evaluation:

There is no formal evaluation for the course.

Course Content Materials:
Course materials for this class can be found on www.forestry.ubc.ca/biometrics website and then click on Advanced Biometrics.

Recommended Textbooks:
There are many other textbooks on using regression to fit models, and on experimental design. When choosing a textbook as a reference for your future use, you may wish to choose texts that have biological examples, in particular. Here are a few recommendations for textbooks and other materials.

Introductory books:
Freese, F. 1963. Elementary statistics for foresters. USDA Forest Service (See www.forestry.ubc.ca/biometrics and then "Links" to find this short book). [This gives an overview of statistics that should be familiar to students at the beginning of this course, and includes forestry examples. All examples are given using imperial units, rather than metric units, since this is a very old book. However, the methods used are still relevant.]

Linear Models:
Class notes for FRST 430/533 taught at the University of British Columbia www.forestry.ubc.ca/biometrics and then click on the course number for the lecture notes

Experimental Design:

1. LeMay, V. and A. Robinson. 2004. Inventory: Design, performance and analysis of experiments. J. Burley, J. Evans, and J.Z. Youngquist, eds. Elesvier Academic Press, San Diego 1: 158-163.
2. Biometrics Pamphlets by the British Columbia Ministry of Forests (www.forestry.ubc.ca/biometrics and then **click on "links**" to find the biometrics pamphlets by the British Columbia Ministry of Forests, Research Branch) [excellent with a variety of forestry examples for each experiment]

Using R:

1. Crawley, Michael J. 2005. Statistics: An introduction using R. Wiley. This book covers simple analyses using R, including multiple regression. There is very little on experimental design, however. The examples are easy to follow.
2. Robinson, A. 2008. IcebreakerR. This is an excellent introduction to R and fitting models and obtaining graphs using R. Dr Robinson teaches and does research at the University of Melbourne, Australia see: http://www.ms.unimelb.edu.au/~andrewpr/

Other Reference Materials:

Hicks, C.R. 1993. Fundamental concepts in the design of experiments, 4th edition. Saunders College Publishing, Toronto.

Kutner, M.H., C.J. Nachtsheim, J. Neter, and W. Li. 2005. Applied linear statistical models, 5th edition. McGraw-Hill Irwin. [This is a comprehensive textbook covering linear models including regression analysis and experimental design. This is a very good reference for the topic, but may be difficult if this is your first experience in using linear models.]

Kleinbaum and Kupper, Applied regression analysis, Duxbury Press. [This text is a bit easier to read than the Kutner et al. book, but is therefore less comprehensive.]

Dalgaard, P. 2002. Introductory statistics with R. Springer. [This book is a little harder to follow than the book by Crawley, but may be preferred by some students.]

NOTE: The schedule will likely change somewhat, as it depends upon the background and interests of students.

Day 1: 4 hours + 2 hours of homework

Short Review (Revision) of Probability and Statistics

Lecture (classroom; 1 hour):

- Descriptive statistics
- Inferential statistics using known probability distributions: normal, t , F, Chi-square, binomial, Poisson
- Examples

Practical Exercises (classroom or computer lab; 1 hour):

Exercise 1: Basic statistics, "hand" calculations

Fitting Linear Models

Lecture (classroom; 2 hours):

- Dependent variable and predictor variables
- Purpose: Prediction and examination
- General examples
- Simple linear, multiple linear, and *nonlinear regression*
- Objectives in fitting: Least squared error *or Maximum likelihood*

Simple Linear Regression:

Definition, notation, and example uses

- dependent variable (y) and predictor variable (x)
- intercept, and slope, predicted y variable ($\hat{y}$) and error
- Graphs

Least squares solution to finding an estimated intercept and slope

Derivation, normal equations, equations for intercept and slope for one predictor variable

Assumptions of linear regression and properties when assumptions are met

- Residual plots to visually check the assumptions that:
  - 1. Relationship is linear MOST IMPORTANT!!
  - 2. Equal variance of y around x values (equal "spread" of errors around the "line")
  - 3. Observations are independent (not correlated in space nor time)
- Normality plots to check assumption that:
  - 4. Normal distribution of y around x value (normal distribution of errors around the "line")
- Sampling and measurement assumptions:
  - 5. x values are fixed
  - 6. random sampling of y occurs for every c x value
- Properties when all assumptions are met versus some are not met

Transformations and other measures to meet assumptions

- Common Transformations for nonlinear trends, unequal variances, percents, rank transformation
- Outliers: unusual observations

Measures of goodness-of-fit

- Graphs
- Coefficient of determination (r^2) [and Fit Index, I^2]
- Standard error of the estimate (SE_E) [and SE_E']

Estimated variances, confidence intervals

- For the intercept and slope
- For the mean of the dependent variable given particular values for the x variable

Hypothesis tests

- Is the equation significant?

Multiple Linear Regression:

- dependent variable (*y*) and more than one predictor variable (*x*)
- intercept, and slopes
- Least squares solution to finding an estimated intercept and slopes
- Assumptions and transformations
- Measures of goodness of fit (R^2 and SE_E)
- Estimated variances, confidence intervals
- Hypothesis tests: 1. Is the equation significant? 2. Is each variable significant, given the other variables?
- Tools to help select x variables

Other methods: *nonlinear least squares, weighted least squares, general least squares, general linear models*

Practical exercises (homework; 2 hours): Exercise 1 (con't) Hand calculations of intercept, slope, standard errors for estimated intercept and slope (s), and confidence intervals, etc. for a simple linear regression to compare to Day 2 R results.

Day 2: 4 hours

Practical Exercises (Computer Lab; 1 hour): Exercise 2
Redo your basic statistics and your fit of your simple linear regression model using R—transforming data, graphs, fitting equations, assessment of assumptions, transformations and refitting, goodness of fit, confidence intervals and hypothesis tests. Compare to hand calculations

Lecture (Computer lab; 0.5 hours)
Discussion of R outputs

Practical Exercises (Computer Lab; 1 hour): Exercise 3:
Multiple linear regression using R —transforming data, graphs, fitting equations, assessment of assumptions, transformations and refitting, goodness of fit, confidence intervals and hypothesis tests.

Lecture (classroom or computer lab; 0.5 hours)
Discussion of Exercise 2

Lecture (classroom or computer lab; 1 hour): Other topics.
Adding class variables as predictors
- *Dummy variables to represent a class variable*
- *Interactions to change slopes for different classes*

Methods to aid in selecting predictor (x) variables

- All possible regressions
- Stepwise methods

Selecting among alternative models

- Process to fit an equation using least squares regression
- Meeting assumptions
- Measures of goodness-of-fit: Graphs, Coefficient of determination (r^2) or I^2, and Standard error of the estimate (SE_E) or SE_E’
- Significance of the regression, and variables
- Biological or logical basis
- Parsimony and cost

Day 3: 4 hours

Practical Exercises (lab; 1 hour): Complete (or practice) any R exercises from Days 1 and 2.

Practical Exercises (lab; 1 hour): Exercise 4: Use R and stepwise methods to select possible x variables. Do a full regression analysis on the selected variables. Discussion of results.

Practical Exercises (lab; 1 hour): Exercise 4a: Use R to get a regression by species. Subset the data and do regression analysis on each subset.

Practical Exercises (lab; 1 hour): Exercise 5: Use R for a number of graphs using a large (500 trees) dataset.

Day 4: 5.5 hours

Experimental Design and Analysis

Lecture (classroom; 0.5 hour)

- Sampling versus experiments
- Definitions of terms: experimental unit, response variable, factors, treatments, replications, crossed factors, randomization, sum of squares, degrees of freedom, confounding, samples
- Variations in designs: blocking, number of factors, fixed versus random effects, split-plot, nested factors, subsampling, covariates
- Main questions in experiments for fixed versus random effects

Lecture (classroom; 1 hour)

Completely Randomized Design (CRD**)**

Definition: no blocking and no splitting of experimental units

One Factor Experiment, Fixed Effects

- Main questions of interest
- Notation and example: observed response, overall (grand mean), treatment effect, treatment means, error
- Data organization and preliminary calculations: means and sums of squares, box plots
- Assumptions regarding the error term: independence of observations, equal variance among treatments, normality within treatments

- Transformations if assumptions are not met
- Test for differences among treatment means using one-way analysis of variance (ANOVA) : degrees of freedom, mean squares, F-test
- If there are differences, which treatments differ, using pairs of means t-tests and correction of alpha
- Confidence intervals for treatment means

Power of the test, expected values under the assumptions

<u>Lecture (classroom; 1 hour)</u>

<u>Two Factor Experiment, Fixed Effects</u>

- Introduction: Separating treatment effects into factor 1, factor 2 and a possible interaction between these
- Example layout
- Notation, means and sums of squares calculations
- Assumptions, and transformations
- Test for interactions and main effects: ANOVA table, expected mean squares, hypotheses and tests, interpretation
- Differences among particular treatment means
- Confidence intervals for treatment means

<u>Two Factor Experiment, One Fixed and One Random Effect</u>

- *Explanation*
- *Example layout*

Practical Exercises (lab; 1 hour): Exercise 6: Use R to repeat your calculations for the one factor experiment. Include a box plot, and simple statistics for each treatment. What are your conclusions based on the descriptive and inferential statistics? Were the assumptions of one way analysis of variance met?

Lecture, lab; 0.5 hours):
Discussion of Exercise 6 results

Practical Exercises (lab; 1 hour): Exercise 7: R calculations for two factor experiment, both are fixed-effects factors. What does this mean? How would you extend your analysis to three fixed-effect factors?

Lecture (lab; 0.5 hours)
Discussion of Exercise 7: CRD with two fixed-effects factors

Day 5: 4 hours

Restrictions on Randomization

Lecture (classroom; 1.5 hours)

Randomized Block Design (RCB) with one fixed factor

Introduction, example layout, data organization, and main questions

- Notation, means and sums of squares calculations
- Assumptions, and transformations
- Differences among treatments: ANOVA table, expected mean squares, hypotheses and tests, interpretation
- Differences among particular treatment means
- Confidence intervals for treatment means

Randomized Block Design with other experiments

- RCB with replicates in each block – Generalized RCB
- Two fixed factors
- One fixed, one random factor

Incomplete Block Design

- Definition and Example

Latin Square Design: restrictions in two directions

Definition and examples

Split Plot Design

- Definition and examples
- Notation and assumptions
- Model and analysis of variance table

Practical Exercises (classroom; 1 hour): Exercise 8. Using R for RCB one factor experiment. 1. How do these differ from the two factor CRD as far as (1) calculations; and (2) interpretation of the results? Discussion of results

Lecture (classroom; 1.0 hour)

Nested and Hierarchical Designs

CRD: Two Factor Experiment, Both Fixed Effects, with Second Factor Nested in the First Factor

- Introduction using an example
- Notation
- Data organization
- Modification for RCB

CRD: One Factor Experiment, Fixed Effects, with sub-sampling

- Introduction using an example
- Notation
- Analysis methods: averages, least squares, *maximum likelihood*
- Data organization and preliminary calculations: means and sums of squares
- Modification for other designs: e.g., RCB

Adding Covariates (continuous variables)

Analysis of covariance

- Definition and examples
- Notation and assumptions
- Expected mean squares
- Hypotheses and confidence intervals for main questions if assumptions are met
- Allowing for Inequality of slopes

Lecture (classroom; 30 minutes)

Summary