Weekly Learning Objectives - BIO144, Data Analysis for Biologists

Owen Petchey & Stephanie Muff

2/3/2017

Contents

Overarching learning objectives **1**

Nested learning objectives **2**

Unit 1 - Introduction **2**

Unit 2 - Linear Regression **2**

Unit 3 - Multiple linear regression part 1 **3**

Unit 4 - Multiple linear regression part 2 **3**

Unit 5 - ANOVA **3**

Unit 6 - ANCOVA & Matrix alegbra **4**

Unit 7 - Model selection **4**

Unit 8 - Interpretation, causality, and cautionary notes **4**

Unit 9 - Analysing count data **4**

Unit 10 - Analysing binary data **5**

Unit 11 - Measurement error; repeated measures and random effects **5**

Unit 12 - Recap and outlook **6**

Overarching learning objectives

By the end of the course you will be able to:

- Plan how to make use of quantitative data to solve biological problems.
- Translate a biological question into a quantitative problem.
- Collect and arrange data for efficient processing.
- Reliably, accurately and efficiently manage and manipulate data.
- Make clear and informative visualisations of data.
- Select, perform, validate, and interpret an appropriate statistical model or test.
- Understand why and when linear models are useful, and what to do when they are not.
- Clearly communicate the answer to your biological question.
- Research and learn about other tools for data analysis.
- Recognise the limitations of data from experimental and observational studies.

Nested learning objectives

Below you will see that some learning objectives refer to sets of previous ones. E.g. "You will be able to do the same as before, but with a difference of some kind". Be sure to cover all of the previous learning objectives referred to.

Unit 1 - Introduction

By the end of this week you will be able to:

- Describe the aims, importance, and applications of data analysis in biology and biomedicine.
- Recall and use previous relevant learning.
- Choose graphical tools appropriate to question and data.
- Describe what is a model, and what it is not.
- Identify important features of data, such as skew and correlation.
- Describe the equation for a straight line.
- Recall the general workflow for data analysis.
- Help yourself when R / RStudio.
- Be able to work with add-on R packages.
- Fix simple errors in R code.
- Confidently and reliably read a data file into R / RStudio.
- Use the dplyr functions `select`, `slice`, `mutate`, `filter`, `arrange`, `group_by`, and `summarise`.
- Make simple graphs using the ggplot function.
- Relate the difference between statistical and biological significance.

(Please note that these skills are assessed in the Graded Assessment 0 & 1)

Unit 2 - Linear Regression

By the end of this week you will be able to:

- Describe the type of biological / biomedical question that linear regression could help answer.
- What we should have done, before we do a linear regression.
- Understand how the intercept and slope are estimated in linear regression.
- Fit and interpret a linear regression model to data in R / RStudio.
- Recall the assumptions of linear regression.
- Evaluation the assumptions of normality and of independence of residuals.
- Interpret the biological meaning of the estimated parameters.
- Quantify how good is the linear regression model.
- Perform a relevant hypothesis test and get the confidence interval of the slope estimate.
- Describe the difference between a confidence range and a prediction range.
- Be able to calculate yourself the degrees of freedom of error for bivariate regression.
- Get fitted values, residuals, and arbitrary predictions of a linear regression in R / RStudio.
- Appropriately communicate, using text and graphically, the findings of a questioned answered with linear regression.
- Perform linear regression in R / Rstudio, including all of the above.

(Please note that these skills are assessed in the Graded Assessment 2)

Unit 3 - Multiple linear regression part 1

By the end of this week you will be able to:

- Describe the assumptions of linear regression.
- Use QQ-plots and Tukey-Anscombe plots to assess validity of these assumptions.
- **Do all the things listed for week 2, except with multiple predictor / explanatory variables.**
- In particular, describe the type of biological / biomedical question that multiple linear regression could help answer.
- In particular, interpret the R model `summary` table for models with multiple predictor variables.
- Start to recognise the implications of having correlated predictor variables.
- Relate what are binary and factor type predictor / explanatory variables (sometimes called covariates).
- Start to do all the above for models with binary and factor covariates.
- In particular, calculated degrees of freedom for error with binary and factor covariates.
- In particular, use the F-test to compare models with/without a factor covariate.
- In particular, interpret the R model `summary` and `anova` table for binary and factor covariates.

(Please note that these skills are assessed in the Graded Assessment 3)

Unit 4 - Multiple linear regression part 2

By the end of this week you will be able to:

- Describe the type of biological / biomedical question that linear regression with interactions among predictor variables could help answer.
- **Do all the things listed for week 2, except with interactions among multiple predictor / explanatory variables.**
- In particulary, find and interpret the interaction terms in the R model `summary` table.
- In particular, calculated degrees of freedom for error with interacting predictor variables.
- Use the four common graphs for assessing validity of model assumptions (i.e. Tukey-Anscombe plot, QQ-plot, scale-location plot, and leverage plot).
- Use the ggfortify add-on package to easily produce these four diagnostic graphs.
- Recognise and fix two problems: non-normal residuals, and outliers.

(Please note that these skills are assessed in the Graded Assessment 4)

Unit 5 - ANOVA

By the end of this week you will be able to:

- Describe the type of biology / biomedical question that one-way ANOVA could help answer.
- Describe the type of biology / biomedical question that two-way ANOVA (with and without interaction) could help answer.
- **Do everything as with previous models, with one- and two-way ANOVA.**
- Understand and perform an F-test (the really important one for ANOVA).
- Describe what hypothesis is tested by an F-test.
- Understand when to use, and how to perform and interpret post-hoc hypothesis tests.
- Recognise that ANOVA is just another linear model, as is linear regression, and all the models you've seen so far.
- In particular, recognise that the model checking strategy is the same as for linear models from the previous weeks.
- Relate why ANOVA is often applied to data from manipulative experiments.

(Please note that these skills are assessed in the Graded Assessment 5)

Unit 6 - ANCOVA & Matrix alegbra

By the end of this week you will be able to:

- Describe the type of biology / biomedical question that one-way ANCOVA could help answer.
- **Do all the things listed for previous models, except now for ANCOVA.**
- In particular, interpret the R model `summary` table for ANCOVA type linear models.
- Recall basic concepts of linear algebra (e.g. vectors, matrices).
- Relate why linear algebra is useful in data analysis.
- Do some linear algebra in R.

(Please note that these skills are assessed in the Graded Assessment 6)

Unit 7 - Model selection

By the end of this week you will be able to:

- Describe what is model selection.
- Describe the type of biology / biomedical question that require model selection.
- Relate why model selection is difficult, fraught with danger, and perhaps more an art than a science.
- Understand the meaning, use and importance of p-values, AIC, AICc, BIC for model selection
- Describe what is forward, backward, and automatic selection.
- Relate the difference between explanatory and predictive models.
- Understand why automatic model selection procedures are not recommended for explanatory models.
- Recognise the importance of a priori hypotheses for escaping the nightmare that is model selection in explanatory models.

(Please note that these skills are assessed in the Graded Assessment 7)

Unit 8 - Interpretation, causality, and cautionary notes

By the end of this week you will be able to:

- Understanding the P-value.
- Critique, use and misues of p-values.
- Debate statistical versus biological significance (i.e. practical relevance).
- Describe the importance of the statement: “one cannot prove the null hypothesis”.
- Understand that the p-value is useful if it is interpreted properly.
- Appropriately assess and describe the importance of regression terms.
- Describe how causality, correlation, and effect are related.
- Describe what are effect sizes, and appropriately report them.
- Understand how to decompose r-squared among mulitple predictor variables, including via LMG.
- Recall the nine Bradford-Hill-Criteria for causal inference.
- Describe the difference between observational and experimental studies, and its importance for inference.

(Please note that these skills are assessed in the Graded Assessment 8)

Unit 9 - Analysing count data

By the end of this week you will be able to:

- Describe the type of biological / biomedical questions that involve count data.

- Relate why using a standard linear model to analyse such data could be a bad idea.
- Describe how a generalised linear model can differ from a standard linear one.
- Decribe the meaning of the terms *family*, *linear predictor*, and *link function*.
- Give the family, linear predictor, and link function most often used for count data.
- Fit a GLM (generalised linear model) to count data in R / Rstudio.
- **Do all of the same things as one does for a linear model.**
- In particular, understand how to interpret the regression coeffcients and check model diagnostics.
- Understand and perform Chi-squared-tests using the anova function.
- Say what is different about the model summary table produced by R / RStudio.
- Understand and check for a common problem with count data: overdispersion.

(Please note that these skills are assessed in the Graded Assessment 9)

Unit 10 - Analysing binary data

By the end of this week you will be able to:

- Describe the type of biological / biomedical questions that involve binary data.
- Use a conditional density plot to explore binary data.
- Relate why using a standard linear or Poisson model to analyse such data could be a bad idea.
- Recall the use of Chi-squared test for binary data (contingency tables).
- Understand and calculate an odds, and odds-ratio.
- Give the family, linear predictor, and link function most often used for binomial/binary data.
- Express a binary response appropriately to model it in R / RStudio.
- Fit a GLM (generalised linear model) to binary data in R / Rstudio (i.e. do logistic regression in R).
- **Do all of the same things as one does for a Poisson model.**
- In particular, understand how to interpret the regression coeffcients and check model diagnostics.
- Say what is different about the model summary table produced by R / RStudio.
- Understand and perform an Chi-squared-test (using the anova function).
- Understand and check for overdispersion.

(Please note that these skills are assessed in the Graded Assessment 10)

Unit 11 - Measurement error; repeated measures and random effects

By the end of this week you will be able to:

- Understand that in regression modelling the covariates are assumed to be error-free
- Understand that this is often not the case
- Be aware of the effect that the violation of this assumption has
- Know the most important error structures (classical and Berkson error)
- Be able to account for classical measurement error in simple cases (linear regression)
- Know at least one tool in R that can be used for error modelling
- Identify explanatory variables as better being included in models as *fixed* or as *random* effects.
- State what is a mixed model.
- Describe the type of biological / biomedical questions that required mixed models to analyse.
- Determine if a statistical test is probably pseudoreplicated.
- In R / RStudio, perform, check and interpret a simple linear mixed model.
- **Do all of the same things as one does for a linear model.**
- Say what is different about the model summary table produced by R / RStudio.
- Interpret the R model `summary` table for ANCOVA type linear models.

(Please note that these skills are assessed in the Graded Assessment 11)

Unit 12 - Recap and outlook

By the end of this week you will be able to:

- Describe the type of biological / biomedical questions that require other statistical methods, including:
  - Time series analysis
  - Multivariate analysis, including ordination, clustering, and classification
  - Breakpoint analysis
  - Nonlinear regression
  - Structural equation modelling / path analysis
  - Generalised linear mixed models
  - Bayesian methods
  - Generalised addative models
  - Meta-analysis
  - Survival analysis
  - Non-parametric analyses
  - Spatial analyses / statistics
  - Power analysis
  - Randomisation based methods