Introduction and statistical packages

Based on a book by Julian J. Faraway

University of Iceland

The book

The material in this course is based on a freely available book by Julian J. Faraway.

The book can be downloaded from here:
`https://cran.r-project.org/doc/contrib/Faraway-PRA.pdf`

Software

- We will use the statistical package R which can be downloaded from here:
  `https://cran.r-project.org/mirrors.html`
- We recommend using RStudio which can be downloaded from here:
  `https://www.rstudio.com/products/rstudio/download/`
- A short introduction to R can be found in Appendix C in the book and many resources are available online such as the Cookbook for R:
  `http://www.cookbook-r.com/`

Before you start

Statistics starts with a problem, continues with the collection of data, proceeds with the data analysis and finishes with conclusions.

It is a common mistake of inexperienced Statisticians to plunge into a complex analysis without paying attention to what the objectives are or even whether the data are appropriate for the proposed analysis.

Look before you leap!

Formulation

1. Understand the physical background. Statisticians often work in collaboration with others and need to understand something about the subject area. Regard this as an opportunity to learn something new rather than a chore.
2. Understand the objective. Again, often you will be working with a collaborator who may not be clear about what the objectives are. Beware of "fishing expeditions" - if you look hard enough, you'll almost always find something but that something may just be a coincidence.
3. Make sure you know what the client wants. Sometimes Statisticians perform an analysis far more complicated than the client really needed. You may find that simple descriptive statistics are all that are needed.
4. Put the problem into statistical terms. This is a challenging step and where irreparable errors are sometimes made. Once the problem is translated into the language of Statistics, the solution is often routine. Difficulties with this step explain why Artificial Intelligence techniques have yet to make much impact in application to Statistics. Defining the problem is hard to program.

Data Collection

- Are the data observational or experimental? Are the data a sample of convenience or were they obtained via a designed sample survey. How the data were collected has a crucial impact on what conclusions can be made.
- Is there non-response? The data you don't see may be just as important as the data you do see.
- Are there missing values? This is a common problem that is troublesome and time consuming to deal with.
- How are the data coded? In particular, how are the qualitative variables represented.
- What are the units of measurement? Sometimes data is collected or represented with far more digits than are necessary. Consider rounding if this will help with the interpretation or storage costs.
- Beware of data entry errors. This problem is all too common — almost a certainty in any real dataset of at least moderate size. Perform some data sanity checks.

Initial Data Analysis

This is a critical step that should always be performed. It looks simple but it is vital.

- Numerical summaries - means, sds, five-number summaries, correlations.
- Graphical summaries
  - One variable - Boxplots, histograms etc.
  - Two variables - scatterplots.
  - Many variables - interactive graphics.

When to use Regression Analysis

- Regression analysis is used for explaining or modeling the relationship between a single variable Y , called the response, output or dependent variable, and one or more predictor, input, independent or explanatory variables, $X_1, ..., X_p$.
- When $p = 1$, it is called simple regression but when $p > 1$ it is called multiple regression or sometimes multivariate regression.
- When there is more than one Y, then it is called multivariate multiple regression which we won't be covering here.

When to use Regression Analysis

We will only look into models where we have one continuous response variable but the explanatory variables can be of different types:

- When the explanatory variables are all numerical (quantitative): regression
- When the explanatory variables are all categorical (qualitative): analysis of variance (ANOVA)
- When the explanatory variables are a mix of numerical and categorical: analysis of covariance (ANCOVA)

When to use Regression Analysis

Regression analyses have several possible objectives including

1. Prediction of future observations.
2. Assessment of the effect of, or relationship between, explanatory variables on the response.
3. A general description of data structure.

Before we start...

Before we start diving into multiple linear regression (MLR) we will review some basic concepts (slides: Basic concepts and introduction to statistical inference) and simple linear regression (SLR) (slides: Simple linear regression (SLR))