Cornell University
Cornell Statistical Consulting Unit

Stats Happening - December 2014

1) CSCU Winter Session Workshops

A) Data Carpentry Bootcamp at Cornell: Jan 15th & 16th 2015

The Cornell Statistical Consulting Unit (CSCU) is hosting a Data Carpentry Bootcamp this winter on Jan 15th and 16th. This workshop will be similar to the one we held in June but with a greater emphasis on data management and analysis workflows. The workshop is aimed at academic researchers in all fields and at all career stages. Data Carpentry is designed to teach basic concepts, skills and tools for working more effectively with data. The workshop will cover the following topics:

- How to use spreadsheet programs (such as Excel) more effectively and the limitations of such programs.
- Getting data out of spreadsheets and into more powerful tools (R).
- Using databases, including managing and querying data in SQL.
- Workflows and automating repetitive tasks.

For more information on workshop content, see http://datacarpentry.org/ or contact cscu@cornell.edu.

We especially encourage registration for those who may be less familiar with the above topics. There is no prerequisite as to what computing skills and knowledge is required. Participants will be able to work on library computers, but are encouraged to bring and use their own laptops to insure the proper setup of tools for an efficient workflow once you leave the workshop. To help defray the costs there will be a

$40 registration fee, with some allowances for scholarships. Please watch the CSCU webpage this month and for an email announcement for information on how to register.

B) Multilevel Modeling Workshop: Jan 13th & 14th 2015

Start the spring semester early by attending the CSCU Multilevel Modeling workshop right before classes begin: Tuesday January 13th and Wednesday January 14th 2015. Multilevel models (also referred to as mixed models or hierarchical models) are used when observations are not independent. Data clustering occurs with many experimental designs (e.g., split plot designs), with social science data collected simultaneously on different units of analysis (e.g., households and their individual members), with measurements taken on the same units over several time periods (longitudinal studies) and with spatial data. The purpose of this workshop is to introduce the concepts that form the basis of these models, the underlying statistical model and the estimation techniques that are used. Many examples will be presented during the workshop to enable participants to recognize such models when encountered in their own research, to analyze them and interpret the results. The workshop is intended for participants who have the equivalent of two semesters of statistics and some previous experience with ANOVA and linear regression. The workshop will be taught through a combination of lectures and hands-on computer exercises. The lecture will be on Tuesday January 13th from 9:00am until 12:00am in 100 Savage Hall. For the hands-on session in Mann library B30A you have a choice to sign up for Tuesday 13th from 2:30-4:00pm or on Wednesday from 10:00-11:30am.

For more information and registration: http://www.cscu.cornell.edu/workshops/multilevel.php

2) Open Refine for taming messy data

When using secondary data, researchers often are faced with data sets are often not formatted in a way that is conducive to statistical analysis. OpenRefine is a powerful program that can aid in formatting messy data. It can handle basic data exploration, data cleaning, transformations, reclassifying categories, splitting or merging variables and more. It runs in a web browser from an executable file on your computer, so all data is secure. It can handle a wide variety of formats, can interface with SQL, JSON, and several online databases. Every action you take is recorded so you always know how your data had changed and so you can repeat processing with similar files. Open Refine has a large online user community that is eager to share techniques for specific disciplines.
http://openrefine.org/

3) Overdispersed Poisson Regression & OLRE

Count data is commonly modeled using a poisson distribution. By definition, in a poisson distribution the mean is equal to the variance. This is often not the case in practice and can be addressed by modeling a poisson distribution with an additional overdispersion parameter. Failing to correctly model this overdispersion yields biased parameter estimates. Most commonly used software packages have the ability to estimate an overdispersed poisson regression. A statistical software that can estimate an overdispersed poisson with random effects is however not readily available. One way to model the overdispersion in such model is by adding an observation-level random effect (OLRE). The following article describes OLRE and assesses its efficacy in dealing with overdispersion. It shows that OLRE is effective in dealing with overdispersion (unless there is a high level of overdisperson) but that it fails to reduce bias with zero-inflated poisson.

Using observation-level random effects to model overdispersion in count data in ecology and evolution.
Xavier A. Harrison
Institute of Zoology, Zoological Society of London, London, UK
Harrison (2014),PeerJ, DOI 10.7717/peerj.616
https://peerj.com/articles/616.pdf

4) Extracting data out of published papers

Have you ever wanted to replicate a classic analysis, or try fitting a new model to published data? Many published articles do not come with data appendices. There are several types of software that allow you to start with a pdf or scan of an article and extract data from tables or figures.

Extracting Data out of Tables: It is easy to copy and paste a table from a pdf, but the formatting afterwards can be quite challenging. **Tabula** is a free, opensource software that will preserve the formatting for you. It needs to be downloaded and run through an .exe file, but will open in your browser. Upload a pdf, select a table from the thumbnail with a rectangular tool, and it will provide a spreadsheet or csv of the data. It appears to work via onscreen character recognition, so annotations can limit accuracy. http://tabula.nerdpower.org/
Extracting Data out of Figures: All of the software options below allow you to hand-digitize points or lines after defining axes. All can handle log scales. **Plot Digitizer** is very intuitive, but can only handle XY plots. It comes with an auto-trace option for auto-digitizing. http://plotdigitizer.sourceforge.net/ **Data Thief** digitizes points and traces lines. It is less intuitive, but comes with a thorough manual. By paying the shareware fee, you can unlock more features, such as a polar coordinate system. http://datathief.org/ **Webplot Digitizer** runs in most browsers, no need to download or install. It can handle XY plots, polar coordinates, ternary diagrams, and maps (with scale bar). It has capabilities of hand-digitizing and auto detection. The onscreen instructions are intuitive and easy to follow, but also comes with a thorough manual and tutorials. This is open source software available on GitHub. http://arohatgi.info/WebPlotDigitizer/

5) Finite Mixture Models

If you have performed a regression but are unable to meet the assumptions despite transforming variables, a finite mixture model (FMM) might offer you the solution you have been looking for. A FMM approximates a distribution by a mixture of distributions. This can be very useful if you are faced with very skewed or multi-modal distribution that actually is created by overlaying several symmetric distributions coming from different subpopulations e.g. length of specific species of fish grouped by gender when gender is unknown, or distribution of hemoglobin across various populations. Typically the number of groups and how observations are grouped are not known so FMM (also known as latent class models) are considered an unsupervised learning technique. Besides several packages available in R to estimate FMM, SAS and STATA now also implement FMM. For a user-friendly introduction see: https://support.sas.com/resources/papers/proceedings12/328-2012.pdf

6) Julia: A New Language and Statistical Computing Toolkit

Julia is a high-level, high-performance dynamic programming language for technical computing. It is open source and its syntax is familiar to users of other statistical computing environments such as R or Python. Julia is built for analysis of very large databases: it is designed for parallelism and distributed computation and provides a sophisticated compiler. An extensive mathematical function library, largely written in Julia itself, also integrates mature, best-of-breed C and Fortran libraries for linear algebra, random number generation, signal processing, and string processing. The Julia developer community is contributing a number of external packages through Julia's built-in package manager at a rapid pace, such as Douglas Bates' MixedModels package. http://julialang.org/

7) Treatment Discontinuation and Missing Data

A recent invited lecture at Cornell by Prof. Rod Little outlined some important distinctions between treatment discontinuation and missing data in the context of clinical trials. Data collection is often stopped after treatment discontinuation, but outcome data can still be recorded on individuals after they discontinue treatment. Conversely, outcome data may be missing for individuals who do not discontinue treatment, as when there is loss to follow up or missed clinic visits. Missing outcome data is a standard missing data problem, but Prof. Little argues that treatment discontinuation is better viewed as a form of noncompliance and best analyzed in a causal inference framework on noncompliance. For more details on this topic as well as the estimators recommended, see the full article by Little and Kang: (wileyonlinelibrary.com) DOI: 10.1002/sim.6352: http://onlinelibrary.wiley.com/doi/10.1002/sim.6352/full

8) Geospatial science and technology at Cornell

The first annual Cornell Geospatial Forum (CUGEO) took place this past October. This was a one-day gathering of faculty, staff, students, and community members to discuss the past, present, and future of geospatial science and technology at Cornell. Spatial analyses and place-based studies are more important than ever in the world today, and the digital technologies that support such work are becoming much more commonplace and in high demand. Although CSCU can address questions related more specifically to spatial statistics, we are not equipped to handle all the questions related to GIS. So we are happy to inform you that following the CUGEO forum, a Geographic Information System (GIS) consulting desk was set up at Mann library and is open for walk-in consulting most weekday afternoons. For detailed information about this and other GIS resources at Cornell see: http://mannlib.cornell.edu/research-help/gis