Example 1: Effects of Spatial Correlation on the Analyses of Field Experiments.

Suppose a field experiment is conducted to compare the effect of a treatment (or several treatments) on some response. The traditional analysis of variance assumes that the observations from such an experiment are independent; however, this assumption is often not met because the observations are gathered at different locations and are thus correlated spatially. Though methods of accounting for this spatial correlation do exist and are easily implemented in some statistical software packages, many researchers still use traditional analyses which assume independence. How serious are the implications of this departure from model assumptions? To investigate this question, data from a field experiment with known spatial correlation structures will be simulated. Simulation studies will be used to investigate the type I error rate and power of both a traditional analysis which assumes independence and also a spatial analysis which accounts for the correlation structure.

Bibliography

- Clark, Isobel and William V. Harper (2000). Practical Geostatistics 2000. Geostokos (Ecosse) Limited, Scotland.
- Li, Erning, Dennis Boos, and Marcia Gumpertz. *Simulation Study in Statistics (Draft).* Available online: http://www4.stat.ncsu.edu/~reich/st810A/simulationinstatistics.pdf.

Example 2: Investigating Rankings of NCAA Basketball Teams

The ranking of college athletic teams is often a hotly debated issue, especially when so many different variables such as strength of schedule, whether or not a team is playing on their respective home court or field, etc., can have an impact on the results of a game or match. How do we fairly determine what team deserves to be ranked where? One approach is to use Bradley-Terry models, which essentially evaluates pairwise comparisons (in terms of wins and losses) between teams. In this project, the student will use Bradley-Terry models to rank basketball teams in the Big Ten. In addition to learning about the theory involved with such models, the student will also gain experience scraping data from the web, cleaning data, and fitting models. The ultimate goal of the project is to obtain rankings based on the Bradley-Terry model and to compare these to existing polls or rankings.

Bibliography

- Agresti A (2002). Categorical Data Analysis. 2nd edition. John Wiley & Sons.
- Bradley RA (1984). "Paired Comparisons: Some Basic Procedures and Examples." In PR Krishnaiah, PK Sen (eds.), Nonparametric Methods, Volume 4 of Handbook of Statistics, pp. 99- 326. Elsevier.
- *http://cran.r-project.org/web/packages/BradleyTerry2/vignettes/BradleyTerry.pdf*

Example 3: Investigating the Validity of the Survey of Attitudes Toward Statistics

The Survey of Attitudes Toward Statistics (SATS) is a tool widely used by statistics educators to help gain insight into students' attitudes and how they impact teaching and learning in introductory statistics courses. Three instructors at Winona State University have been administering this survey to students both at the beginning and end of several semesters since 2011. This study will involve an analysis of the data collected in these courses to investigate students' attitudes towards statistics and how they change throughout the semester. The results will also be compared to national norms. Finally, an exploratory factor analysis will be conducted using the data collected from Winona State University courses to investigate the construct validity of the SATS tool.

Bibliography

- http://www.evaluationandstatistics.com/index.html

Example 4: Measuring and Modeling Batted Ball Quality

The main idea behind this capstone project is to learn which variables impact the quality of a hit ball in the MLB. Using StatCast data, I will be able to obtain information on every single pitch for every single hitter in the MLB. Here is some of the information included in the StatCast data: player name, pitcher name, pitch type, event, description, spin direction, spin rate, break angle, break length, pitch location, hit distance, hit speed, hit angle, and much more. This data source is a very rich place to find information about the game of baseball. For my project, I will look into how to define what a well hit ball is. With that being said, exit velocity is the response variable that I foresee myself using in much of my analysis. For my project, I will focus on three specific avenues: learning about variable importance, learning and comparing different model performance, and a simple application where one can compare different hitters and their respective variables that influence a well hit ball.

Bibliography

- James, Witten, Hastie, and Tibshirani. An Introduction to Statistical Learning. BaseballSavant.com.

Example 5: Analyzing Video Game Sales

This project will analyze characteristics of different video games and how their sales vary regionally, across platform and genre. A major component of this project will be scraping and cleaning data off the Internet, as there is no available data in analysis-ready format. Accordingly the student will be required to implement data-scraping capabilities of software such as import.io or R. Once the data have been scraped, the student will create interactive visualizations to investigate factors that influence sales. Ideally, the student will also obtain data on video game ratings and determine whether correlation exists between a game's rating and its sales, and investigate the characteristics of any outlying games.

Bibliography:

- http://vgsales.wikia.com/wiki/List_of_video_game_sales_websites
- http://www.vgchartz.com/
- http://gamerinvestments.com/video-game-stocks/

Example 6: Biomarkers of Inflammation and Mortality in the CHS

The Cardiovascular Health Study is a government sponsored longitudinal study of adults aged 65 or older in four communities. CHS is designed to determine the importance of conventional cardiovascular disease (CVD) risk factors in older adults, and to identify new risk factors in this age group, especially those that may be protective and modifiable.

Data on a subset of variables on 5000 participants is available for analysis. Interest lies in two biochemical markers of inflammation: the C reactive protein and fibrinogen. This project will assess whether patients with higher levels of these biomarkers are at increased risk of death in general and cardiovascular disease in particular, and what associations exist between these biomarkers and other known risk factors of cardiovascular disease. Accordingly this project will analyze the association of biomarkers with known risk factors and survival time. Potential questions of interest may include the following:

1. What associations exist between the inflammatory biomarkers and other known risk factors for cardiovascular disease?
2. Are the inflammatory biomarkers predictive of overall mortality and cardiovascular disease specific mortality either individually or in combination?
3. Does any association between the inflammatory biomarkers and mortality differ between short-term (death within 3 years) and long-term (death after having survived at least 3 years)?
4. Does any predictive value of the biomarkers differ between the sexes?
5. Do any associations found above persist after adjustment for known risk factors for cardiovascular disease?

The project will implement both exploratory and inferential analyses. The former will exploit the use of Kaplan-Meier curves and scatterplots; the latter will include fitting conventional survival models such as the Cox Proportional Hazards Model.

Bibliography:

- Fried, LP *et al*. The cardiovascular health study: design and rationale. *Annals of Epidemiology* 1(3): 263-276. 1991.
- Kleinbaum, D. G., and M. Klein. *Survival analysis. A self-learning approach.* Springer, New York, USA. 2005.

Example 7: Money Spent on College Education

The student will investigate whether students attending US undergraduate colleges and universities are getting their "money's worth" regarding their college education. This is of interest with the continually increasing cost for students to obtain an undergraduate degree; are some students at some schools getting an education comparable to its cost. Several factors (such as acceptance rate, average debt upon graduation, and job placement rate) will be collected on each institution for this analysis. The student will utilize methods and techniques learned in his coursework (e.g., STAT 325, STAT 350, STAT 360) to complete this project. The student will also gain experience in creating/managing a data set in addition to then using multiple techniques to determine at which school students get the most for their money.

Bibliography:

- Cook, Dennis. Weisberg, Sanford. *Applied Regression Including Computing and Graphics*. John Wiley & Sons, Inc. New York. 1999.
- Kutner, et al. *Applied Linear Statistical Models*. McGraw-Hill Irwin. Madison. 2005