Cornell University
Cornell Statistical Consulting Unit

Stat Happenings #10: Statistics News You Can Use
Cornell Statistical Consulting Unit
October 2009

1. Statistical workshops this semester

There is still time to register for the following workshops:

- **Introduction to Logistic Regression**

This workshop will be offered Thursday ***October 15*** from 1:30 PM 3:00 PM in the Stone Computing Room in Mann Library. For more information and to register visit: http://www.cscu.cornell.edu/workshops/schedule.php.

CSCU is also offering the following workshops this semester:

- **Getting Started with Data Analysis:** Tuesday, ***October 20*** from 6:00 PM - 8:00 PM
- **NEW this Fall! Univariate Repeated Measures vs MANOVA:** Monday **October 26** from 12:00 to 1:00 PM and the second part on Monday **November 2**, also from 12:00 to 1:00 PM
- **Refresher on Interpreting Linear Regression Parameters:** Tuesday, ***October 27*** from 11:30 AM - 1:30 PM
- **Introduction Logistic Regression for Responses with More than Two Categories:** Thursday, ***October 29*** from 1:30 PM - 3:00 PM
- **Introduction to Multilevel Models:** during the January break.

For more information and to register for these workshops go also to:
http://www.cscu.cornell.edu/workshops/schedule.php.

2. Bayesian statistical analysis SAS 9.2.

Are you trying to implement Bayesian methods in your statistical analyses?

SAS 9.2 offers Bayesian analysis for ANOVA, logistic regression, Poisson and Cox regression. It is implemented in the GENMOD, LIFEREG and PHREG procedures. A BAYES statement for Bayesian analysis via Gibbs sampling is available with these procedures. PROC MIXED can also perform a sampling-based Bayesian analysis through the use of a PRIOR statement that currently operates only with variance component models.

More information can be found at: http://support.sas.com/rnd/app/da/bayesproc.html.

3. Multiple Imputation with Stata 11

Stata 11 offers a new multiple imputation command, the MI command. In addition to the previously existing ICE command for analyzing data sets with missing values, Stata now allows for multiple imputations using the joint multivariate normal distribution approach. The MI approach assumes a joint multivariate normal distribution of the variables.

4. Exact logistic regression with Stata

Have you tried to run a logistic regression model with a small data set or with one-way causation, such as the case where all females are observed to have a positive outcome? Estimation using the usual asymptotic maximum likelihood methods may be inadequate when sample sizes are small or the data are sparse. If you have such a data set, then you may find the Stata *exlogistic* command useful. For more information visit: http://www.stata.com/help.cgi?exlogistic

5. GEE modeling with R - A Word of Caution When Using GEE Models in R

The Generalized Estimating Equations (GEE) approach is useful in the analysis of correlated data where the outcome has a distribution such as Binomial, Normal (Gaussian), Poisson and Negative Binomial. An example of a correlated data structure is when measurements are taken on students nested within schools. Measurements taken on students from the same school are more likely to be similar to each other (i.e. they are correlated). GEE models can be implemented in statistical package R using either a "gee" function in the package "GEE" or a "geeglm" function in the package "GEEPACK". Both functions share a similar syntax and provide similar output. However, a user must be warned that when the data set is not sorted by the clustering variable (the school variable in our example), incorrect parameter estimates will be produced without warning. Therefore, sorting the data set by the clustering variable is **essential** prior to GEE analysis in R in order to obtain correct results.

6. Useful macro for variable selection for logistic regression with correlated data with SAS

The selection of predictor variables when a large number of potential variables exists is often cumbersome. Logistic regression models can be implemented in SAS using PROC LOGISTIC and PROC GENMOD procedures. PROC LOGISTIC offers an automated explanatory variable selection option in the MODEL statement. Unfortunately, PROC LOGISTIC cannot handle correlated data. The GENMOD procedure can model correlated data with binary outcome using Generalized Estimating Equations (GEE), but does not have

an automated model selection tool. It is however possible to automate the model selection process in PROC GENMOD, and thus speed up model building process in logistic regression with correlated data, with the use of a SAS macro. The macro can be downloaded at http://www.nesug.org/proceedings/nesug07/cc/cc26.pdf.

7. Improved graphing capabilities in JMP and SAS 9.2

- The new Graph Builder option in JMP 8 is an easy way to create graphical displays of your data by dragging and dropping variables. It is interactive and high dimensional. For more information and a tutorial on the Graph Builder feature visit: http://www.sasconsig.com/software/jmp8/demos/sall_graphbuilder.shtml.

- Producing publication quality statistical graphics in SAS has become easier. SAS 9.2 offers improved graphing capabilities with the new Statistical Graphics procedures SGPLOT, SGSCATTER, SGPANEL, and SGRENDER available in the ODS Graphics module. It is also possible to edit and annotate the graphs using the ODS Graphics Editor. For more information see the following article: http://www2.sas.com/proceedings/forum2008/235-2008.pdf

8. Access to specialized statistical software: Mplus

CSCU can give you access to the latest version of Mplus. Mplus is a statistical program for estimating a wide range of models containing latent, or unobserved, variables. Mplus can handle models with both continuous and categorical latent variables. Mplus also can handle multilevel data, complex survey data, and missing data. If you wish to use the software or try it out before purchasing it or if you, feel free to contact us.

9. PASW Statistics 18 (formerly SPSS Statistics) available soon at Cornell

The new version of PASW (formerly SPSS Statistics), version 18, will be soon available at Cornell. For current subscribers of SPSS, the new version will be made available to you at no additional cost as soon as it is available to Cornell. We are looking forward to testing new modules, namely the Boostrapping module and the Statistics Developer module, which promise an easy approach for users who wish to work with R packages and share procedures with others. For more information visit: http://www.spss.com/statistics/.

10. Recommended readings

- Do you want to assess mediation using multilevel data sets?

A mediator variable mediates the observed relationship of a predictor variable with the outcome variable. In multilevel data sets, such as students nested within schools, data were collected at more than one level. In this case, the usual methods for testing mediation can produce erroneous results, as they do not take the multilevel structure into account.

For a comprehensive discussion on dealing with mediation in the context of hierarchical linear models, read *Testing Multilevel Mediation Using Hierarchical Linear Models*, Zhen Zhang et al., that can be found at: http://orm.sagepub.com/cgi/rapidpdf/1094428108327450v1 .

- Will you be testing multiple hypotheses in your study?

The following article illustrates how multiple hypotheses testing can produce associations with no clinical plausibility: *Testing multiple statistical hypotheses resulted in spurious associations: a study of astrological signs and health*, Austin P.C.,et. al., Journal of Clinical Epidemiology 59 (2006). It is available at: http://www.centrocochranedobrasil.org.br/mestradoProfissional/aulas/29junho/Testing_multiple_statistical_hypotheses_resulted_in_spurious_association_-_a_study_of_astrological_signs_and_health.pdf.

10. Cornell's Dropbox

Did you know that Cornell has a safe way of transporting large data files between users?

The Cornell Dropbox is a service that allows users to send and receive large files (up to 1.5 GB). This service is available to any person with a Cornell NetID, and allows Cornell members to send and receive files of up to 1.5 GB, from other Cornell users or from external users without a NetID. For more information visit: https://dropbox.cornell.edu/.