University of Pennsylvania
School of Social Policy & Practice
SW 899
Applied Linear Modeling

Spring 2010

I.   Course Description

This course deals with the underlying assumptions and applications of the general linear model with social science, education, and social welfare policy related questions/data. The first half of the course begins by covering simple linear regression and the assumptions of the general linear model, assumption diagnostics, consequences of violation, and how to correct for violated assumptions. This will also include methods of incomplete case analysis (i.e. missing data analysis). Then various aspects of regression analysis with multiple independent variables will be covered including categorical explanatory variables (e.g. to estimate group differences), interaction effects, mediating effects (e.g. to estimate the indirect effect of social processes), and non-linear effects. The course will then cover some of the applications of the general(ized) linear model including logistic regression, some elements of path modeling (structural equation modeling), multilevel analysis (hierarchical linear modeling), and longitudinal modeling (both growth modeling and time-series analysis). The course will be taught using SAS, but students are welcome to use any statistical package of comfort. Pre-requisite: Introductory Graduate Statistics.

II.   Educational and Learning Objectives

By the end of Applied Linear Modeling students are expected to demonstrate:

1. a strong foundation in simple linear and multiple regression and how to arrive at the Best Linear Unbiased Estimates (B.L.U.E.);
2. their knowledge of the main assumptions of the general linear model and the implications they have on estimation when they are violated;
3. how to conduct diagnostics and correct for the violation of the assumptions of the general linear model;
4. how to interpret various coefficients;
5. how to estimate moderating (i.e. interaction) and mediating effects (and their relationship to path modeling);
6. have an understanding of the various incomplete case analytic techniques and be able to conduct an incomplete case analysis;
7. how to analyze binary outcomes in logistic regression;
8. what nested data structures are and the basics of how to analyze them within multilevel (or hierarchical linear) modeling;
9. a strong competence in the selection of appropriate applications of the general(ized) linear model given particular research questions;
10. how to effectively communicate, in writing, the results for a professional publishable journal article (including tables, coefficients, etc.).

III. Course requirements:

<u>**Expectations**</u>

Classroom learning is a fundamental component of your professional education. Students are therefore expected to attend each class, arrive to class on time, and be in attendance for the full class. In the event that you are unable to attend class for any reason, you must notify the instructor in advance and learn how you are to make up the content you missed. Excessive absenteeism (i.e., missing more than two classes) is considered a serious problem the instructor will handle by meeting with the student and determining whether the student's educational adviser should be notified. Excessive absenteeism could result in course failure.

Students are expected to: (A) participate substantively in class discussions; (B) read on a weekly basis and come to class prepared to apply and discuss the reading assignments; (C) submit assignments by the due date and in accordance with the specified format.

Grades will be based on four major assignments written as would be reported in the results section of a publishable manuscript. The first three assignments will be worth 20% of your course grade and the final assignment will be worth 40% of your overall course grade.

<u>**Assignments**</u>

Students will be responsible for four, graded written assignments during the semester. These assignments will require a specific analytic task of applied linear modeling and students will be required to perform and show all work. This will include submitting print outs of the output to analyses and writing up the results as you would report in the results section of a publishable manuscript. Specific instructions for completing each assignment will be provided during the semester.

Format

All papers must be typewritten, in 12-point font, double-spaced, page-numbered, with 1" margins at the left, right, top and bottom. The cover page (not included in the page limit) should include the title of the paper, student's name, assignment number, professor's name and date submitted. Papers should be stapled, not paper-clipped. Papers **must be proofread** carefully for clarity, organization, spelling, punctuation, and other potential errors before submission.

I**n-text citations following APA style guidelines are required** for all written assignments, with the specific source including authors' last names and year of publication, regardless of whether you are paraphrasing or using specific quotes. Direct quotes must have the specific source as above but with page number(s). **A list of references cited or consulted must be included at the end of each paper in proper APA bibliographic form**. Footnotes may be used where appropriate to further explicate a concept or issue. **American Sociological Association style may be used, but this must be consistent throughout the assignment.**

You should keep a copy of each paper submitted. The instructor will gladly answer any questions regarding format, citing or organization. Papers written for other classes may not be submitted for written assignments in this course. Direct substitution of papers between courses may result in a failing grade for that assignment.

Plagiarism
Students are expected to conduct themselves consistent with the University of Pennsylvania's Code of Academic Integrity, which presents standards regarding plagiarism, multiple submissions and other actions. Students are expected to be familiar with the Code, which can be found at http://www.vpul.upenn.edu/osl/acadint.html

Evaluation
Assignments will be evaluated based on the following criteria:
1. demonstrates understanding of the research question and how to appropriately analyze;
2. effective use of SAS or any other statistical software package;
3. shows all analyses;
4. utilizes tables and reports data/coefficients appropriately, as would be expected for the results section of a peer-review journal article;
5. demonstrates understanding of the particular task/application of the general(ized) linear model;
6. and, appropriately interprets data.

Grading Policies
The final course grade is based on the student's performance on all assignments. Students whose performance is minimal or failing at midterm will be notified in writing.

Readings
It is expected that students will read required class assignments from the recommended texts and/or articles, and from relevant materials of their own choosing.

Required Texts
Cohen, Jacob, Patricia Cohen, Stephen G. West, and Leona S. Aiken, L.S. 2003. *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences,* Third Edition. Mahwah NJ: Lawrence Erlbaum Associates.

Delwiche, Lora D., and Susan J. Slaughter. 2003. *The Little SAS® Book: A Primer,* Third Edition. Cary, NC: SAS Institute Inc.

Blackboard: Additional readings will be posted on-line via blackboard at: https://courseweb.library.upenn.edu/

TOPIC AND READING LIST BY WEEK

Section I: Introduction to Course and The General(ized) Linear Model

Week 1: January 14
Course Introduction. Review of the syllabus, and course requirements. What is the general(ized) linear model? What is correlation? What is regression? Introduction to the SAS program.

Week 2: January 21
Bivariate Correlation and Regression.

Readings:
Cohen, et.al., text, Chs. 1 & 2

Delwiche & Slaughter, text, Chs. 1, 2, & 3

Section II: Multiple Regression and Assumptions of the General Linear Model

Week 3: January 28
Multiple Regression with Two or More Independent Variables.

Readings:
Cohen, et.al., text, Chs. 3 & 5

Week 4: February 4
Multiple Regression with Categorical Independent Variables.

Readings:
Cohen, et.al., text, Ch. 8

Week 5: February 11   *Assignment 1 Due: Multiple Regression*
Assumption Diagnostics and Violation Correction.

Readings:
Cohen, et.al., text, Ch. 4

Week 6: February 18
Assumption Diagnostics and Violation Correction Continued: Model Specification.

Readings:
Cohen, et.al., text, Ch. 6

Week 7: February 25
Outliers and Multicollinearity.

Readings:
Cohen, et.al., text, Ch. 10

Week 8: March 4
Missing Data/Incomplete Case Analysis

Readings:
Cohen, et.al., text, Ch. 11

Peng, Chao-Ying Joanne, Michael Harwell, Show-Mann Liou, and Lee H. Ehman. 2007. "Advances in Missing Data Methods and Implications for Educational Research." in *Real Data Analysis.* edited by S.S. Sawilowsky. Charlotte, NC: Information Age Publishing. (Pp. 31-78)

Note: No class on March 11 (Spring Break).

Section III: Applications of the General(ized) Linear Model

Week 9: March 18	*Assignment 2 Due: Assumption Diagnostics &*
Interaction Effects.	*Incomplete Case Analysis*

Readings:
Cohen, et.al., text, Chs. 7 & 9

Week 10: March 25
Indirect/Mediating Effects & Path Modeling.

Readings:
Cohen, et.al., text, Ch. 12

Week 11: April 1
Indirect/Mediating Effects & Path Modeling Continued.

Readings:
Loehlin, John C. 2004. *Latent Variable Models: An introduction to factor, path, and structural equation analysis* (Fourth Edition). Mahwah NJ: Lawrence Erlbaum Associates. Chapter One: "Path models in factor, path, and structural equation analysis." (Pp. 1-34)

Week 12: April 8	*Assignment 3 Due: Estimating Interaction Effects*

The Analysis of Binary Outcomes: Logistic Regression.

Readings:
Cohen, et.al., text, Ch. 13

Agresti. Alan. 1996. *An Introduction to Categorical Data Analysis.* New York NY: John Wiley & Sons. Chapter Five: "Logistic Regression." (Pp. 103-144)

Week 13: April 15
The Analysis of Nested Data Structures: Multilevel/Hierarchical Linear Modeling.

Readings:

Raudenbush, Stephen, and Anthony Bryk. 2002. *Hierarchical Linear Models: Applications and Data Analysis Methods* (Second Edition). Thousands Oaks CA: Sage Publications.
Chapter Two: "The Logic of Hierarchical Linear Models" (Pp. 16-37)

Singer, Judith. 1998. "Using SAS PROC MIXED to Fit Multilevel Models, Hierarchical Models, and Individual Growth Models." *Journal of Educational and Behavioral Statistics*. 24(4):323-355.

Recommended Reading:
Singer, Judith D., and John B. Willett. 2003. *Applied Longitudinal Data Analysis: Modeling Change and Event Occurrence.* New York NY: Oxford University Press.
Chapter One: "A Framework for Investigating Change over Time" (Pp. 3-15)
Chapter Three: "Introducing the Multilevel Model for Change" (Pp. 45-74)

Week 14: April 22
The Analysis of Longitudinal Data: Time Series Analysis.

Readings:
Yaffee, Robert and Monnie McGee. 2000. *Introduction to Time Series Analysis and Forecasting with Applications of SAS and SPSS.* New York NY: Academic Press.
Chapter One: "Introduction and Overview" (Pp. 1-13)
Chapter Four: "The Basic ARIMA Model" (Pp. 101-149)

Assignment 4 Due April 28th (No Later Than 5pm): Application of the General Linear Model

Assignment 1: Multiple Regression (20%)
Due: Thursday February 11th

There has been increasing research evidence on the meaningful effect nutrition and family income between the ages of 0 to 5 has on child development (Birch & Gussow 1970; Duncan & Brooks-Gunn 1997). You have been hired as a research consultant for an evaluation study examining the effect of the Women, Infant and Children (WIC) Nutrition Program and Aid to Families with Dependent Children (AFDC) program participation during pregnancy on child reading achievement. This is a large evaluation study using a national probability sample, the Child Development Supplement to the Panel Study of Income Dynamics, and thus has meaningful policy implications. Using the constructed data set economic.good, the evaluation team has asked you to examine the effect and unique contribution of (1) WIC program participation during pregnancy (WICpreg) on child reading achievement (readss97) over and above the child's age in 1997 (age97), family income (faminc97), low birth weight status (bthwht), and parenting practices (HOME97) and (2) AFDC program participation during pregnancy (AFDCpreg) on child reading achievement (readss97) over and above the child's age in 1997 (age97), family income (faminc97), low birth weight status (bthwht), and parenting practices (HOME97). The evaluation team is interested in (1) each variables individual effect, (2) the total variation of child reading achievement accounted for by the model, (3) the unique contribution of both program participations on the variance of child reading achievement, and (4) which program has a larger effect on child reading achievement.

Thus, you are asked to write a report similar to the results section of a journal article reporting the descriptive statistics (i.e. means, sds, and frequencies) of the variables then multiple regression analyses and interpretation. You are also asked to provide in a one to two paragraph concluding statement your recommendations on the importance of each policy for child achievement. In addition, in an appendix, the evaluation team would like to see all of your output.

Definition of Variables:
WICpreg – Women, Infant and Children (WIC) Nutrition Program participant during pregnancy: 0 = No, 1 = Yes.
AFDCpreg – Aid to Families with Dependent Children (AFDC) program participant during pregnancy: 0 = No, 1 = Yes.
readss97 – Woodcock-Johnson Revised Reading Achievement Test Age Standardized Score. Minimum = 47.5, Maximum = 165.5.
age97 – The child's age in 1997. Minimum = 3, Maximum = 13.
faminc97 – Total family income in 1997 (in 2002 constant dollars). Minimum = $-72296.26, Maximum = $784610.59.
bthwht – Low birth weight status of the child. 0 = non-low birth weight child, 1 = low birth weight child.
HOME97 – A composite total score of the emotional and cognitive stimulation at home. Minimum = 7, Maximum = 27.

Assignment 2: Assumptions Diagnostics & Violation Corrections (20%)
Due: Monday March 18th

You are still working as a research consultant for an evaluation study examining the effect of the Women, Infant and Children (WIC) Nutrition Program participation during pregnancy, but now on the raw scores of child mathematics achievement in 1997. Using the Child Development Supplement to the Panel Study of Income Dynamics (economic.good), the evaluation team has asked you to (1) diagnose the assumptions of the general linear model for the following model:

$$mathraw97 = B_0 + B_1 Age97 + B_2 fa\min c97 + B_3 bthwht + B_4 WICpreg + \varepsilon$$

that is, evaluate whether the assumptions of linearity, homoscedasticity, and normality of residuals are violated and (2) make any necessary and appropriate corrections. In addition, the evaluation team would like to assess whether parenting practices (HOME97) is a relevant variable for the model as well as diagnostics (and corrections, if necessary) for outliers and multicollinearity. The evaluation team would like for you to report the multiple regression model results before respecification/corrections and after respecification/corrections

The report should include: the descriptive statistics (i.e. means, frequencies, standard deviations, and correlations) of the variables; a multiple regression analysis before assumption diagnostics and corrections; plots, diagnostic statistics, and full discussion of diagnostics and corrections; and, lastly, a multiple regression analysis after respecification and corrections. You are also asked to provide in a one to two paragraph concluding statement your thoughts on the changes in the results of the before and after analyses and what they suggests. In an appendix, the evaluation team would like to see all of your output.

Definition of Variables:
- WICpreg – Women, Infant and Children (WIC) Nutrition Program participant during pregnancy: 0 = No, 1 = Yes.
- mathraw97 – Woodcock-Johnson Revised Mathematics Achievement Test Raw Score. Minimum = 0, Maximum = 98.
- age97 – The child's age in 1997. Minimum = 3, Maximum = 13.
- faminc97 – Total family income in 1997 (in 2002 constant dollars). Minimum = $-72296.26, Maximum = $784610.59.
- bthwht – Low birth weight status of the child. 0 = non-low birth weight child, 1 = low birth weight child.
- HOME97 – A composite total score of the emotional and cognitive stimulation at home. Minimum = 7, Maximum = 27.

Assignment 3: Interaction Effects & Incomplete Case Analysis (20%)
Due: Monday April 8th

You are still working as a research consultant for an evaluation study examining the effect of the Women, Infant and Children (WIC) Nutrition Program participation during pregnancy on the raw scores of child mathematics achievement in 1997. Using the Child Development Supplement to the Panel Study of Income Dynamics (economic.good), the evaluation team has asked you to examine if the effect of WIC program participation during pregnancy is moderated by (1) family income, (2) race, and (3) the current age of the child. Although assumptions still need to be evaluated it is not necessary to report the evaluation of the assumptions as was asked for in assignment 2. Simply state what changes were made to the data and model in order to account for assumption violations.

The report should include: the descriptive statistics (i.e. means, frequencies, standard deviations, and correlations) of the variables including a reporting of the amount of missingness; the multiple regression analyses with complete data and then with incomplete data. You are also asked to provide in a one to two paragraph concluding statement your interpretation of the results and what they suggest. In an appendix, the evaluation team would like to see all of your output.

Use the following syntax to recode chrace to a centered binary variable called race:
```
if chrace = 1 then race = .5; *white;
if chrace = 2 then race = -.5; *black;
if chrace > 2 then race = .;
```

Definition of Variables:
- WICpreg – Women, Infant and Children (WIC) Nutrition Program participant during pregnancy: 0 = No, 1 = Yes.
- Race – Centered Binary Coding of Race: -0.5 = Black, 0.5 = White.
- mathraw97 – Woodcock-Johnson Revised Mathematics Achievement Test Raw Score. Minimum = 0, Maximum = 98.
- age97 – The child's age in 1997. Minimum = 3, Maximum = 13.
- faminc97 – Total family income in 1997 (in 2002 constant dollars). Minimum = $-72296.26, Maximum = $784610.59.
- bthwht – Low birth weight status of the child. 0 = non-low birth weight child, 1 = low birth weight child.
- HOME97 – A composite total score of the emotional and cognitive stimulation at home. Minimum = 7, Maximum = 27.

Assignment 4: Application of the General(ized) Linear Model (40%)
Due: Wednesday April 28th by 5pm

You are now a full-time academic professor for a tier one research university and have a major program of research. For this major program of research you have received NIH grant funding, have collected your data or are using pre-existing data, and are now ready to begin analyses. Using any data set (if you do not have any data to work with you are more than welcomed to use the PSID good.sas7bdat data set, but you may also consider using the National Longitudinal Survey of Youth, the Health and Retirement Study, the National Longitudinal Study of Adolescent Health, or any other pre-existing national probability sample), you are to examine any particular research question(s) of interests that will call for you to use one of the applications of the general(ized) linear model covered the second half of the semesters. These applications include path modeling (using OLS regression), logistic regression, multilevel modeling, or growth modeling. You must use at least one application of the general(ized) linear model but are not limited to one.

In your report, you need to provide a brief introduction that frames and states your research question(s). Describe your data as you would in a methods section, report your analyses as you would in a results section, and provide a discussion of your interpretation and implications of the results. In the methods section you need to explain which modeling technique you are using, why it is appropriate for your research question(s), and present the equation for the model with all variables included in model. Assumptions need to be evaluated but it is not necessary to report everything done. Simply state if any assumptions were violated and what changes were made to the data and/or model in order to account for assumption violations. You also need to use an incomplete case analytic method. The results section should report: the descriptive statistics (i.e. means, frequencies, standard deviations, and correlations) of the variables including a reporting of the amount of missingness then the modeling with incomplete data. Please show all of your output in an appendix.

Given that this assignment is open to many research questions and the use of any data, please talk to the instructor first about what you plan to do. This assignment can be a good opportunity to work on the analyses for a journal article, an MA thesis, or an on-going research project. Please feel free to use this assignment as an opportunity to sharpen, strengthen, or simply complete the analyses for your existing on-going work.

GOOD LUCK & HAVE FUN! ☺