INTRODUCTION
From the major topic headings you will note that we will focus on three statistical procedures: OLS regression analysis, logistic regression, and factor analysis. But the course is primarily a course on multiple regression and related procedures that go by different names such as quadratic regression, path analysis, etc. This course will provide exposure to what multivariate statistics is all about. After you have taken this course you should know when to use each technique, how to set up Stata runs using the technique, how to interpret what you see in computer printouts using the technique, and how to clearly summarize what you have found in a written report. I also want you all to be able to critically evaluate books and journal articles using the techniques we cover. To that end you will be asked to read and critically evaluate the statistical analysis in published articles and books using each of these techniques. This year my plan is to also provide a quick overview of discriminant analysis in class, but I do not plan to require that you do your own analysis using that method (however I do provide the reading with the suggestion and hope that some of you will find the time to cover this topic on your own for extra credit. We will also cover how to deal with missing data using a technique called multiple imputation. It is a sophisticated way to deal with problem the of missing data for some variables for some cases.

COURSE REQUIREMENTS
Approximately every other week I will give an assignment. I strongly urge you to finish the assignment by the specified due date. In some cases, I will ask you to read and comment on an article, a book chapter, or a table in addition to doing and writing up Stata runs relevant to the material being covered. There will be 5 assignments for this course. Assignments 1 and 2 will be handed in together as your midterm paper a few days before the spring break and your mid-term grade will be based on that midterm paper. Similarly, for the second half of the course there will be three assignments (3, 4 and 5) that will be due one week after the end of classes for this course. These three assignments will be your final paper for the course.

I will suggest that you work on one of the datasets that I have made available for the course as it will save you a lot of time and be easier for me to understand your analysis. However, if you come to see me, it will, in some cases, be possible to obtain permission to use your own data at least for some of the assignments. I will provide two alternative datasets with a huge number of variables in each one. The choice of dependent variables and independent variables will be entirely up to you. When selecting variables keep in mind that most of the procedures you will be using will be based on the subset of countries for which you have no missing data for any of the variables you are considering. Thus, if you select variables with much missing data, there will be adverse consequences when it comes to testing for statistical significance.

GRADING
Grading will be based on my assessment of how much of the material we have covered that you seem to understand, how well you understand it, and how well you communicate that you understand it based on your mid-term and your final papers. This will include how much of what I cover in class you manage to cover in your statistical runs, how thoroughly, accurately, and clearly you write up the results of your statistical computer runs. It will also depend on how impressed I am by your comments on the articles that I assign for you to read and comment on. In class I want you to feel free to ask for clarification at any time or to ask what you might fear is a “dumb” question at any time without feeling that there will be a penalty for showing your ignorance. If you do not understand something that I seem to think is important, please ask, and the sooner you ask the better. Always feel free to interrupt me in class at any time.

You should be careful to allocate the space in your reports so that you at least touch on all of the important concepts and issues that come up in class. I am not going to give you a list. You need to be able to figure that out and you can discuss the issue with others in the class. As I need to be able to check your work, grading will be influenced by the clarity not only of the writing, but also the clarity of efforts to present relevant documentation from your Stata runs (including the relevant material from syntax files) to support assertions in the text of your paper. I need to be able to find that documentation very quickly.

It is entirely unacceptable to “borrow” another student’s assignments from this year or a prior year as a “guide” to your analysis. It is also not acceptable to hire a consultant who in essence does much of the assignment for you. You can ask anyone you like to help you get your own Stata runs to work, to help you understand material in the runs that I distribute in class, and how to write up results found in the outputs discussed in class. I encourage you to work with friends in the class on those things and to help one another on those tasks. But you should not be reading preliminary or final drafts of the answers from other students who have taken or currently are taking the course. Similarly, you should not be asking other students to read and make suggestions on
your early drafts to answers to questions in the various assignments that you will be dealing with this semester. Feel free to discuss with others in the class how you would answer the statistical analysis questions that I put to you based on the computer runs that I distribute in class. That will never get you in trouble. **However, all of your written work must be your own work.** If you suspect that what you are doing might be stepping over the line, come to see me and we can discuss whether or are doing or are considering doing is acceptable. Another suggestion if you are having trouble running or summarizing your findings based on your own variables, you can ask another person how the question would be answered based on the parallel output on the website and discussed in class.

If you are not fluent in English, you have permission (and encouragement) to get help with your English from the Connors Family Learning Center (617 552 0611) located in Room 200 in O'Neill Library. Please go to the following website for information about the issue of academic integrity: http://www.bc.edu/integrity. **I take this issue very seriously.**

**MID-TERM PAPER** (Due Tuesday February 27 @3:00 pm). This paper will consist of Assignments 1 and 2. It will make up 40% of your grade for the course. I will need both a hard copy MS word copy and an electronic MS word copy by this time and date. What you must write must be in 12-point font and can be either single or double spaced, and the body of your paper must not run more than 60 pages including the imbedded tables *(TYPICALLY YOU DO AN ELECTRONIC CUT AND PASTE FROM YOUR OUTPUT TO ADD A TABLE TO YOUR PAPER. I BELIEVE YOU WILL FIND THAT THERE IS A WAY TO DO THIS THAT ALLOWS YOU TO SIZE THE TABLE EITHER BEFORE OUR AFTER YOU INSERT IT INTO THE BODY OF YOUR WORD DOCUMENT. Before each question based on output tables that you will be submitting, insert the syntax used to generate the table(s) you will be discussing. Do not include in your paper any table that you will not be discussing. The 60-page upper limit on each paper is an upper limit, not a lower limit. The penalty for a late midterm will be minimal if I have both the hard and electronic copy by 9 am on February 28th. If it is more seriously late, but both versions are in by 3pm February 29, there will be no reduction in grade, but my comments will be more general and much less extensive. If you are very seriously late, but have it in by Tuesday March 13 at the latest, there will be a one grade level reduction in grade.

**FINAL PAPER** (Due Thursday May 10 @ 3:00 pm). This final paper will consist of Assignments 3, 4, and 5. It will make up 60% of your grade for the course. I will again require both a hard copy and an electronic copy. As with the midterm paper what you write must be in 12 point font, single or doubled spaced, and the body of your paper must not run more than 60 pages including the imbedded tables. Before each question insert the syntax used to generate the tables used to address that question. The penalty for a late paper will be minimal if in by 9 am on May 11. If it is more seriously late, but in by 3 pm on May 12: my comments will be more general and much less extensive. If your paper is in later than that, there will be a reduction of one grade level.

**SUGGESTED TEXTS**

1. Polit, Denise F. 2010. Statistics and Data Analysis for Nursing Research (2: Edition). Upper Saddler River, NJ: Pearson Education. (More Basic) *(Available at bookstore or on-line.* I suggest that you buy this book for the discussion of the various statistical methods. It will cover most, but not all of the statistical material.

2. Norusis, Marija J. 2012. IBM SPSS Statistics 19.0 Statistical Procedures Companion. Upper Saddle River, NJ: Pearson/Prentice Hall. This will be a very useful book as it will discuss various statistics that are presented in your output, but it is geared to SPSS output making it less useful than if it had been geared to Stata. That said, it will still be very useful. If you buy online, you may want to buy a less expensive (and less recent editions). The changes from edition to edition tend to be very modest and four our purposes they will be entirely cosmetic. I do not think you will run into problems if you are working with and edition of this manual based on the 16- or later edition. *(You may want to use Amazon.com to get a second-hand copy of this book.)*

3. Acocck, Alan C. 2012. A Gentle Introduction to Stata (Revised Third Edition). This book is recommended primarily for help with Stata commands of various sorts. You can also get help using the “help” facility that is built into the Stata software. I would recommend buying a copy of the third, fourth, or fifth edition. You should be able to get a second hand copy from amazon for under $20. It will be very useful to learn how to use the help facility in Stata and it might be preferable long term to use that for your Stata source.

**SUGGESTED FOR AN EXCELLENT, BUT A BIT MORE ADVANCED TREATMENT THAN THE WARNER BOOK:**

Tabachnick, Barbara G., & Linda S. Fidell. 2014. Using Multivariate Statistics. 6: Edition. Boston: Pearson/Allyn and Bacon. This text is suggested for a more advanced treatment than the other two statistics texts that I have discussed above. It is often cited as a source in other multivariate statistics books. Again, I would buy it second hand from Amazon. You would buy this book for the statistics, not the programing. It shows runs using SPSS and SAS, but not STATA. I would recommend the sixth or the fifth edition.

**SUGGESTED FOR SPECIFIC TOPICS**

Keith, Timothy K. 2006. Multiple Regression and Beyond. Upper Saddler River, NJ: Pearson Education. This book provides a thorough treatment of many of the issues we cover in this course. It is not written at an overly sophisticated level. You might find this book very useful for multiple regression, path analysis, and related procedures.

Namboodiri, Krishnan. 1984. Matrix Algebra: An Introduction. Sage. (GSSW). This is a good place to go if you want to learn the basics of matrix algebra. *(O’Neill Course reserves.* The Tabachnick, Barbara G., & Linda S. Fidell, 2014 source makes some use of matrix algebra, but most of the book is not based on matrix algebra and there is an appendix that provides a primer on matrix algebra.
TOPICS AND READINGS

Items on reserve at O’Neill are indicated by (O’Neill on Reserve) at the end of the citation. While just (O’Neill) is used to designate a book that is in O’Neill, but not on reserve for this course. Most assigned articles, STATA runs, “how to” documents, and datasets can be downloaded from the course Google Drive folders. It is best to download files after I tell you to, so that you will have this year’s (and not last year’s version) of the file.

Items that I suggest that everyone read are marked with an asterisk. Each of you will read items without an asterisk selectively depending on whether you are looking for a more basic or a more advanced treatment and depending on which version of STATA you are working with. I recommend that you start by reading either the Polit or the Norusis sources.

Bottom Line: I will be leaving it up to you to decide which of the various sources to use in connection with each procedure and topic. It will depend on how sophisticated a treatment you want. When I cover a technique, concept, statistic, option, etc., after class look it up and read about it in at least one of the sources that I have provided. If you are having trouble finding a relevant discussion in one of the two required books, see me for help in figuring out where to find relevant material to read. On a few highly specialized topics you will need to get it from the class notes alone or from an item on reserve at O’Neill. Past experience suggests that 80% of the students in the class will get 90% of the basic stats they need from the combination of the book by Polit and the book by Norusis. In general if you want more depth, try Tabachnick, Barbara G., & Linda S. Fidell. 2014. Using Multivariate Statistics. 6th Edition. Boston: Pearson/Allyn and Bacon. If you think you might want to go on to courses like SOCY7704, SOCY7705, SOCY7706, I would suggest that you spend at least some time with one of the more advanced sources such as Tabachnick & Fidell.

MULTIPLE REGRESSION & RELATED TECHNIQUES

OBJECTIVES: To be able to do multiple regression and path analysis using STATA. After we finish this section of the course you should understand such concepts as:

<table>
<thead>
<tr>
<th>slope</th>
<th>intercept</th>
</tr>
</thead>
<tbody>
<tr>
<td>least squares line</td>
<td>simple vs. multiple regression</td>
</tr>
<tr>
<td>regression coefficient vs. partial regression coefficient</td>
<td>the multiple correlation coefficient</td>
</tr>
<tr>
<td>the coefficient of determination</td>
<td>R-square</td>
</tr>
<tr>
<td>adjusted R-square</td>
<td>unstandardized partial regression coefficients</td>
</tr>
<tr>
<td>standardized partial regression coefficients</td>
<td>beta weights</td>
</tr>
<tr>
<td>t values</td>
<td>hierarchical multiple regression</td>
</tr>
<tr>
<td>stepwise multiple regression</td>
<td>forward inclusion</td>
</tr>
<tr>
<td>backward elimination</td>
<td>partial correlation</td>
</tr>
<tr>
<td>part correlation</td>
<td>collinearity</td>
</tr>
<tr>
<td>multicollinearity</td>
<td>tolerance</td>
</tr>
<tr>
<td>residual analysis</td>
<td>standardized residuals</td>
</tr>
<tr>
<td>assumptions about residuals</td>
<td>normality</td>
</tr>
<tr>
<td>linearity</td>
<td>homoscedasticity</td>
</tr>
<tr>
<td>partial plots</td>
<td>the construction and use of dummy variables</td>
</tr>
<tr>
<td>interaction terms</td>
<td>analysis of interaction</td>
</tr>
<tr>
<td>data transformation</td>
<td>quadratic regression</td>
</tr>
<tr>
<td>panel regression</td>
<td>path analysis</td>
</tr>
<tr>
<td>path coefficients</td>
<td>direct effects</td>
</tr>
<tr>
<td>indirect effects</td>
<td>spurious effects</td>
</tr>
<tr>
<td>causal analysis</td>
<td>path diagrams</td>
</tr>
</tbody>
</table>

You should be able to interpret a STATA do-file file for a multiple regression run and you should be able to write a STATA program to do multiple regression. In all assignments, please turn in the appropriate do-file file corresponding to any and all STATA output tables presented. Extract the tables that you will be discussing from the STATA output using copy and paste to insert the material into a word document at the point you discuss the output and results. Just prior to each output table (or set of tables) paste in the syntax (do-file) commands used to generate those tables.

READING:

Textbooks

*Polit, 2010:
Chapter 4: Preparing Your Data
Chapter 5: Transforming Your Data
Chapter 9: Correlation and Simple Regression.
Chapter 10: Multiple Regression (a good basic introduction.)
Chapter 14: Missing Values (Good discussion of alternatives for dealing with missing values)

*Norusis, 2012.

Chapter 4: Preparing Your Data (This is very basic. Read it when we discuss “data cleaning” several weeks into the course)
Chapter 11: Correlation
Chapter 12: Bivariate Linear Regression
Chapter 13: Multiple Linear Regression

Tabachnick & Fidell. 2014.

Chapter 4: Cleaning Up Your Act: Screening Data Prior to Analysis (good discussion of various issues related to getting the data ready for statistical analysis)
Chapter 5: Multiple Regression (a place to go for a more advanced and more detailed treatment of the topic than you will find in the Polit book). If the library does not have the 6th edition ask for the 4th edition (2001) or the 5th edition (2007) keeping in mind that the chapter numbers may change. This book provides a more advanced treatment of multiple regression than you will find in Polit. Do not be scared by your lack of ability to fully understand the sections that make use of matrix algebra. You will not need to understand the matrix algebra to understand most of what is discussed in this chapter. I do recommend that at some point you take the time to learn the basics of matrix algebra. I suggest you use the Namboodiri (1984) book below, but there are many other books that cover the topic just as well. For a very brief (I think too brief) introduction to matrix algebra see Appendix A of the Tabachnick & Fidell book. (O’Neill)


See this book for a more detailed overview of matrix algebra. (O’Neill course reserves)

Articles


Pampel, Fred C., John B. Williamson, and Robin Stryker. 1990. "Class Context and Pension Response to Demographic Structure in Advanced Industrial Democracies." Social Problems 37:535-550. (This is an example of multiple regression analysis checking for and reporting interaction effects.) (O’Neill) (MyFiles)


Shen, Ce. and John B. Williamson. 1999. "Maternal Mortality, Women’s Status, and Economic Dependency in Less Developed Countries: A Cross-National Analysis." Social Science & Medicine 49:197-214. (This is an example illustrating the use of multiple regression analysis and path analysis.) (O’Neill) (MyFiles)

Book

*Wilkinson, Richard and Kate Pickett. 2010. The Spirit Level: Why Greater Equality Makes Societies Stronger. New York: Bloomsbury Press. You will be asked to make use of this book some assignments. You can probably get a copy at a good price on Amazon.com. I have a few copies that I can borrow, but your will need to return the book so it can be used again next year.

LOGISTIC REGRESSION

OBJECTIVES: To be able to do logistic regression using STATA. After we finish this section of the course you should understand such concepts as:

<table>
<thead>
<tr>
<th>the logistic regression curve</th>
</tr>
</thead>
<tbody>
<tr>
<td>logit coefficients</td>
</tr>
<tr>
<td>standard error of the logit coefficient</td>
</tr>
<tr>
<td>Wald statistic</td>
</tr>
<tr>
<td>odds ratio</td>
</tr>
<tr>
<td>assessing goodness of fit of the model.</td>
</tr>
</tbody>
</table>

READING:

*Polit, 2010.

Chapter 13, pp. 389-412.
This is a simple introduction to Logistic Regression. (O’Neill)

*Norusis, 2012.

Chapter 16: Logistic Regression Analysis
This is a good statistical overview of logistic regression
Tabachnick & Fidell. 2014 (or 2007).
Chapter 10 Logistic Regression.
This is a more advanced treatment of logistic regression than you will find in Polit.

Articles:

Gay, David, and John Lynxwiler. 1999. "The Impact of Religiosity on Race Variation in Abortion Attitudes." Sociological Spectrum 19:359-377. (This is a simple example of logistic regression.) (O’Neill) (MyFiles)


DISCRIMINANT ANALYSIS (This year this material will be optional for extra credit)

OBJECTIVES: To be able to do discriminant analysis (also called discriminant function analysis) using STATA. You should understand such terms as:

<table>
<thead>
<tr>
<th>discriminant function</th>
<th>unstandardized canonical discriminant function</th>
</tr>
</thead>
<tbody>
<tr>
<td>discriminant score</td>
<td>coefficients</td>
</tr>
<tr>
<td>group centroids</td>
<td>standardized canonical discriminant function</td>
</tr>
<tr>
<td>eigenvalue</td>
<td>coefficients</td>
</tr>
<tr>
<td>Wilks’ lambda</td>
<td>structure matrix</td>
</tr>
<tr>
<td>canonical correlation</td>
<td></td>
</tr>
</tbody>
</table>

READING

Textbooks

*Polit, 2010.
Chapter 11, pp. 293-298: This is a simple introduction to discriminant analysis.

*Norusis, 2012.
Chapter 15: Discriminant Analysis
This is a good statistical overview of discriminant analysis and what you get when you run it in SPSS, but it does not say anything about how to set up your run.

Tabachnick & Fidell. 2014.
Chapter 9 (Chapter 11 in the 2007 edition): This is a more advanced treatment of discriminant analysis than you will find in Polit or Warner. (O’Neill on reserve)

Articles


FACTOR ANALYSIS

OBJECTIVES: To be able to do factor analysis using STATA. By the time we finish this topic you should understand such terms as:

<table>
<thead>
<tr>
<th>common factors</th>
<th>oblique rotation</th>
</tr>
</thead>
<tbody>
<tr>
<td>communality</td>
<td>factor scores</td>
</tr>
<tr>
<td>eigenvalue</td>
<td>factor score coefficient matrix</td>
</tr>
<tr>
<td>scree plot</td>
<td>principal components analysis</td>
</tr>
<tr>
<td>factor loadings</td>
<td>principal axis factoring</td>
</tr>
<tr>
<td>factor pattern matrix</td>
<td>plots of factor loadings</td>
</tr>
<tr>
<td>factor structure matrix</td>
<td>reproduced correlation matrix</td>
</tr>
<tr>
<td>factor rotation</td>
<td>anti-image correlation matrix</td>
</tr>
<tr>
<td>orthogonal rotation</td>
<td></td>
</tr>
</tbody>
</table>

READING

Textbooks

*Polit, 2010.
Chapter 13: Factor Analysis. This source provides a simple overview.

*Norusis, 2012.
Chapter 18: Factor Analysis. This is a good statistical overview of factor analysis and what you get when you run it in SPSS, but it does not say anything about how to set up your run.

Tabachnick & Fidell, 2014.
Chapter 13: This is a more advanced treatment of factor analysis than you will find in Polit. (O’Neill on reserve)

Articles


SUGGESTIONS FOR THE MID-TERM AND FINAL PAPERS (AS WELL AS ASSIGNMENTS 1-5)

1. It is very important that you include the tables (that I will need to check what you have done). Do not include tables that you do not discuss in your report. It is important to make it very easy for me to find the relevant material in the tables presented. Highlighting specific numbers in a table that you discuss in the narrative can help a lot.

2. I will give you a page limit for each assignment. So that I can read your work, I am going to require that in all assignments your writing be in a 12-point font, and the words and numbers in the tables must be very easy to read by me in hard copy (and I fine print hard to read!).

3. The most important time during the course to avoid getting behind is toward the end, specifically when doing assignments 4 and 5. Those assignments often take longer than the others. By that time in the semester the course will seem to be moving very quickly and most of you will experience increased time demands from other courses.

The solution: keep up, particularly toward the end of the course.

4. Does it make sense to try to show that you have learned some tricks that we did not cover in class? If you have the time to do some of the optional supplementary reading and if you understand what you have read and if you can find a way to do some computer runs using this new material, then be sure to find a way to show this in your assignments. If you elect to learn optional advanced material, it is up to you to learn this new material on your own or if you like working with another student in the class. If you do so and understand what you have read, then there will be some extra credit. Be sure to include a discussion of the runs to illustrate that you understand how to interpret what you have found. There is no extra credit for just doing the run that does something new. Also there is no extra credit for doing something early; that is, doing it before we get to it in class, so feel free to ask if we will be covering the topic later in the course. If you are not sure whether or not we are going to cover the topic in class later in the course, please ask me. When you are doing something in the hope of getting some extra credit for your efforts, be sure to put that material in bold to make sure I notice it and read it with care.

5. Why do extra credit material? (1) intellectual curiosity, (2) you will learn more from the course, (3) it will help compensate for credit lost due to errors that you do make or for the omission or inadequate coverage of topics that I feel are important, and (4) as an insurance policy (particularly for those who received an A on the mid-term paper). That said, there is typically at least one person in the class who manages to get an A on either the mid-term or the final paper without having done any extra credit work. This can happen if your presentation of the basic material covered in class is thorough and nearly flawless.

6. Is there a penalty for turning in a late mid-term for final paper? Except for very unusual circumstances (hospitalization, etc.) THERE IS A PENALTY when a paper comes in late. Think of it this way, (1) I do not want to create an incentive for turning in a paper late in a effort to get a better grade and (2) I want to protect students from the serious completion problems that in the past have come up when I have allowed people to turn in their assignments late. If you are considering turning in your assignment late, I urge you to first review the discussion of grading earlier in this syllabus. Due to the technical nature of what we cover in this course many students start to forget material that was once very clear. Within a few weeks after the end of the semester it can become very hard to finish assignments and after a few months, it becomes very hard.

7. If you find yourself in a situation in which it becomes obvious that you are not going to be able to finish your mid-term or for final paper on time, my advice is to focus on the most important material that we have covered and leave out some of the more difficult material. If you do not leave out too much of that material, you are very likely to pass albeit with a lower grade.