

**RUTGERS, THE STATE UNIVERSITY OF NEW JERSEY**

**EDWARD J. BLOUSTEIN SCHOOL OF PLANNING AND PUBLIC POLICY**

Department of Urban Planning and Policy Development

34:970:527, Fall 2006  
Advanced Multivariate Methods

Michael Greenberg  
Office: 932-4101 x673  
Home: 249-0222  
mrg@rci.rutgers.edu  
CSB 183

Teaching assistant  
Office hours to  
be determined  
at first class meeting

**Civic Square third floor computer lab (33 Livingston Avenue, next to the State Theater)**  
**Monday 6:00 pm - 8:40 pm**

**Purpose:**

This course will introduce some of the most useful multivariate methods and will specifically concentrate on linear-based models. The models I have chosen are factor, cluster, discriminant, and various forms of linear regression analysis, including hierarchical linear modeling. In addition to the methods I teach in this course, you should also learn various forms of logistic regression. These are taught by excellent faculty in the Bloustein School and the School of Public Health.

I want you to leave the course with an understanding of the methods and an ability to read the literature in your field. If you read an article that applies one of these methods, you should be in a position to assess if the application was appropriate. More important, I want you to be able to use these methods in your research. This means that I stress analysis and interpretation of actual data.

I will talk about and illustrate how you can bring multiple methods to bear to answer the same question, so that your results are not dependent on the method you chose.

The teaching philosophy in the course is hands-on interaction between student, data, methods, and computer. The course will not be a theoretical or mathematical introduction to advanced methods, although we will do some of the math. The course will require learning the weaknesses of the data, formatting data, and discussing different ways of minimizing the affects of data weaknesses.

We will consider such questions as:

1. What are the options when there are missing data?
2. What do you do when some of the data in a multivariate analysis is skew?
3. What are the advantages and disadvantages of transforming data from interval to ordinal and nominal scales?
4. What are the advantages and disadvantages of different methods?
5. What method(s) will enable you to most effectively test your ideas and theory?

At the end of the course, you will be able to use factor, principal components, and regression analysis in the case where components are used in independent variables, and you will have an

introduction to cluster analysis, discriminant, and hierarchical linear modeling. You will also have in introduction to SPSS.

Many more useful methods exist. I recommend a course by Dr. Radha Jagannathan in the Spring semester that fits nicely with this course. Also, the School of Public Health has a good set of courses.

### **Prerequisites:**

Knowledge of linear statistical methods including correlation and bivariate regression analysis. We will not start the course with review sessions.

### **Requirements:**

Presentation of two written reports on an analysis of a data chosen by you. The first report is about factor analysis. It should be 10 to 15 pages of writing. Printouts should accompany the writing (see below for more detailed suggestions). The paper should state the purpose of the analysis, describe relevant literature, explain the method(s), and present results. A discussion should include what you would do if you had more time and data.

Choose a data set of about 50-100 observations by 8 to 15 variables. Do pick a data set that is of interest to you. The data should include at least one variable that can be a dependent variable in a regression/correlation analysis. The first paper will be due somewhere between the 5th and 7th weeks (depending upon your progress). The other paper will be about the use of principal components in regression analysis. The instructions for the assignment are the same as for the factor analysis. It will be due between the 11th and 13th week. Class participation is quite important. You will learn a great deal from listening to each other. Student presentations are welcome. We will also spend a good deal of time using the computer lab.

Some students do not come with a data set. Consequently, I have posted some for you use. It is imperative that you take one of these data sets, or work with one that you have. The sooner you work with the data, the better.

### **Reading:**

All class members should eventually purchase the most recent SPSS-PC manuals. You need not purchase them for this class. These manuals are the base system (syntax reference guide), professional statistics, advanced statistics and base-systems user's guide. They turn out a new version every year. The newer versions have a few more methods but are not notably different from the older ones. In fact, version 7.5 works just fine on most programs. Eventually you need to get the new version because the company will not support the old versions. But you do not necessarily need to buy a new set of manuals. Periodically, a new application comes out, such as power analysis and missing data. These are usually one time purchases.

Those of you who are SAS and STATA fans should note that they work just fine, though SPSS is

better for some of these methods. However, SPSS is so widely available that you should use it. My experience is that some of these packages are better for one method but worse for others. I'll discuss this during the semester.

Required books about the methods: S.K. Kachigan, Multivariate Statistical Analysis, Radius Press, NY, 1991; and A. Afifi and V. Clark, Computer Aided Multivariate Analysis, Wadsworth, 1984 and later versions. Both of these books cover the methods we will cover. I will suggest specific reading assignments because the reading is obvious. The limitation of these books is that they are not adequate with regard to the underlying math. The books are available at the Rutgers Bookstore, located in the Ferren Mall on Albany Street (across from the New Brunswick train station). I have not asked you to buy any of the method-specific books because of their expense. If you want more methods-specific books, I'll be happy to recommend some, such as those below.

In addition, I have posted some papers that illustrate the use of these methods.

M. Greenberg, Individual-oriented and neighborhood protecting health actions: the critical role of environmental education seeking behavior, *The Environmentalist*, 23, 2003, 159-173. Good example of use of factor analysis.

M. Greenberg and K. Crossney, Perceived neighborhood quality in the United States: measuring outdoor, housing and jurisdictional influences, *Socio-Economic Planning Sciences*, in press. Factor analysis and multistage regression.

M. Greenberg, Environmental protection as a U.S. National priority: analysis of six annual public opinion surveys, 1999-2004, *Journal of Environmental Planning and Management*, 48(5), 733-746. Illustration of ordinal regression.

M. Greenberg and Reya Sinha, Government Risk Management Priorities: A Comparison of the Preferences of Asian Indian Americans and Other Americans. In review, *Risk Analysis*. This paper contains a useful application of Factor Analysis to create independent variables

Other valuable books include:

R.J. Rummel, Applied Factor Analysis, Northwestern, 1970 (strong mathematical presentation).

R.M. Thorndike, Correlation Procedures for Research, Gardner, 1978.

H.H. Harman, Modern Factor Analysis, U. of Chicago, 1960 and later versions (strong mathematical presentation).

L. Legendre and P. Legendre, Numerical Ecology, Elsevier, 1983 (ecology and biology examples).

G. Dunteman, Multivariate Analysis, Sage, 1984 (Useful mathematical presentation).

R.J. Johnston, Multivariate Statistical Analysis in Geography, Longman, 1978 (Nice overview that goes from simple statistics to multivariate methods).

C. Huberty, Applied Discriminant Analysis, Wiley, 1994 (satisfactory mathematical presentation).

D. Bartholomew, et al., The Analysis and Interpretation of Multivariate Data for Social Scientists.

CRC Press, 2002 (good on cluster and factor analysis).

### **Parking:**

After the normal beginning of semester chaos, all students normally are able to park in the underground parking garage beneath Civic Square (a right hand turn off of New Street, one block from Livingston Avenue), or elsewhere close to the building. This sometimes changes because of events and competition with the police department for parking spaces. Guess who wins that competition. Also, Rutgers does not own the parking garage, so we have relatively little leverage. But we are usually able to work out the parking.

### **Scheduling**

I know that some of you will miss some classes because of trips and religious holidays. Feel free to ask someone else for notes, to tape the class, and to set up an appointment with your teaching assistant.

### **Sequence of Topics**

#### **1. Introduction**

- a. Introduction of professor and students. What are your interests?
- b. Course requirements. Who should and should not be in the class.
- c. What is multivariate analysis? From exploratory analysis to forecasting.  
Classification: factor analysis, cluster, Chronbach's Alpha; Causal: discriminant, logistic, canonical, path, econometric, regression analysis, hierarchical linear models.
- d. Computer options, trying to simplify confusion: SPSS, SAS, STATA, BMD, mainframe, windows and DOS, lotus, excel, other special packages from public health and economics.
- e. Reading.
- f. Teaching order: examples, to printouts, to jargon, to math.
- g. What methods lead into multivariate methods: central tendency; dispersion; correlation; parametric and non-parametric.
- h. Some SPSS and simple methods: crosstabs and one-way a good place for you to start
- i. Posted data sets
- j. Student data sets

#### **2. What is Factor Analysis and Demonstration of SPSS-Windows**

- a. How to enter, save, and edit data. How to check on accuracy of data entry.
- b. Syntax: compute, recode and other commands.
- c. Data analysis: a simple introduction with cross-tabs and descriptive statistics
- d. Further discussion of students project
- e. Initial discussion of factor analysis if time permits, perhaps a demonstration in the

- computer lab with some data.
- f. Discuss weighting of data

**3. Understanding Underlying Dimensions of Your Data: Cronbach's Alpha, Factor Analysis, Principal Components, and Cluster Analysis (Classes 3-8)**

- a. Similarities and differences of the two methods.
- b. Examples of their use: examples, printouts, math.
- c. Class analysis
- d. Do a set of questions form a single scale? Correlations, Cronbach's and Factor Analysis demonstration of results
- e. PC and FA demonstrations on the computer for class in lab
- f. Mathematics of the method
- g. Different types of rotation
- h. Cluster Analysis
- i. Student presentations of factor analysis woven in to class as they are produced.

**4. When You Have a Continuous Dependent Variable: Regression and Component Analysis (Class 9-12)**

- a. Introduce method
- b. Examining descriptive statistics and associations
- c. Compare model with all variables to model with components to stepwise
- d. Interpreting and choosing a model
- e. Student projects – picking a dependent and independent variable

**5. When You Have a Categorical Dependent Variable: Discriminant, Logistic, and Ordinal Methods, and Hierarchical Linear Modeling (Classes 13-15)**

- a. Comparison with other methods.
- b. Examples.
- c. Student presentations.

# PROTOCOLS

## Data Management

### 1. Check for bad data

Out of range check by running frequency.

Eliminate cases and variables with 20+% missing data.

### 2. Filling- in missing data

Never use mean unless the data are normally distributed.

Use mode or median for categorical data and median for continuous data.

Regression-based program can fill in data. But be cautious.

If you have filled in a lot of missing data, be sure to run your analysis with and without it. Running analyses with missing data should produce more conservative results.

Code missing data with a unique value (I usually use -1). Careful with string variables. I'll show you how to recode them.

### 3. Transformations

Check frequency with frequency and/or plot.

Is the distribution normal? If so, no transformations are required.

Is the distribution skew? An option for negative skewness is to square the data. For positive (right skewness), the log base 10 is recommended. If these do not work, the data can be ranked, or broken into categories? I'm not a big believer in more exotic transformations, such as negative reciprocal root or even reciprocal because I'm not sure what the data mean. Is the distribution multi-modal? If so, categorical ranking.

Be especially careful about categorical variables that contain separate measures (e.g., 1=doctor, 2=lawyer, 3=nurse, 4=student). These are four separate dummy variables (1,0).

## Factor and Principal Component Analysis Protocol

1. What kind of relationships do you want to explore? Think about your research questions.
2. Run PC with all the defaults. If it ran and you don't have out-of-range crazy results, then move forward.
3. Examine the correlation matrix. Look for highly correlated variables ( $r \geq 0.4$ ). These will be part of factors.
4. Run factor analysis with eigenvalue set at 1. Did it run? If so, continue. If not, change the default number of iterations to 35, and then again 50 if it won't run. If it fails to converge, then go back to the principal components run. I'll explain why factor analysis sometimes does not work.
5. Examine the total variance explained table. What does it suggest?
6. Look at the unrotated matrix and interpret the factor loadings.
7. Use the regression method to save the factor scores. If they are places or times, they may mean something to you. Are they skew, try to figure out why? A factor score  $>4$  SDV from the mean can create a factor. You may need to eliminate it to get a more meaningful set of factors. You can map factor scores, chart them, save them and use them in other analyses.
8. Interpret the factors looking both at the variables and cases. Have you created anti-factor(s) by using variables that are defined from the same data? Sometimes the cases won't help you interpret the factors.
9. Now rotate the unrotated using the varimax criteria. Interpret the results? How do they compare to the unrotated solution?
10. Go back to the original extraction table? Was there a factor with an eigenvalue of 0.85 or 0.92? If so, tell the computer to give you another factor. Is it useful? Was there a factor with an eigenvalue of 1.1? If so, drop it and see what happens to the solution. Have you destabilized the solution by making these changes?
11. Now do an oblimin rotation, and look at the correlations among the factors. Are any correlated at  $r > 0.3$ ? If so, you may need to use this solution.

## **Suggested Outline for Factor Analysis Paper**

- Introduction: Context and general research questions (1-3 pages)
- Brief literature review (2-3 pages)
- Source of data: table or description of variables. Are you recoding them? What did you do with missing data?
- Table of runs (you do not have to detail every run, just the more important ones).
- Describe key run in detail (your correlation matrix, your factor loadings, your factor scores). most of your space is here.
- How did your most important run compare with the other runs, such as oblimin?
- Suggested tables:
  - (1) list of variables, explanation, coding, descriptive statistics
  - (2) correlation table
  - (3) summary of PC run
  - (4) summary of factor analysis run
  - (5) summary of varimax run
  - (6) summary oblimin run, including correlation of factors table
  - (7) illustrative factor scores

As you write the document, be sure to name the factors, clearly describe the factors and link them back to theory and the real world. We are looking for an insightful description. The paper should be about 15-20 pages, not including printouts. It should be double spaced and 12 point. See also factor analysis examples in the literature that are provided below.

## **Protocol for regression analysis using conventional single variables and principal components**

1. Choose a dependent variable that is continuous. If it is categorical, we'll need to use logistic regression or ordinal regression.
2. Why is this a good variable to choose?
3. What variables do you think are associated with this dependent variable? What theories justify your hypotheses?
4. How are these measured? Are they continuous, ordinal or dichotomous?
5. Run descriptive statistics on all the variables? Do any of the data need to be converted? Deal with any missing data.
6. Run a full model with all the variables entered. What were the strong associations, what were the weak? Do the results make sense?
7. Run a model with the independent variables entered by theoretical construct. If you have four constructs (e.g., location, socioeconomic status, family status, occupation), then you would have four runs. What do these results show?
8. Run a stepwise model with an entry level of  $p < 0.10$ . What was included and what was eliminated? Do the results adequately capture the impacts of the full set of variables, or are important variables excluded?
9. Run principal components analysis on independent variable set? Do the components capture key interactions and thereby more fully the underlying theory?
10. Picking a combination of factor scores and original data, run another regression. Is the overall  $r^2$  value lower? Are the results more interesting and compact?
11. Which works best for you in this case?

## **Suggested Outline for OLS Regression using Conventional and PC Variables Paper**

1. Introduction (1-3 pages) – Research questions and interests. Introduce data set.
  2. Literature review (2-3 pages) – theory, frame research questions, and how these are related to the data.
  3. Results
    - 3.1. Discuss data set, sources
    - 3.2. Describe variables, including descriptive statistics, levels of measurement, missing data, dependent vs. independent
    - 3.3. Bivariate correlation description
    - 3.4. Introduce OLS and interpret first enter model
    - 3.5. Describe theoretical construct models
    - 3.6. Describe stepwise model
    - 3.7. Introduce PCA results
    - 3.8. Describe PCA model
    - 3.9. Compare PCA results with the above in terms of clarity, utility, relationships to theory and expectations
  4. Discussion
- Future research – directions , additional questions. What would you do with more resources?

Notes: Paper will be 15-20pages, including all graphics. Double space and use 12 point.

### Suggested tables:

- (1) list of variables, explanation, coding, descriptive statistics
- (2) correlation table
- (3) regression of all variables enter run
- (4) regression of runs by theoretical concept
- (5) Stepwise regression of all individual data
- (6) PC of independent variables
- (7) Regression based on PC run
- (8) Analysis of regression residuals (if appropriate)

## Discriminant Analysis Protocol

1. *Define key variables*
  - 1a. Define dependent variables. Are there sufficient cases in each category? Make sure that you start with default priors=equal.
  - 1b. Define explanatory variables. Use simple bivariate tests as a screen, such as ANOVA and independent tests of means.
  - 1c. Look at discriminant analysis printout. Study means to get an idea of differences.
  - 1d. Go to table with F values at stage 0 to get strongest bivariate predictors.
  - 1e. Look at correlation matrix to determine interactions among independent variables.
  - 1f. Look at final variables in model to determine which variables are in the model.
  - 1g. Look at predictor or classification table to determine how well it predicted.
  
2. *Look at functions*
  - 2a. Look at canonical correlation and significance of each function.
  - 2b. Look at structure matrix
  - 2c. Look at group centroids. Integrate the results from the structure and centroid matrices.
  - 2d. Try to name functions.
  - 2e. Do you want to rotate? If so, add command rotate=structure.
  - 2f. Re-interpret the results if you have rotated.
  - 2g. Look at case-wise results, if they are relevant.
  
3. *Predict*
  - 3a. Look at unstandardized or standardized canonical discriminant function coefficients.
  - 3b. Look at Fischer's for raw data.
  - 3c. Predict, if appropriate.



## DATA SETS

1. New Jersey municipal data, 1990.

A data set I created for a variety of studies. It has all the NJ municipalities (n=567 at that time) and 31 variables. These include land area, water area, population size, racial/ethnic breakdowns, age, housing units, value and rent of units, family structure. Some of these are already converted to percents and means, other are not transformed. Particularly good data set for those interested in mapping.

2. Environmental public perception in NJ, 2004.

A sample of Non-Hispanic White respondents (n=431) from a survey I did in NJ. I dropped a lot of the variables, but left quite a few (see SPSS data file). We also have some location variables. Would be a good data set for those interested in working with survey data.

3. High school student behavioral data set, 2003.

A sample I drew of 1500 drawn from a 15000 sample survey. The behaviors include everything from heroin use to driving with seatbelts. You will also see a codebook (yrs 2003), which defines the variables. I removed the weighting and their summary variables.