This course is the second of a two-course sequence (BUSI 808, BUSI 809) that covers fundamental issues in conducting empirical research in the social sciences, focusing on business administration (e.g., organizational behavior, corporate strategy, marketing, international business, operations management, behavioral accounting) and allied disciplines (e.g., psychology, sociology). This course sequence is designed for doctoral students who intend to conduct empirical research publishable in scholarly journals. The course sequence is organized according to stages in the research process, from the initial framing of a research question through design, measurement, and analysis, and culminating in publishing the results of one’s work. This second course begins with model building and analysis, covering topics such as variance partitioning, categorical independent variables, testing and interpreting interaction and curvilinear effects, difference scores and their alternatives, multivariate procedures, categorical dependent variables, the study of change, and path analysis. The course then integrates measurement and model analysis through structural equations modeling. The course concludes with strategies for managing the publishing process. As with Applied Research Methods I (BUSI 808), students are assigned core readings, and, when appropriate, are given empirical data from published studies to apply the method under discussion. For each topic, students are also given an extensive annotated bibliography and a detailed written synopsis. As a prerequisite, students should have a full understanding of basic statistics, including correlation and regression analysis, and should have taken Applied Research Methods I or its equivalent.

Course Requirements:

Students are required to fully participate at weekly class meetings. Each meeting will include discussion of specific readings and reporting of any assigned homework, including data analyses. Students are expected to read all materials and complete all homework prior to class. At each class meeting, students should also turn in written answers to the assigned homework questions. These answers should be brief (usually 2-4 pages, single spaced) and to the point. When analyses are involved, written answers should include supporting results (e.g., excerpts from computer output, brief summary tables). These written answers serve two purposes. First, this course emphasizes hands-on application of various methodological tools, and for learning purposes, there is no substitute for using these tools and summarizing the results in writing. Second, written answers to homework questions provide an important means of monitoring learning throughout the course. When these answers show that certain topics are not well understood, these topics will be reviewed in class. Students may consult with each other regarding the concepts and principles underlying the methods used. However, written answers should represent the work of each individual student. Moreover, students must not obtain answers to homework questions from students who have previously taken this course; doing so is considered a violation of the honor code. Also, as a general assignment for all class meetings, all students should bring one or more discussion questions based on the assigned readings and should be prepared to initiate a discussion of the questions (these questions need not
be included in the written answers to the homework assignments). These questions will be addressed as time permits. Written answers to homework assignments will constitute 100% of the final grade.

For assignments that require data analysis, students must have access to statistical software. I will conduct analyses using SYSTAT, SPSS, SAS, and LISREL, all of which are available on the business school network. I will provide guidance regarding the use of this software (e.g., sample input files, answers to basic questions). However, students may also find it useful to seek help from documentation provided with this software, the helpdesk of the UNC Information Technology Services (962-5261, help@unc.edu), and web pages and listservs devoted to the software, which include the following:

SPSS: http://support.spss.com/  
SAS: http://support.sas.com/techsup/intro.html  
LISREL: http://www.ssicentral.com/lisrel/resources.html

Students may use other software (e.g., STATA) but will be responsible for converting the data files provided and for running the software.

**Required Reading Materials:**

Required reading materials for the course include books, book chapters, and journal articles. Books for the course are the following:


Pedhazur and Schmelkin (1991) will be the primary texts for the course. The books by Booth et al. (2008) and Williams (1995) provide background that will be very useful for written assignments for this class and papers for other degree requirements and for publication. Other reading material will consist of chapters and journal articles, which I will distribute via email.

**Summary of Topics:**

1. **Preparing Data for Analysis**: Making sure data is clean; detecting usual patterns; handling missing data.

2. **Categorical Independent Variables**: Dummy coding, effect coding, and contrast coding; costs of dichotomizing continuous variables.

3. **Interaction Effects and Multi-Group Analyses**: Meaning of interaction; estimating interactions between various combinations of categorical and continuous variables; higher-order interactions; scaling issues.
4. **Curvilinear and Piecewise Linear Effects**: Specifying and estimating polynomial regression equations; using piecewise linear regression equations to depict discontinuities and changes in slope.

5. **Difference Scores and Profile Similarity Indices**: Problems with difference scores and profile similarity indices; alternative procedures based on polynomial regression analysis and response surface methodology.

6. **Categorical Dependent Variables**: Problems of OLS regression with categorical dependent variables; discriminant analysis and its association with MANOVA and canonical correlation; estimation and interpretation of logistic regression equations.

7. **Multivariate Regression and Canonical Correlation**: Advantages of multivariate analysis over separate univariate analyses; canonical correlation as an extension of univariate regression, discriminant analysis, MANOVA, and principal components analysis; multivariate regression and multivariate hypothesis testing; set correlation.

8. **Repeated Measures and the Study of Change**: Advantages and disadvantages of repeated measures designs; the debate over the use of change scores; growth curve approach to the study of change; panel analysis.

9. **Mediation**: The meaning of mediation; simple and complex mediation models; the causal steps approach to mediation; mediation as indirect effects; procedures for testing mediated effects; combining mediation and moderation.

10. **Path Analysis**: Exogenous and endogenous variables; direct and indirect effects; spurious and unanalyzed associations; decomposition and reproduction of correlations; assessing model fit.

11. **Structural Equation Modeling I**: Comparisons between path analysis and structural equation modeling; alternative methods for handling measurement error; specification and identification of structural equation models.

12. **Structural Equation Modeling II**: Estimation of structural equation models; assessing model adequacy and fit; comparing alternative models.

13. **The Publication Process**: Preparing your manuscript; choosing an outlet; responding to reviews; reacting to rejection; celebrating acceptance.

*Indicates classes that involve data analysis.
Session 1: Preparing Data for Analysis

Required Reading:

Smith et al. (1986)

Abelson (1995), ch. 5

Pigott (2001)

Supplemental Reading:

Schafer and Graham (2002)

Graham (2009)

Assignment Questions:

1. Read the Smith et al. (1986) article, which describes 14 rules for generating clean data. How would you implement each of these steps in your own research? Have Smith et al. (1986) omitted any steps that you consider particularly important? Next, read the chapter from Abelson (1995), and again consider how the points raised in the chapter would apply to your research. Are there other potential sources of fishiness that we should add to those identified by Abelson (1995)?

2. Read the article by Pigott (2001). What are the relative advantages and disadvantages of the various methods for handling missing data? Which of these advantages and disadvantages are more important, and why? What can be done to minimize missing data in the first place?
References for Preparing Data for Analysis


Session 2: Categorical Independent Variables

Required Reading:

Pedhazur and Schmelkin (1991), ch. 19

Cohen (1983)

Questions:

1. Read chapter 19 from Pedhazur and Schmelkin (1991), paying particular attention to the procedures for dummy coding, effect coding, and contrast coding (which is discussed in the section on multiple comparisons among means). How should the regression coefficients from these procedures be interpreted? Next, read Cohen (1983), which concerns the costs of dichotomizing continuous variables. In your opinion, why do researchers continue to dichotomize continuous variables, despite the drawbacks of this procedure?

2. The files MEDICAL.SAV, MEDICAL.SYD, and MEDICAL.XLS contain mortality rates (deaths per 100,000 people) from various causes for the 50 states, broken down by region (numbered 1-4) and census division (numbered 1-9). Analyze differences in death rates due to cancer across regions, using dummy coding, effect coding, and contrast coding. For the contrast coding, construct three contrasts, the first testing whether cancer is higher along the east coast, the second testing whether it is higher where sunbathing is more prevalent, and the third testing whether it is related to cost of living (hint: each contrast compares two regions with two others, and the three contrasts are mutually orthogonal). The map on the following page may be useful to you. Draw conclusions that are appropriate for each of these analyses.

3. Test all pairwise comparisons between the mean death rates for accidents (ACCIDENT), cardiovascular disease (CARDIO), cancer (CANCER), and pneumonia/flu (PNEU_FLU) across the four regions. One way to conduct these tests is by letting each region take its turn as the group for which all of the dummy variables equal zero. Which regions differ from one another on the four death rates?

4. Using dummy coding, test differences in death rates due to accidents, cardiovascular disease, cancer, and pneumonia/flu, first by region and then by division. For each cause of death, test whether separating regions into divisions is justified by comparing the $R^2$ values from the region and division equations, using the following formula:

$$F = \frac{(R^2_D - R^2_R)/(df_R - df_D)}{(1 - R^2_D)/df_D}$$

where $R^2_D$ is the $R^2$ from the division equation, $R^2_R$ is the $R^2$ from the region equation, $df_D$ is the residual degrees of freedom from the division equation, and $df_R$ is the residual degrees of freedom from the region equation. To obtain a probability value for the $F$ statistic, use $df_R - df_D$ as the numerator degrees of freedom and $df_D$ as the denominator degrees of freedom. What conclusions can you draw? If it is justified to separate the regions into divisions for any of the outcome variables, test the pairwise comparisons among the divisions for that outcome, using dummy coding and letting each division take its turn as the group for which all dummy variables equal zero. What do these tests tell you?
5. Compute a correlation matrix of the seven death rates (i.e., ACCIDENT, CARDIO, CANCER, PULMONAR, PNEU_FLU, DIABETES, LIVER). Then, dichotomize each variable at its mean and recompute the correlation matrix (to calculate means in SPSS, select Analyze, Descriptive Statistics, Descriptives; in SYSTAT, select Statistics, Descriptive Statistics, Basic Statistics). What differences do you see between the correlation matrices? How do these differences correspond to what you would expect, based on Cohen (1983)? How would your substantive conclusions differ, depending on which correlation matrix you examined?
References for Categorical Independent Variables


Session 3: Interaction Effects and Multi-Group Analyses

Required Reading:

Cohen (1978)

Jaccard et al. (1990)

Edwards (2008)

Questions:

1. Read Cohen (1978), which discusses the basic principles of using product terms to represent interactions in multiple regression. Next, read Jaccard et al. (1990), which explains how to compute and test simple slopes using estimates from regression equations with product terms. Finally, read Edwards (2008), which addresses various misconceptions concerning the analysis and interpretation of interaction effects in regression analysis.

2. Last weekend, I experimented with the effects of salt on generating bubbles in two beers, Guinness and Heineken. I bought two six packs of each and chilled them to the recommended temperature. I then chose two identical glasses, rinsed them with cold water, filled one with Guinness and the other with Heineken, waited 30 seconds, and took a photograph of both glasses in the same lighting conditions. I then rinsed the glasses again, filled them with beer, added 1/4 teaspoon of salt, waited 30 seconds, and took another photograph. I repeated this procedure until all the beer was gone. I counted the bubbles in the glasses in the photographs and put the data the files BEER.XLS, BEER.SYD, and BEER.SAV. As you will see, the data are organized to make it easy to dummy code the beer and salt variables. Analyze this data to determine which beer reacts more to the salt, and to what degree. Change the coding of beer and salt from 0,1 to -.5,+.5. How do the results differ? How do you interpret both sets of results?

3. The files WFINT.XLS, WFINT.SYD, and WFINT.SAV contain data from 1,397 working men and women who completed measures of anxiety (ANX), the amounts of control, relationship quality, and security at work (WACON, WAREL, WASEC, respectively), the amounts of control, relationship quality, and security with the family (FACON, FAREL, FASEC, respectively) and the importance of these six attributes (WICON, WIREL, WISEC, FICON, FIREL, FISEC). Gender (GEN) was coded 0 for men and 1 for women. Using these data, answer the following questions:

   a. The amount of control, relationships, and security people experience can provide a sense of stability and therefore can affect anxiety. Using six separate regression equations, analyze these effects for control, relationships, and security for both work and family, and interpret your results.

   b. Research on gender roles suggests that men and women may place different emphasis on work and family. Therefore, the amounts of control, relationships, and security might have different relationships with anxiety for men and women. Test this notion with six moderated regression equations, using gender as the moderator variable. If you find significant interactions for gender, compute the intercepts and slopes for men and women. What do these results mean? You might find it helpful to plot the functions for men and women.
c. One explanation for gender differences in reactions to work and family experiences is that men and women ascribe different levels of importance to work and family experiences. Therefore, the effects of control, relationships, and security on anxiety could differ not because of gender per se, but because of the importance of work and family experiences. Using six moderated regression equations, test whether the effects of control, relationships, and security for work and family depend on the importance of these attributes. If you find significant interactions, compute intercepts and slopes at low and high values of the importance variable (as before, you may want to plot these functions). What can you infer from these analyses?

d. Overall, what conclusions can you draw regarding the effects of gender and importance on the relationships of work and family control, relationships, and security on anxiety?

Note: To plot simple slopes, you can use the file SIMSLOPE.XLS, which will be sent to you with the data. Be sure to read the formulas in the cells that generate the values used to produce the plots so you understand how the plotted lines correspond to the regression coefficients you estimate.
References for Interaction Effects and Multi-Group Analyses


Session 4: Curvilinear and Piecewise Linear Effects

Required Reading:

Excerpt from Neter et al. (1989), ch. 10

Pedhazur and Schmelkin (1991), pp. 451-459

Supplemental Reading:

Pedhazur (1982), ch. 11, pp. 404-430

Questions:

1. Read the excerpt from the Neter et al. (1989) chapter on piecewise linear regression. Next, read the indicated pages from Pedhazur and Schmelkin (1991). What are the relative advantages and disadvantages of piecewise linear regression and polynomial regression for the analysis of nonlinear effects?

2. The files NONLIN.SYD, NONLIN.SAV, and NONLIN.XLS contain data from 185 MBA students who were interviewing for jobs. For each job, the students reported satisfaction with six job attributes (commute time, pay, span of control, closeness of supervision, travel, and variety) and perceived discrepancies for the same attributes, meaning the degree to which the job provided more or less of each attribute than the student considered adequate (variable names each job attribute are listed on the following page). The response scale for the satisfaction measure ranged from 1 (very dissatisfied) to 7 (very satisfied), and the response scale for the perceived discrepancy measure ranged from -3 (much less than adequate) to +3 (much more than adequate). With these data, conduct six piecewise linear regression analyses using the procedure on pp. 370-372 of Neter et al. (1989), testing the hypothesis that satisfaction is greatest when actual amount equals adequate amount and decreases when actual amount deviates from adequate amount in either direction. When coding the dummy variable for each job attribute, consider different options for coding cases for which the discrepancy equals zero. Do the regression functions appear symmetric around the point of zero discrepancy? Can you think of a way to test whether each function is symmetric? Next, test whether the functions are discontinuous, using the approach on pp. 373-374 of Neter et al. (1989). Which functions showed evidence of discontinuity? How do these results affect your thinking about whether the hypothesis has been supported? To help you interpret your results from both sets of analyses, you can plot the functions you obtained using the Excel file NONPLOT.XLS.

3. Reanalyze the job attributes using polynomial regression, again focusing on the hypothesis that satisfaction is greatest when actual amount matches adequate amount. Do your results support the hypothesis? How do these results differ from those from the piecewise linear regression analyses? For each job attribute, which polynomial function best describes the relationship between the perceived discrepancies and satisfaction? Again, to help you interpret your results, you can plot the functions you obtained using the Excel file NONPLOT.XLS.

4. Taking all your analyses into account, what can you conclude regarding the relationship between perceived discrepancies and satisfaction?
Variable Names for Nonlin Data

Each variable begins with three letters indicating the job attribute. These letters and the corresponding job attributes are as follows:

COM = Commute time
PAY = Pay
SOC = Span of control
SUP = Closeness of supervision
TVL = Travel
VAR = Variety

The fourth letter of each variable name is either D (perceived discrepancy) or S (satisfaction). Thus, AUTD is the perceived discrepancy for autonomy, and AUTS is satisfaction with autonomy.
References for Curvilinear and Piecewise Linear Effects


Session 5: Difference Scores and Profile Similarity Indices

Required Reading:

Edwards (2002)

Edwards (2001)

Supplemental Reading:

Edwards (1994)

Edwards & Parry (1993)

Questions:

1. Read Edwards (2002), paying particular attention to problems with algebraic and squared difference scores and the procedures proposed as alternatives to these scores. Then, read Edwards (2001), which attempts to dispel misconceptions regarding the use of difference scores, polynomial regression, and response surface methodology.

2. The files CAP.SYD, CAP.SAV, and CAP.XLS contain measures of job dissatisfaction (DIS), workload dissatisfaction (WLDIS), and boredom (BOR), and scale-centered measures of actual and desired amounts of job complexity (COMEC, COMPC) and quantitative workload (WOREC, WORPC) from the classic person-environment fit study conducted by Caplan et al. (1980). Construct scores representing the algebraic and squared difference between actual and desired job complexity and actual and desired quantitative workload. Next, calculate the correlation between these difference scores and the three measures of strain (DIS, WLDIS, BOR). What do these correlations seem to suggest?

3. Conduct confirmatory tests of the models underlying the algebraic and squared difference scores, as described in Edwards (2002, p. 363). You can test constraints using the LMATRIX in the GLM procedure of SPSS, the AMATRIX in the MGLH procedure of SYSTAT, or by conducting an F-test of the difference in $R^2$ values from the constrained and unconstrained equations. You can test sets of higher-order terms by comparing the $R^2$ values for equations of different orders (e.g., linear versus quadratic, quadratic versus cubic) using an F-test. For each difference score, which of the four conditions were satisfied, and which conditions were rejected?

4. For the models that do not survive the confirmatory test, determine the equation (linear, quadratic, cubic, etc.) that best fits the data. What is your interpretation of this equation? You can plot surfaces for each equation using the Excel file SURFACE.XLS, which requires you to enter the coefficients from the unconstrained equation and the minimum and maximum values of the X and Y scales. Note that you can rotate the resulting surface by right-clicking the area just outside the box that surrounds the surface and selecting 3D-View. Alternately, you can use the fplot command in SYSTAT (see syntax on reverse side of this sheet), and you can rotate the surface using View, Dynamic Explorer. These plots should help you interpret the results.
SYSTAT syntax for plotting functions:

This command plots a simple 3-D surface:

```
fplot z=3+.5*x+.3*y+.25*x^2 .5*x*y+.3*y^2;xmin=1,xmax=5,ymin=1,ymax=5,axes=9
```

This command plots a 3-D surface but reverses the scaling of the X and Y axes:

```
fplot z=3+.5*x+.3*y+.25*x^2 .5*x*y+.3*y^2;xmin=1,xmax=5,ymin=1,ymax=5,axes=9,xrev,yrev
```

This command plots a 3-D grid surface, like those in Edwards (1994) and Edwards and Parry (1993):

```
fplot z=3+.5*x+.3*y+.25*x^2 .5*x*y+.3*y^2;xmin=1,xmax=5,ymin=1,ymax=5,axes=9,surface=xycut
```

Note that, for all commands, the variables x, y, and z can be replaced with names that convey the meaning of a variable (e.g., "actual", "desired", "boredom"). Also, be sure to specify the correct minimum and maximum values for the independent variables used in the analyses.
References for Difference Scores and Profile Similarity Indices


Session 6: Categorical Dependent Variables

Required Reading:

Dillon & Goldstein (1984), ch. 10

Pindyck & Rubinfeld (1981), ch. 10

Questions:

1. Read the Dillon and Goldstein (1984) chapter, focusing on the basic concept of discriminant analysis, the calculation of discriminant functions, classification rules, and tests of significance for individual variables, subsets of variables, and all variables collectively (this will take you through p. 375). I would like to note two issues with the chapter that may create confusion. First, the chapter describes Mahalanobis' $D^2$ as the distance between group centroids. In fact, the distance between group centroids is actually Mahalanobis' $D$ (i.e., the square root of Mahalanobis' $D^2$). You can obtain Mahalanobis' $D^2$ by squaring the difference between the group centroids reported by SPSS. Use this value in Equations 10.2-6 and 10.3-3 in Dillon and Goldstein (1984) (although the authors report the correct equations, they incorrectly use Mahalanobis' D rather than $D^2$ in their examples). Second, the chapter states that the discriminant function weights are calculated as:

$$b = S^{-1}(\bar{x}_1 - \bar{x}_2)$$

where $b$ is a $p \times 1$ column vector of discriminant function weights, $S$ is the $p \times p$ pooled within-group covariance matrix, and $\bar{x}_1$ and $\bar{x}_2$ are $p \times 1$ column vectors of the means of the $p$ variables for groups 1 and 2, respectively. The chapter does not mention that these weights are usually scaled such that:

$$b'Sb = 1$$

This is accomplished by calculating $b'Sb$ for the original weights, taking the square root of this value, and dividing the original weights by the obtained square root.

2. Read the Pindyck and Rubinfeld (1981) chapter, focusing on the basic concept of using a categorical dependent variable, the drawbacks of ordinary least-squares when the dependent variable is dichotomous, and the principles underlying probit and logit models (just skim the examples and the section on multiple-choice models). Note that you will be using maximum likelihood logit for this assignment, which is mentioned on p. 294 and discussed further in the appendix. Also, the formula for the $R^2$ on p. 312 should be $1 - \log L_0/\log L_{\text{max}}$ (i.e., the correct formula uses logs of the likelihood values rather than the likelihood values themselves). You need not compute this $R^2$ (called a pseudo $R^2$) for the assignment, but please note this correction.

3. The files ENTRESUB.SAV, ENTRESUB.SYD, and ENTRESUB.XLS contains a categorical dependent variable indicating whether the respondent has engaged in entrepreneurial activities (i.e., started a business), two demographic variables (age and gender, coded 1 for male and 0 for female), four socioeconomic variables (years of education, household income, number of entrepreneurs known by respondent, employment status), and four work values (autonomy, challenge, wealth, home), based on data from 555 residents of Wisconsin. Use this data to address three research questions: (1) to what extent do demographics predict entrepreneurial activities; (2) to what extent
do socioeconomic factors predict entrepreneurial activities, after taking demographics into account; and (3) to what extent do work values predict entrepreneurial activities, after taking demographics and socioeconomic factors into account. To address the research questions, run three sets of analyses:

a. Run the analyses using OLS (i.e., standard regression analysis). Examine the results. What anomalies do you see? (Hint: See if any information available under the "Save..." icon within linear regression might be useful.)

b. Conduct three discriminant analyses (see "Statistics", "Classify", "Discriminant"), entering predictors according to the hierarchical scheme indicated by the research questions. How well does each discriminant function perform in terms of classification accuracy? (Hint: check the options available under the "Classify..." icon.) Is each function significant? Using formula 10.3-3 from Dillon and Goldstein (1984, p. 374), test the significance of adding each set of predictors. Finally, determine whether the individual predictors are significant. What do these tests tell you?

c. Use logistic regression (see "Analyze", "Regression", "Binary Logistic"), testing predictors in blocks according to the hierarchical scheme (note that the predictors are labeled "Covariates"). How well does each logistic regression equation perform in terms of classification accuracy? Determine the significance of each equation, each set of predictors, and each individual predictor. Note that the equation can be tested using the chi-square statistic reported in the output, and sets of predictors can be tested using a chi-square difference test drawing from the chi-square statistics reported for equations with and without the predictors. What do these tests tell you? How do the results of the logistic regression differ from those obtained from the discriminant analysis? Taking both sets of analyses into account, what conclusions can you draw?

Variables in ENTRESUB file:

ENTRE Has started a business (0=no, 1=yes)
AGE Age (in years)
GENDER Gender (0=female, 1=male)
EDYRS Years of education
HHINC Household income (in thousands of dollars)
NETSIZE Size of network (i.e., number of entrepreneurs known by respondent)
EMP Employed (0=no, 1=yes)
AUT Work value: Autonomy
CHA Work value: Challenge
WEL Work value: Wealth
HOM Work value: Home
References for Categorical Dependent Variables


Session 7: Multiple Dependent Variables

Required Reading:

Dwyer (1983), ch. 8

Thompson (1991)

Sherry and Henson (2005)

Questions:

1. Read the Dwyer (1983) chapter, focusing on pp. 163-188. Then read the Thompson (1991) and Sherry and Henson (2005) articles, paying close attention to the interpretation of the results from the example. I suggest you reproduce the Thompson example using SYSTAT, SPSS, or SAS, using the data in THOMPSON.SYD, THOMPSON.SAV, or THOMPSON.XLS, using the syntax on the last page of this assignment.

2. The files TAHEALTH.SYD, TAHEALTH.SAV, and TAHEALTH.XLS contain data from a study by Edwards and Baglioni (1991) involving four measures of health (anxiety (ANX), depression (DEP), somatic complaints (SOM), self-reported angina (ANG)) and four measures of Type-A components, including the speed and impatience from the Bortner (BORSI), hard-driving and competitive from the Bortner (BORHC), time pressure from the Framingham scale (FRATP), and hard-driving and competitive from the Framingham scale (FRAHC). All measures used 5-point scales with higher scores indicating greater amounts of each variable. Research suggests that the components of Type-A behavior are detrimental to health, and your task is to determine whether this finding is supported by the data provided to you.

Conduct a MMR analysis, with the four health measures as dependent variables and the four Type-A measures as independent variables. In SYSTAT, the commands for estimating the model are as follows:

MGLH
MOD ANX,DEP,SOM,ANG=CONSTANT+BORSI+BORHC+FRATP+FRAHC
ESTIMATE

To test each independent variable and all four variables as a set, use the HYPOTHESIS subroutine and enter an AMATRIX to select the appropriate coefficients. The syntax for testing all four coefficients is:

HYPOTHESIS
AMATRIX [0 1 0 0 0;
         0 0 1 0 0;
         0 0 0 1 0;
         0 0 0 0 1]
TEST

To test each independent variable separately, conduct four additional hypothesis tests, each using one row of the AMATRIX shown above. These HYPOTHESIS commands must be entered after you have estimated the model.
Commands for conducting a MMR analysis in SPSS are as follows:

**GLM**

```plaintext
anx dep som ang WITH borsi borhc fratp frahc
/METHOD = SSTYPE(3)
/INTERCEPT = INCLUDE
/PRINT = DESCRIPTIVE PARAMETER
/LMATRIX = "OMNIBUS" borsi 1; borhc 1; fratp 1; frahc 1
/DESIGN = borsi borhc fratp frahc .
```

The **LMATRIX** statement conducts the same test indicated by the **AMATRIX** in SYSTAT. Without the **LMATRIX** statement, SPSS will report tests of the independent variables individually and jointly, but like the SYSTAT **AMATRIX**, the **LMATRIX** also allows you to test subsets of independent variables (e.g., you could test two of the four independent variables using an **AMATRIX** with two rows or a **LMATRIX** statement that lists two of the independent variables).

Using either SYSTAT or SPSS, what do the analyses tell you about the relationships between the independent and dependent variables?

3. Conduct a CCA of the data in the TAHEALTH.SYD file. For these analyses, you can modify the SYSTAT, SPSS, or SAS commands from the Thompson example. Summarize your results by constructing a table analogous to Table 3 in Thompson (1991). Conduct significance tests of the canonical correlations as follows:

   a. Calculate Wilks' lambda \( W \) for all four canonical correlations using the following formula. This value should match the Wilks’ lambda reported in your output.

   \[
   W = (1 - R_{c1}^2)(1 - R_{c2}^2)(1 - R_{c3}^2)(1 - R_{c4}^2)
   \]

   b. Calculate Bartlett's chi-square using the following formula:

   \[
   \chi^2 = -[N - 1 - .5(p + q + 1)]\log_e(W)
   \]

   where \( N \) is the sample size, \( p \) is the number of independent variables, \( q \) is the number of dependent variables, \( W \) is Wilks' lambda, and \( \log_e \) is the natural logarithm. The degrees of freedom are \( pq \).

   This is an omnibus test of all four canonical correlations as a set. If the test is significant, you should conclude that at least the first canonical correlation is significant. To test the remaining canonical correlations, calculate another Wilks' lambda, this time omitting the first canonical correlation, and use this value to calculate Bartlett's chi-square. The degrees of freedom will now be \( (p - 1)(q - 1) \). Repeat this procedure in the same fashion to test the remaining two canonical correlations. Focus your interpretation on the coefficients from only those variates that yielded significant canonical correlations. What information is yielded by these analyses?

4. Based on your analyses, what are the comparative advantages of MMR and CCA in determining the relationships between two sets of variables? Which method yields results that are more readily interpretable, and why?
Canonical Correlation Syntax for Thompson Example

SYSTAT:

SETCOR
MODEL CHA6 INT6 OTH6 = CHA2 INT2 OTH2
ESTIMATE

SPSS:

INCLUDE '[installdir]/Samples/English/Canonical correlation.sps'.
CANCORR SET1=varlist1 /
    SET2=varlist2 /.

The two variable lists must be separated with a slash.
[installdir] is the installation directory.

SAS:

PROC CANCORR DATA=THOMPSON ALL;
VAR CHA6 INT6 OTH6;
WITH CHA2 INT2 OTH2;
RUN;
References for Multiple Dependent Variables


Session 8: The Study of Change

Required Reading:

Liker et al. (1985)

Finkel (1995), ch. 2

Bryk and Raudenbush (1987)

Questions:

1. Read Liker et al. (1985), focusing on the distinctions between the first-difference approach (Eq. 2) and the two-wave, two-variable (2W2V) approach which uses the initial level of the dependent variable as a predictor (Eq. 11). Next, read the chapter from Finkel (1995), noting that the unconditional change score model (Eq. 2.2) and the static-score or condition change model (Eq. 2.5) correspond to the first-difference approach and 2W2V approach, respectively, described by Liker et al. (1985). Finally, consider the hierarchical linear model approach described by Bryk and Raudenbush (1987), paying particular attention to the distinctions between the within-subject and between-subject models and the logic of using parameter estimates from the former model as outcome variables for the latter model. Overall, which approach do you find most compelling for the study of change? Are there specific conditions that determine when one approach should be preferred over the others?

2. The files SDE.SYD, SDE.SAV, and SDE.XLS contain data from a four-wave panel study of 73 second-year MBA students collected monthly during the semester in which the students were seeking full-time employment. Each month, the students described the perceived amount of opportunity for self-development in their current job prospects (PSDE) on a 7-point scale ranging from none at all to very much, and also described their feelings regarding the amount of self-development opportunity available to them (FSDE), ranging from -3 (extremely distressed) to +3 (extremely delighted). Reliabilities for the PSDE measures across the four panels were .864, .915, .903, and .886, and reliabilities for the FSDE measures across the four panels were .910, .927, .928, and .947. The file also contains intercept and slope coefficients (A0, A1, respectively) for the within-subject model in which the four FSDE measures were estimated using a linear function with time coded 0, 1, 2, and 3 for the four time periods. Other variables in the file include an ID number and GENDER (0 = female, 1 = male). Using these data, assess the effect of perceived self-development opportunities on feelings regarding self-development opportunities, using the first-difference (unconditional change) approach, the 2W2V (static-score) approach, and the hierarchical linear model approach. Specific guidelines for each set of analyses are as follows:

   a. Use the first-difference approach to examine the effects of change in perceived self-development opportunities on change in feelings regarding self-development opportunities. Focus your analyses on changes between measures from adjacent waves (i.e., wave 2 vs. wave 1, wave 3 vs. wave 2, wave 4 vs. wave 3). What conclusions can you draw from these analyses? How do the reliabilities of the change scores compare to the reliabilities of the separate measures? To answer this question, recall the formula for the reliability of a difference:
\[ \alpha(x-y) = \frac{\sigma_x^2 \alpha_x + \sigma_y^2 \alpha_y - 2 \sigma_{xy}}{\sigma_x^2 + \sigma_y^2 - 2 \sigma_{xy}} \]

where \( \alpha_x \) is the reliability of \( X \), \( \alpha_y \) is the reliability of \( Y \), \( \sigma_x^2 \) is the variance of \( X \), \( \sigma_y^2 \) is the variance of \( Y \), and \( \sigma_{xy} \) is the covariance between \( X \) and \( Y \). Given that the coefficient linking two change scores actually represents the relationships among four variables, can you think of a procedure for revealing these relationships and testing the assumptions embedded in the change scores?

b. Apply the static-score model. What do the results tell you about the effects of perceived self-development opportunities on feelings regarding self-development opportunities and changes in these feelings over time? What appears to be the appropriate time lag for the effects of perceived self-development opportunities on feelings regarding self-development opportunities? Compare your results from those obtained using change scores as dependent variables. Is there evidence for regression toward the mean? Can you test the constraint on the coefficient on \( Y_{t-1} \) that is imposed when a change score is used as a dependent variable?

c. Apply the HLM approach. To what extent are perceived self-development opportunities related to the intercept and slope of the function describing initial status and change in feelings regarding self-development opportunities? Do the intercept and slope differ for men and women? What do these results tell you?

d. Taking all of your analyses into account, what conclusions can you draw regarding the effects of perceived self-development opportunities on feelings toward self-development opportunities?
References for The Study of Change


Chan, D. (1998). The conceptualization and analysis of change over time: An integrative approach incorporating longitudinal mean and covariance structure analysis (LMACS) and multiple indicator latent growth modeling (MLGM). *Organizational Research Methods, 1*, 421-483.


Namboodiri, N. K. (1972). Experimental designs in which each subject is used repeatedly. *Psychological Bulletin, 77*, 54-64.


Session 9: Mediation

Required Reading:

Baron and Kenny (1986)

MacKinnon et al. (2007)

Supplemental Reading:

Shrout and Bolger (2002)

Questions:

1. Read Baron and Kenny (1986), focusing on pp. 1176-1179. Consider the criteria recommended by Baron and Kenny (1986) for establishing complete and partial mediation and the regression equations for evaluating these criteria. Do you think all of the criteria are equally important? Are some of the criteria unnecessary? Are there situations in which applying the criteria can lead to incorrect conclusions? Next, Read MacKinnon et al. (2007) and consider their recommendations for testing mediation. How do these recommendations differ from those of Baron and Kenny (1986)? To what extent are these recommendations similar to your evaluation of the Baron and Kenny (1986) procedure?

2. The files MEDIATE.SYD, MEDIATE.SAV, and MEDIATE.XLS contain data from 970 working adults who completed measures of: (a) the degree to which the culture of the organization values pay (PAYO), security (SECO), authority (ATHO), and prestige (PREO); (b) the degree to which the organization rewards its employees with pay (PAYH), security (SECH), authority (ATHH), and prestige (PREH); and (c) job satisfaction (JOBSAT), turnover intentions (TOI), organizational identification (IDENT), and organizational citizenship behavior (OCB), which refers to prosocial actions by employees that help the organization. Research suggests that the cultural values of an organizational influence how the organization rewards its employees and that culture and rewards both influence outcomes such as job satisfaction, turnover intentions, organizational identification, and organizational citizenship behavior. Hence, rewards may be viewed as a mediator of the effect of organizational values on outcomes.

   a. Using the criteria set forth by Baron and Kenny (1986), analyze rewards as a mediator of the effect of values on outcomes. You should conduct 16 separate analyses in which organization values cause rewards for each of the four content dimensions (e.g., pay values cause pay rewards) which in turn cause each of the four outcomes. Based on the Baron and Kenny (1986) criteria, which of these analyses provide evidence for complete mediation, partial mediation, or no mediation?


   c. Overall, based on your analyses, what do you conclude about rewards as a mediator of the effect of organizational values on the outcomes you examined?
References for Mediation


Session 10: Path Analysis

Required Reading:

Pedhazur (1982), ch. 15

Questions:

1. Read ch. 15 from Pedhazur (1982). Make sure you understand the calculation of path coefficients, the decomposition of correlations into their four components, and testing overidentified models for fit.

2. The files AMBSTD.SAV, AMBSTD.SYD, and AMBSTD.XLS contain data from 211 respondents who completed measures of ambition, competitive drive, exposure to time pressure, hurried/impatient behavior, and anxiety (all variables are standardized). The attached two figures show path models indicating relationships among these variables. The first model (Theoretical Model A) shows that ambition (AMB) leads to competitive drive (CD) and time pressure (TP). Competitive drive leads to hurried/impatient behavior (HI), as well as to time pressure, which also influences hurried/impatient behavior. Finally, time pressure and hurried/impatient behavior both influence anxiety (ANX). Using the data provided, please analyze this model as follows:
   a. Calculate the coefficients for each path in the model, including the paths that relate the theoretical variables to one another and the paths for the residuals.
   b. Decompose the ten correlations among the five variables into direct effects, indirect effects, spurious components, and unanalyzed components.
   c. Assess the fit of the model, including the overall model fit using the formulas on p. 619 of Pedhazur and the correspondence between the observed correlations among the variables and the reproduced correlations yielded by summing the direct effects, indirect effects, spurious components, and unanalyzed components from question 2b.
   d. Do your analyses support each of the hypothesized paths in the model? Do your analyses confirm that the omitted paths should be excluded from the model?
   e. Prior analyses showed that the reliabilities for the AMB, CD, TP, HI, and ANX measures are .663, .793, .631, .629, and .745, respectively. What implications do these reliabilities have for the results of the path analysis?
   f. Draw any substantive conclusions you think are warranted from your analyses.

3. Theoretical Model B differs from Theoretical Model A in that ambition and competitive drive are treated as correlated exogenous variables (i.e., ambition does not cause competitive drive, but rather these two variables are correlated causes of other variables in the model). Calculate path coefficients for this model, decompose the ten correlations among the five variables into direct effects, indirect effects, spurious, and unanalyzed components, and retest the fit of the model. How do these results differ from those you obtained for Theoretical Model A? How do these differences influence the substantive conclusions you could draw?
Theoretical Model A: One Exogenous Variable
Theoretical Model B: Two Exogenous Variables
References Path Analysis


Session 11: Structural Equation Modeling I

Required Reading:

Pedhazur & Schmelkin (1991), ch. 24

Review:

Anderson and Gerbing (1988)

Hu and Bentler (1999, pp. 27-28)

Jackson et al. (2009, pp. 9-11, pp. 17-20)

Questions:

1. Read the Pedhazur & Schmelkin (1991) chapter. Make sure you understand the meaning of the three primary equations involved (i.e., the equations representing the measurement models for the exogenous and endogenous variables and the equation representing the structural relationships among the exogenous and endogenous variables) and why path analysis may be viewed as a special case of structural equation modeling.

2. Using the correlation matrix and reliability estimates from the data for the path analysis assignment, calculate disattenuated correlations among the five variables. Recall that a disattenuated correlation is calculated by dividing the original correlation by the square root of the product of the reliability estimates of the two measures:

\[ \phi_{12} = \frac{r_{12}}{\sqrt{r_{11} \cdot r_{22}}} \]

where \( \phi_{12} \) is the disattenuated correlation between \( X_1 \) and \( X_2 \), \( r_{12} \) is the correlation between \( X_1 \) and \( X_2 \) measured with error, and \( r_{11} \) and \( r_{22} \) are reliability estimates of \( X_1 \) and \( X_2 \), respectively. You can enter the disattenuated correlations into a SYSTAT worksheet by selecting `Data` from the `Window` menu, entering the variable names (AMB, CD, TP, HI, ANX) in the top row (i.e., the unnumbered row), entering 1's along the diagonal, and entering the disattenuated correlations below the diagonal. Save the file by choosing `Save As` under the `File` menu, specifying `Correlation` under `Matrix Type`, and specifying a path and file name (e.g., AMBDIS.SYD). In SPSS, create a file with seven columns, the first two containing string variables named `rowtype_` and `varname_`, and the remaining five containing the numeric variables AMB, CD, TP, HI, ANX. The contents of the `rowtype_` variable (reading down the column) should be `N`, `MEAN`, `STDDEV`, `CORR`, `CORR`, `CORR`, `CORR`, `CORR`. The `varname_` variable should contain three blank cells followed by AMB, CD, TP, HI, ANX. The first three cells for the variables AMB, CD, TP, HI, ANX should be 211, 0, 1. The remainder of the file should contain a square symmetric matrix of disattenuated correlations with 1s down the diagonal. Save the file, specifying an appropriate path and file name (e.g., AMBDIS.SAV). To help you set up these files, I have created SYSTAT and SPSS files with the observed correlations (AMBPA.SYD and AMBPA.SAV, respectively), and you can edit these files by entering the disattenuated correlations you compute. The SYSTAT syntax for conducting a regression analysis from a correlation matrix is:
Note that you enter the sample size at the end of the MOD statement. The corresponding SPSS syntax is:

```
REGRESSION MATRIX=IN(*)
/VARIABLES=AMB CD TP HI ANX
/DEPENDENT=ANX
/METHOD ENTER=AMB CD TP HI.
```

You can modify the SYSTAT and SPSS shown above to estimate the required regression equations. Using the disattenuated correlation matrix, reestimate the path coefficients for the theoretical and saturated versions of model A from the previous assignment. How do the results using the disattenuated correlations differ from those you obtained using the raw data? Based on these results, how would you alter your substantive conclusions regarding the relationships depicted in model A?

3. The file AMBPA.LS8 contains LISREL commands that will conduct a path analysis of theoretical model A. Run the command file using LISREL 8 (note that the three values reported for each parameter are the parameter estimate, its standard error, and a t-test, which is computed by dividing the parameter estimate by its standard error). How well does the model fit the data? Are parameter estimates significant and in the expected direction? Compare the results to those yielded by the path analysis you performed with SPSS or SYSTAT. What additional information does the LISREL output provide?

4. The file AMBPADIS.LS8 contains two sets of LISREL commands for estimating theoretical model A correcting for measurement error. The first set of commands uses disattenuated correlations based on the formula shown in question 2, whereas the second set directly incorporates reliability estimates into the model, such that the ith latent variable has a single indicator with its loading set at $(r_{ii})^{1/2}$ and error variance set at $1 - r_{ii}$. Run the command file using LISREL 8. As before, determine how well the model fits the data and whether the parameter estimates are significant and in the expected direction. Compare the results to those yielded by the path analysis you performed with SPSS or SYSTAT using the disattenuated correlations you calculated. How do the three sets of results compare? Which results do you think are most trustworthy, and why? Overall, what can you conclude regarding the validity of model A?
Session 12: Structural Equation Modeling II

Review:

Anderson and Gerbing (1988)

Hu and Bentler (1999, pp. 27-28)

Jackson et al. (2009, pp. 9-11, pp. 17-20)

Supplemental Reading:

MacCallum and Austin (2000)

Questions:

1. The file AMBMM.LS8 contains LISREL commands that will generate a five-factor measurement model based on the indicators for the five variables used in the path analysis (items are listed on the next page). Analyze the measurement model. Do you see any anomalous results? What do the item loadings and factor correlations tell you? What information can you obtain from the modification indices and residuals? How well does the model fit the data?

2. The file AMBOR.LS8 will produce an orthogonal measurement model in which the five latent variables are forced to be uncorrelated. Compare this model to that yielded by AMBMM.LS8. To what degree has the fit of the model deteriorated? Conduct a chi-square difference test to compare the two models (i.e., calculate the difference in the chi-squares for the two models, and refer this quantity to a chi-square table, with df equal to the difference in df for the two models). What does this tell you?

3. The file AMBSE.LS8 will produce estimates for a structural equation model corresponding to the path analysis from AMBPA.LS8. Compare the results from the structural model to those from the two measurement models, focusing on the completely standardized parameter estimates. Are the item loadings stable? Have any problems emerged? How well does each model fit the data? Can you quantify the comparison of the structural model to the two measurement models using chi-square difference tests? Next, examine the structural parameters. How do these results compare to those yielded by the path analysis, the analysis of the model using disattenuated correlations, and the model that directly incorporated reliability information? Which paths are supported? Which are not? Why?

4. Rerun the structural model, sequentially deleting or adding structural paths among the latent variables and stopping when the modification indices and t-tests suggest no further changes (limit your model modifications to the gamma matrix and parameters below the diagonal of the beta matrix). How do the new results compare to the original results? How have the parameter estimates changed? Overall, how well does the model fit the data? How does the fit of this model compare to that yielded by the original structural model and the two measurement models? Taking all of your analyses into consideration, what substantive conclusions can you draw?
Items for five variables used in path analysis:

Ambition:
1. Ambitious
2. Strong need to excel
3. Take work too seriously
4. Put forth more effort than others
5. Take work more seriously than others

Competitive Drive:
1. Hard-driving and competitive
2. Hard-driving and competitive when younger
3. Hard-driving and competitive now
4. Spouse says I am hard-driving and competitive

Time Pressure:
1. Usually pressed for time
2. After work, pressed for time
3. After work, feel stretched to my limits
4. Trouble finding time for a haircut
5. Faced with daily deadlines at work

Hurried/Impatient Behavior:
1. Always rushed
2. Emphatic speech (may pound desk)
3. Act immediately under pressure
4. Always hurry
5. Do most things in a hurry
6. Hurry more than others

Anxiety:
1. Feel upset for no obvious reason
2. Felt as though I might faint
3. Feel uneasy and restless
4. Feel really panicky
5. Would say I am a worrying person
6. Feel 'strung-up' inside
7. Have the feeling I am "going to pieces"
8. Have bad dreams which upset me when I wake up
References for Structural Equation Modeling


Session 13: The Publishing Process

Required Reading:

Graham and Stablein (1995)
Schneider (1995)
Ketchen (2002)
Seibert (2006)
Campbell (1982)

Four sets of reviews and replies

Questions:

1. Write a list of questions and concerns regarding publishing that you face now or expect to face in your early career. The chapter by Graham and Stablein (1995) should help stimulate your thinking.

2. Read the Schneider (1995) chapter. How well do his propositions address your questions and concerns? Do his propositions raise any additional issues?

3. The articles by Ketchen (2002), Feldman (2004), and Seibert (2006) provide specific advice for handling the review process, and the article by Campbell (1982) discusses the review process from an editor's perspective. What are the most important ideas that you take away from these articles?

4. Attached are four sets of exchanges between myself and the reviewers and editors of various journals regarding four manuscripts that were eventually published. These exchanges involve theoretical, methodological, and empirical manuscripts, tell stories of rejection and acceptance, and illustrate the long and often difficult path from submission to publication. For each set of reviews, focus on the dialog between me and the editors and reviewers, and skim the substantive comments in the reviews and my replies to these comments. What do the exchanges tell you about the review process?

5. Based on your assessment of these issues, write your personal "Top 10 Keys to Publishing in Scholarly Journals." Have fun, but be prepared to provide a rationale for each of your points. Bring copies of your "Top 10" list to distribute to everyone (you can make copies at the copy machine near the classroom).
References for The Publishing Process


Eichorn, D. H., & VandenBos, G. R. (1985). Dissemination of scientific and professional knowledge:
Journal publication within the APA. *American Psychologist, 40*, 1309-1316.


Jauch, L. R., & Wall, J. L. (1989). What they do when they get your manuscript: A survey of Academy of


