Intermediate / Advanced Research Design & Statistics

Robert Ploutz-Snyder, Ph.D.
Biostatistician NASA JSC
USRA / Division of Space Life Sciences,
Research Associate Professor of Medicine
SUNY Upstate Medical University
About Biostatistics

HACD Biostatistics is one of several research elements comprising the Human Adaptation and Countermeasures Division (HACD) at the Johnson Space Center. This laboratory provides statistical consulting to HACD and the Space Medicine Health Care Systems Office (SMHCSO), provides opportunities for high school and college students to be directly involved in the analysis and interpretation of biomedical research at NASA, and conducts independent research to address the special challenges raised by the idiosyncrasies of data often gathered on small numbers of human subjects under non-standard environments and test regimens.

Statistical Consulting

Biostatistics provides consulting expertise, mainly to the HACD research laboratories, in the application of statistical theory and practice to ongoing biomedical research. Laboratory personnel offer aid in the preparation of parts of research proposals that describe the experimental design, statistical modeling and subsequent analysis of anticipated research data. Once data is gathered, BSL statisticians also can assist with analysis and interpretation of results to help the investigators extract the most information consistent with the goal of maintaining statistical integrity. A BSL statistician may in fact be a co-investigator in projects requiring sophisticated statistical modeling and/or analysis techniques and will be expected to contribute descriptions of these techniques in forthcoming research papers. In these instances, the participating BSL statistician would be included as a co-author of such papers. Being involved as a consultant to other Biostatistics research laboratories provides an excellent opportunity for the BSL statistician to expand her/his knowledge base in such diverse medical fields as environmental physiology, osteopathy, neurology, pharmacology, microbiology, cardiology, nutrition and psychology. Although HACD research laboratories are the laboratory's main customer, consulting support is also provided to the SMHCSO in support of NASA flight operations.

Outreach

Although the primary customers for the BSL reside within the Office of Biostatistics, statistical consulting support is occasionally given to other organizations within the Johnson Space Center, such as the Engineering Directorate and Human Resources and Education Office. The BSL also provides a venue under which high school or college students, as summer interns, can be directly involved in the analysis and interpretation of NASA biomedical research data.

Statistical Consulting...
Previously at SUNY Cortland

- Built and Directed the Office of Institutional Research & Assessment

- Three Primary Functions of the Office

  - Institutional Research
    - Ex. Enrollment Management
    - Ex. Salary Studies

  - Institutional Assessment
    - General Education
    - Program Majors
    - Administrative Units

  - Institutional Data Warehousing & Reporting
    - SUNY/NYS
    - Middle States & Other Accrediting Bodies
    - The Usual Hodgepodge of others...
More Importantly, *Who are YOU?*

- How many Directors of IR Offices?
  - New Directors?
- How many Associate/Assistant Directors?
  - New to your position?
- How many IR Analysts with 5+ years experience?
- How many IR Analysts with less than 5 years experience?
- Other??
Purpose of This Module

• To provide Institutional Researchers with an understanding of the principles of advanced research design and the intermediate/advanced statistical procedures consistent with such designs
You Will Learn How To Use

- Independent Measures Analysis of Variance (ANOVA)
- Repeated Measures ANOVA/MANOVA Analysis of Covariance (ANCOVA)
  - ANOVA with Covariates
- Simple and Multiple Regression
  - Block Regression
  - Forwards, Backwards, Stepwise Regression
You Will Also be Exposed To

• Exploratory Factory Analysis
  – Principal-Axis Factoring
    • With Varimax Rotation

• Time Series Regression
I Assume That

• This isn’t your first course in statistics?
• That you have the basics covered
  – Foundations I Level Stats
• That you have SPSS loaded on your laptop machines
• That you are interested and motivated to learn!
Format of This Module

- Hands-On!
  - I’ll talk about the statistical tests, assumptions, theory first
  - Then we’ll walk through analyses together
    - Using SPSS
  - We’ll pay a lot of attention to
    - Analytic choices that you have
    - Interpreting the output
    - Presenting the results to your constituents
Quick Review: Gaussian Distribution Function

- A.K.A. The “Normal Distribution”
- A.K.A. The “Bell-Shaped Curve”
- Has known probabilities associated with it,
- Thus all Parametric Statistics are based on the Gaussian Distribution

\[ f_g(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\bar{x})^2}{2\sigma^2}} \]

Where \( \bar{x} = \text{mean} \), and \( \sigma = \text{standard deviation} \)
Quick Review:
Gaussian Distribution Function

- About 68% of all scores fall within 1 SD unit from the mean.
Quick Review:
Gaussian Distribution Function

• About 68% of all scores fall within 1 SD unit from the mean.
• About 95% of all scores fall within 2 SD units from the mean.
Quick Review: Gaussian Distribution Function

- About 68% of all scores fall within 1 SD unit from the mean.
- About 95% of all scores fall within 2 SD units from the mean.
- About 99% of all scores fall within 3 SD units from the mean.
Central Limit Theorem

• States that for any population with mean $\mu$ and standard deviation $\sigma$, the distribution of sample means with sample size $n$ will approach a normal distribution with $\mu$ and SD of $\frac{\sigma}{\sqrt{n}}$ as $n$ approaches infinity.

• REGARDLESS of the shape of the distribution in the population.

• By the time sample sizes hit around 30, sampling distribution of means is close to normal.
Demo of central limit theorem.
• Since we know so much about the Normal Distribution
• And we know that sample summaries (means or otherwise) tend to follow that distribution
  – Even data collected from non-normal samples
  – Especially so with large sample size (big-n)
• We can usually apply our knowledge of the normal distribution to statistical comparisons, estimates, and probability
  – As long as we do some preliminary screening...
Ex. You may recall...

- That we can compare a person’s score to their population with the “Z-Score”
  - Z is a “standardization”
  - Mean = 0
  - SD = 1
  - Probability tables tell us percentiles, probabilities of being “that far” away from the mean.
Z-score quick review

- A student takes a standardized test and scores XXX.
- Can compare their score to the population of all test-takers during that time, given population mean and standard deviation as:

\[ Z = \frac{X - \mu}{\sigma} \], where

\( \mu \) is population mean and \( \sigma \) is population std dev
Z-score quick review

- With their Z-score, we can glean
  - Their percentile rank \[ p(\text{lower}) \]
  - Probability of scoring higher than them
  - Other relevant probabilities.
T-statistic for Comparing Sample to Population

• Where we don’t know SD of the population, but we have sample data
  – Thus sample mean and sd

• And we know from CLT that we can estimate population SD by SE

\[ SE = \frac{s}{\sqrt{n}} \]
T-statistic for Comparing Sample to Population

• Given our sample data, we can calculate
  – Sample mean
  – Sample SD (s)
• Given SE formula
  \[ SE = s_x = \frac{s}{\sqrt{n}} \]
• We can calculate Confidence Intervals on this Estimate also
• And with t-tables…p-values

\[ t = \frac{\bar{X} - \mu}{s_x} \]
Moving to the t-test for comparing two samples

- Used for comparing two samples collected randomly from two populations

- Fairly simple modifications of the t-statistic comparing sample mean to population mean

\[
\begin{align*}
\bar{x}_1 &= \frac{1}{n_1} \sum_{i=1}^{n_1} x_{1i} = 2 \\
\bar{x}_2 &= \frac{1}{n_2} \sum_{i=1}^{n_2} x_{2i} = 6 \\
S_1^2 &= \frac{\sum_{i=1}^{n_1} (x_{1i} - \bar{x}_1)^2}{n_1 - 1} = 2 \\
S_2^2 &= \frac{\sum_{i=1}^{n_2} (x_{2i} - \bar{x}_2)^2}{n_2 - 1} = 2 \\
S_p^2 &= \frac{SS_1 + SS_2}{df_1 + df_2} = \frac{df_1 s_1^2 + df_2 s_2^2}{df_1 + df_2} \\
t &= \frac{\bar{x}_1 - \bar{x}_2}{S_{\bar{x}_1-\bar{x}_2}}
\end{align*}
\]
Dissect the formula: \( s_{-X_1-X_2} \)

\[
t = \frac{(\bar{X}_1 - \bar{X}_2)}{s_{-X_1-X_2}}
\]

where \( s_{-X_1-X_2} = \sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}} \)

\[
s_p^2 = \frac{SS_1 + SS_2}{df_1 + df_2} = \frac{df_1 s_1^2 + df_2 s_2^2}{df_1 + df_2}
\]
Dissect the formula: Numerator

\[ t = \left( \frac{X_1 - X_2}{S_{X_1 - X_2}} \right) \]

The difference between two sample means
Dissect the formula: Denominator

\[ t = \frac{\left( \bar{X}_1 - \bar{X}_2 \right)}{S_{\bar{X}_1-\bar{X}_2}} \]

- The difference between two sample means
- Divided by some measure of standard error of the differences
Dissect the formula: Question?

\[ t = \frac{(\bar{X}_1 - \bar{X}_2)}{S_{\bar{X}_1 - \bar{X}_2}} \]

The difference between two sample means

Divided by some measure of standard error of the differences

Are the differences that I see between my two means unusual, given variability among other sample means of this size?
T-tests on the Computer:

- Software gives us t-score and a p-value
- Allowing us to test hypotheses that the two samples come from the same population (or not)
- And describe the magnitude of the differences (confidence intervals)
- Ex. $t = 4.87$, $p < .001$
  - $H_{null}$: Two samples are from same population
  - $H_{alt}$: Two samples are from different populations
- Reject the Null (alpha < .05) & Report the magnitude of the differences
Virtues of the t-test

- EVERYONE seems to understand it!
- With CLT, it’s easy to apply to lots of different data scenarios
- There are other versions that make it very flexible
  - Formula for “Repeated Measures” designs
  - Formula for problems associated with non-normality and/or variance heterogeneity
Limitations of t-tests

• Alpha risk is .05 for each t-test
  – Probability of falsely rejecting the null, and concluding that there is a difference, when it’s really due to chance.
  – So comparing 3, 4, 5 or more groups is quite problematic!

• With large samples, as with ANY statistical test, “significance” does not necessarily indicate a meaningful difference.
Comparing Three Groups

Group I

Group 2

Group 3
Comparing Three Groups

Group 1

Group 2

Group 3

T-test number 1
Alpha risk = .05
Comparing Three Groups

Group 1
T-test number 1
Alpha risk = .05

Group 2
T-test number 2
Alpha risk = .05

Group 3
Comparing Three Groups

- Group I
  - T-test number 1
  - Alpha risk = .05

- Group 2
  - T-test number 2
  - Alpha risk = .05

- Group 3
  - T-test number 3
  - Alpha risk = .05
Comparing Three Groups

Group I

Group 2

Group 3

T-test number 1
Alpha risk = .05

T-test number 2
Alpha risk = .05

T-test number 3
Alpha risk = .05

Total Alpha Risk ~ .15
Analysis of Variance (ANOVA)

• Can compare unlimited number of groups or occurrences, and still keep alpha risk = .05
• Able to take multiple grouping (or time) factors into account and determine their independent and combined effects
• Can examine “trends” in data, and can test specific (often complex) hypotheses
• The analytic focus is on variance, but the interpretation falls back to means—thus results become intuitive
Assumptions Required of ANOVA

- Data collected randomly from the population, with roughly equal n per cell
  - And sufficiently large n (n>30, common r-o-t)
- Data measured on interval or ratio scale, and is normally distributed
- Homogeneity of variance across groups
- Sphericity for RM designs—variance of the differences between means for any pair of groups is equal to any other pair
Assumption of Randomly Collected Data with Sufficiently Large $n$

- In IR, we don’t always “randomly select”
  - But can we assume that “today’s” data is a random representation of “recent years?”
  - Or can we START randomly selecting a subset of your populations for research?

- How big is big enough?
  - Rule of Thumb... at least 30 per group
  - More is better
    - Cautions about overpowered studies...
  - But BALANCE is critical!!
    - Rule of thumb—smallest group should not be less than 1/3rd the size of the largest group.
Assumption of Interval or Ratio Scale & Normality

• The “bell-shaped” curve—assumption of all parametric statistics
• Studies show that ANOVA is robust to violations of this, but only if sample size is substantially large, and Homogeneity is met
Assumption of Homogeneity of Variance Across Groups

• Variance on the dependant variable should be similar across groups
  – Why?

• Because we’re examining VARIANCE in ANOVA, and so we need for variance in each group to be roughly similar before we can conclude that any differences that we find are attributable to group differences (not mere variability differences).

• Even in Means Comparisons (ex.t-tests), since Means are highly affected by variability, we need for variability to be similar in our groups so that differences that we find can be attributed to true group differences, and not merely by variability differences between our groups.
More on Homogeneity of Variance

- If distributions are normal in one, then should be for all groups.
More on Homogeneity of Variance

- If distributions in 1 group is leptokurtotic (tall and skinny), then it should be for all other groups.
More on Homogeneity of Variance

- If distributions in 1 group is platykurtotic (short & fat) then it should be for all other groups
More on Homogeneity of Variance

• Any Miss-Match is a Problem
  - Because we might interpret a statistical differences to real group differences, when it’s actually due to heterogeneity of variance

• ...Thankfully SPSS will test this assumption for us (stay tuned)
What about skewed data?

- Positive or negative skews in the data can wreak havoc with statistical analysis
  - Thus always recommend thorough data screening
  - Identify outliers—data entry errors?
  - Consider data transformations if necessary
    - More on this later
Two General Types of ANOVA

• **Independent Measures ANOVA (IM-ANOVA)**
  - Data are collected from separate groups of subjects, and comparisons among groups are desired
    - Student GPA by MAJOR
    - Faculty Salaries by DEPARTMENT

• **Repeated Measures ANOVA (RM-ANOVA)**
  - Data are collected from the same group of subjects on multiple occasions/times, and comparisons of occasions are desired.
    - Longitudinal Studies
    - Student Opinions as Fresh, Soph, Jr, Sr
    - Alumni Donations after 1, 3, 5, 7 years post-graduation
IM & RM Designs…

Repeated Measures Designs

Independent Groups Designs
One-Way IM-ANOVA

- For comparing two or more populations
  - Where sample data have been collected
ANOVA: What’s in a Name?

Population 1
\( \mu = \text{?} \)
Sample 1
\( 0 \quad 2 \quad 4 \quad \bar{X} = 2 \)

Population 2
\( \mu = \text{?} \)
Sample 2
\( 1 \quad 4 \quad 7 \quad \bar{X} = 4 \)

Population 3
\( \mu = \text{?} \)
Sample 3
\( 4 \quad 6 \quad 8 \quad \bar{X} = 6 \)

Total Variability

Between Group Variability
Individual Differences (ID’s)
Error
REAL GROUP DIFFERENCES!

Within Group Variability
Individual Differences
Error
Analysis of Variance F-Ratio

• ANOVA is truly an analysis of a measure of variability, called “variance”
  – Within-Groups Variability
  – Between-Groups Variability

• We Evaluate an “F-Ratio” Representing the Ratio of B/T over W/I:

\[
F = \frac{\text{variability between groups}}{\text{variability within groups}} = \frac{\text{ID's + error + group differences}}{\text{ID's + error}}
\]
Recall your Simple Algebra…

- If the same quantity exists in the Numerator and Denominator of a fraction, they “cancel each other out”

The F-Ratio

\[ F = \frac{\text{variability between groups}}{\text{variability within groups}} = \frac{\text{ID's + error + group differences}}{\text{ID's + error}} \]
Recall your Simple Algebra…

- If the same quantity exists in the Numerator and Denominator of a fraction, they “cancel each other out”
- Leaving us with a number (F) that represents Group Differences!

**The F-Ratio**

\[
F = \frac{\text{variability between groups}}{\text{variability within groups}} = \frac{\text{ID's + error + group differences}}{\text{ID's + error}}
\]
Analysis of Variance F-Ratio

• If F=1...
• As F increases...
• How do you know if F is “big enough” to considered significant?
  – How do you know a t-test is significant??

The F-Ratio

\[ F = \frac{\text{variability between groups}}{\text{variability within groups}} = \frac{\text{ID's + error + group differences}}{\text{ID's + error}} \]
Confidence Intervals with the F-test

• CI’s for comparing two groups are straightforward and intuitive
• CI’s for “Omnibus” differences are less so
  – Effect size calculations exist, but non-intuitive to statistically naieve
• Stay tuned for discussions about post-hoc tests, and how they can sometimes help
• Plots will also be very informative
IM-ANOVA Summary Tables

- Purpose is to provide the necessary components of the F-test
  - Variability (SS)
  - Degrees of Freedom (df)
  - Mean Square (MS)
  - F-statistic (F)
  - Probability values associated with F
- Total, Between Groups, Within Groups
IM-ANOVA Summary Tables

• Purpose is to provide the necessary components of the F-test:
  – Variability (SS)
  – Degrees of Freedom (df)
  – Mean Square (MS)
  – F-statistic (F)
  – Probability values associated with F

• Total, Between Groups, Within Groups
IM-ANOVA Summary Tables

- Purpose is to provide the necessary components of the F-test:
  - Variability (SS)
  - Degrees of Freedom (df)
  - Mean Square (MS)
  - F-statistic (F)
  - Probability values associated with F

- Total, Between Groups, Within Groups
IM-ANOVA Summary Tables

- **Purpose** is to provide the necessary components of the F-test
  - Variability (SS)
  - Degrees of Freedom (df)
  - Mean Square (MS)
  - F-statistic (F)
  - Probability values associated with F

- **Total, Between Groups, Within Groups**
IM-ANOVA Summary Tables

- Purpose is to provide the necessary components of the F-test
  - Variability (SS)
  - Degrees of Freedom (df)
  - Mean Square (MS)
  - F-statistic (F)
  - Probability values associated with F

- Total, Between Groups, Within Groups
IM-ANOVA Summary Tables

- Purpose is to provide the necessary components of the F-test
  - Variability (SS)
  - Degrees of Freedom (df)
  - Mean Square (MS)
  - F-statistic (F)
  - Probability values associated with F

- Total, Between Groups, Within Groups
This is what it looks like…

<table>
<thead>
<tr>
<th></th>
<th>df</th>
<th>SS</th>
<th>MS</th>
<th>F</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Between Groups</strong></td>
<td>##</td>
<td>##</td>
<td>###</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Within Groups</strong> (error)</td>
<td>##</td>
<td>##</td>
<td>###</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

(error)
This is where it comes from (Independent Measures Designs)

\[ SS_{total} = \sum_{i=1}^{N} (x_i - \bar{G})^2 = \sum X^2 - \frac{(\sum x)^2}{N} \]

\[ SS_{between} = \sum_{k=1}^{k} n_k (\bar{X}_k - \bar{G})^2 \]

\[ SS_{within} = \sum SS_{inside each group} \]

\[ df_{total} = N - 1 \]

\[ df_{between} = k - 1 \]

\[ df_{within} = N - k \]
This is where it comes from (Independent Measures Designs)

\[ MS_{total} = \frac{SS_{total}}{df_{total}} \]

\[ MS_{between} = \frac{SS_{between}}{df_{between}} \]

\[ MS_{within} = \frac{SS_{within}}{df_{within}} \]

\[ F = \frac{MS_{between}}{MS_{within}} \]

F-tables provide a p value for a given F-statistic, using df\text{between} (numerator) and df\text{within} (denominator).
Example 1

- Compare Faculty Ratings Across 3 Departments
  - History
  - Psychology
  - Math

- Simplest of ANOVA Models, with ONE Independent Factor (department)
The Data:
One-Way Point-n-Click:

Syntax method:
ONEWAY
  eval BY dept
/STATISTICS DESCRIPTIVES HOMOGENEITY
/PLOT MEANS
/MISSING ANALYSIS
/POSTHOC = TUKEY BONFERRONI SIDAK GH ALPHA(.05).
One-Way ANOVA Output

Oneway--IR Example #1: Oneway Independant-Measures ANOVA

Descriptives

<table>
<thead>
<tr>
<th>Student eval of Intro course</th>
<th>N</th>
<th>Mean</th>
<th>Std. Deviation</th>
<th>Std. Error</th>
<th>95% Confidence Interval for Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>Psych 101</td>
<td>263</td>
<td>36</td>
<td>6.65</td>
<td>1.17</td>
<td>35.28 - 36.42</td>
</tr>
</tbody>
</table>

NOTE: The bold, red font above indicates SPSS' 'reductive' feature. Normally, it reduces the Notes output, but it's handy to open it back up if you forget the exact syntax that was used to execute a statistical procedure. Even if you don't always run stats from syntax, it's nice to see that you could duplicate an earlier analysis perfectly by copying the syntax from the Notes output. NOTE: Also that the file location and name are shown here too.
Two-Factor ANOVAS

• What if you want to compare 2+ groups on MORE THAN one factor?
  – Effect of students’ gender *and* race/ethnicity on performance?
  – Effect of students’ major *and* high school on cGPA?
  – Effect of faculty members’ level *and* age on job satisfaction ratings
Two-Factor ANOVA Effects

• Main Effects
  – One per factor...an F-statistic evaluating the impact of each factor in the model
    • Gender effect on performance (M/F diffs?)
    • Race/ethnicity effect on performance

• Interaction Effects
  – One per interaction...an F-statistic evaluating how two (or more) factors interact with one another to affect the outcome
    • Gender “by” Race/Ethnicity interactive effects on performance
    • More complex...often more interesting!
Example 2

- Compare first term GPA by Major and Citizenship
  - U.S. versus non-U.S.
  - Three Majors—Math, Business, US History

- 2 x 3 ANOVA
The Data
Repeated Measures ANOVA

- Only one sample
- Differences based on time, or condition
- Using the SAME subjects time after time
Repeated Measures ANOVA

- Same people... no "individual differences" in the F-ratio
- More powerful statistics

**The F-Ratio**

\[
F = \frac{\text{variability between conditions/times}}{\text{variability within the sample}} = \frac{\text{error + condition/time differences}}{\text{error}}
\]
Repeated Measures ANOVA

- Same people... no "individual differences" in the F-ratio
- More powerful statistics

**The F-Ratio**

\[ F = \frac{\text{variability between conditions/times}}{\text{variability within the sample}} = \frac{\text{error + condition/time differences}}{\text{error}} \]
RM-ANOVA Summary Tables

- Same Concept as IM Table, but now
  - Instead of “Between Groups” effects, we have “Between Treatments” effects
  - And also “Within Treatments”
    - Consist of subject differences (b/t subjects)
    - And error
- One Group measured several times, thus we partition “within group” variability into that which is due to individual differences, and error.
This is where it comes from
(Repeated Measures Designs)

\[ SS_{total} = \text{Same as IM Anova} \]

\[ SS_{between} = \text{Same as IM Anova} \]

\[ SS_{within} = \text{Same as IM Anova} \]

\[ SS_{b/t \ subjects} = \sum \frac{(\text{each person's total across treatments})^2}{k} - \frac{(\sum X)^2}{N} \]

\[ SS_{error} = SS_{within} - SS_{b/t \ subjects} \]

\[ df_{total} = N - 1 \]

\[ df_{between} = k - 1 \]

\[ df_{within} = N - k \]

\[ df_{b/t \ subjects} = n - 1 \]

\[ df_{error} = (N - k) - (n - 1) \]
This is where it comes from
(Repeated Measures Designs)

\[
MS_{between} = \frac{SS_{between}}{df_{between}}
\]

\[
MS_{error} = \frac{SS_{error}}{df_{error}}
\]

\[
F = \frac{MS_{between}}{MS_{error}}
\]

F-tables provide a p value for a given F-statistic, using df\text{between} (numerator) and df\text{error} (denominator).
Example 3

- Compare student satisfaction ratings over time (four time points)
  - Freshman
  - Sophomore
  - Junior
  - Senior
- Same students...different times
Using Covariates in ANOVA

- Sometimes the apparent effects of one factor, can be “explained” by the effects of some other factor—called a covariate
- There is a significant relationship between panty hose wearing behavior and a form of cancer...any ideas what kind of cancer?
Using Covariates in ANOVA

• Let’s look at the Faculty Salaries from a fictitious Biology Department:
  – Comparing Salaries by Gender and Tenure Status
    • A simple 2-factor ANOVA
  – Then considering AGE as a possible covariate in our model
    • In-other-words, covary-out the effect of age and see if the salary differences remain...
Example 4

• Describe effects of sex and tenure on faculty salaries for Biology faculty
  – Sex (male, female)
  – Tenure (tenured, untenured)
  – Use Age as a covariate

• 2 (sex) x 2 (tenure) Model
  – With a covariate
Related Advanced Topics

• ANOVA can handle multiple factors
  – More than most humans can understand!
• Even just three factors can produce SEVEN effects!
  – 3-way interaction
  – A*B 2-way interaction
  – A*C 2-way interaction
  – B*C 2-way interaction
  – A main effect
  – B main effect
  – C main effect
    • Care to interpret that?
• Four factors = 15 effects
• Five factors = 31 effects!!
• $2^n-1$

Three-Factor Example in Monograph:
Salary by Sex, Tenure, & Department
Related Advanced Topics

• Mixed-Model ANOVAs
• It is possible to consider both types of factors in a single model
  – Student satisfaction over time and by major
  – Student performance over time and by teaching modality

Mixed-Model Quick Demo: Intellectual Growth by College over time
Related Advanced Topics

• Simple Non-Parametrics
  – Chi-Square (best for 2x2’s; purely categorical outcomes)
  – Mann-Whitney U (good for 2 groups; rank or ordinal data)
  – Kruskal-Wallis H (if >2 groups, like One-way ANOVA)

• Loglinear ANOVA
  – Useful for more complicated multi-factor designs
    • Recommend additional training & reading
      – Allan Agresti’s Text is a good one
New Advances

• Hierarchical Modeling (a.k.a. HLM, MLM)
  – Maximum Likelihood based, for using random (i.e. not fixed) factors
  – Assessing impact of “layers” of grouping factors
    • Observations within person
    • Person within group
    • Group within larger group
  – Much better at accommodating for missing data
  – Many variations for different distributions
– Unfortunately, SPSS is not well-suited
Day 2: The saga continues…
Shift Gears: Predicting “Y” from “X” (or several “Xs”)

- Z-tests, T-tests, ANOVA
  - All for *comparing* groups, or observations over time

- Now we’ll shift gears and talk about Regression Analysis
Data Relationships with Two Variables

• Getting a Visual on relationships
  – Quantify relationships (Pearson’s r)
    • Strength
    • Significance
  – Background leading into Simple Regression
Are Two Variables Related?

- SES and GPA?
- High School GPA and First Term College GPA?
- SAT Scores and GPA?
- ACT Scores and GPA?
- SAT and ACT Scores?

- We’re looking at PAIRS of variables
- And we’re assessing relationships (non-causal)
NOTE THAT

• We are not comparing means from 2 different groups here!
• We are trying to see if there is a relationship among two different variables
  – Usually continuous in scale
We could look at a table of numbers?

Any relationship between X and Y here?
Scatterplots help us visualize...

- Two continuous variables, X and Y
- Scatterplot shows association of X and Y
Recall basic Correlational Analysis

- Two continuous variables, X and Y
- Scatterplot shows association of X and Y
- Dots represent observations (usually people)
Quantifying the Relationship: Recall Linear Correlation

- Measures the direction and strength of the linear relationship between two variables.

\[ r = \frac{1}{n-1} \sum \left( \frac{x_i - \bar{x}}{s_x} \right) \left( \frac{y_i - \bar{y}}{s_y} \right) \]

- Sign tells you what?
- R-value tells you what?
- P-value tells you what?
- Causality?
Scatter Plots... Footballs, Basketballs, Directionality
Correlation Analysis Quantifies our Understanding

- Group 1: 
  - \( r = 0.90 \)
  - \( p = 0.001 \)

- Group 2: 
  - \( r = 0.90 \)
  - \( p = 0.001 \)

- Group 3: 
  - \( r = 0.00 \)
  - \( p = 0.998 \)

- Group 4: 
  - \( r = 0.40 \)
  - \( p = 0.057 \)
Ordinary Least Squares Regression

- Using *Simple Regression* to Describe a Linear Relationship
- Regression as a Descriptive Tool
- Regression as a Forecasting/Prediction Tool
- Making Inferences from Simple Regression
Simple Linear Regression

• Statistic used to describe linear relationships among variables:

\[ \hat{y} = b_0 + b_1 x \]

\( \hat{y} \) = the dependant (outcome) variable

\( x \) = the independent (predictor, explanatory) variable

\( b_0 \) = the y-intercept of graph of all (x, y) pairs

\( b_1 \) = the slope of the line

Just like HS Algebra:

\[ Y = m(X) + b \]
Consider a simple example of a Height/Age Scatterplot...

- Each green dot represents a person in the sample.
The OLS Regression Equation

- A “best-fitting” line is calculated from the sample data.
- Can be used to predict $Y$, given $X$

$$\hat{y}_{47} = a + 24x_{47}$$

Error $= y - \hat{y}$
OLS Regression Lines Aren’t Perfect!

- Some Predictions are too high:

\[ \hat{y}_{47} = a + 24x_{47} \]

Error = \( y - \hat{y} \)
OLS Regression Lines Aren’t Perfect!

- Some Predictions are too low:

\[ \hat{y}_{47} = a + 24x_{47} \]

Error = \( y - \hat{y} \)
OLS Regression Lines Aren’t Perfect!

- Some Predictions are too high:

\[ \hat{y}_{47} = a + 24x_{47} \]

Error = \( y - \hat{y} \)
Deriving the "Best-Fitting" Line

• Getting the best line is KEY!
• One Approach would be a line that Minimizes the Sum of the Errors:

\[ \sum_{i=1}^{n} (y_i - \hat{y}_i) \]
Why not minimize \( \sum_{i=1}^{n} (y_i - \hat{y}_i) \)?

- Because positive errors cancel out negative errors when minimizing sum of errors...

\[
\sum_{i=1}^{n} (y_i - \hat{y}_i) = 0
\]

- We need a way to consider the sign of the error into the minimization function.
...And Also...

- **All** lines minimizing Sum of Errors pass through the point \((\bar{x}, \bar{y})\).

\[
\sum_{i=1}^{n} (y_i - \hat{y}_i) = 0
\]

- ...Infinite number of lines meet this criteria, but only 1 is really “best”
Another Approach?

• Minimize the Absolute Value of Errors?

\[ \sum_{i=1}^{n} |y_i - \hat{y}_i| \]

Least Absolute Value (LAV)

- No unique LAV line, plus very complicated calculations
Third Approach?

- Minimize the Squared errors?

$$\sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

Least Squares Regression (LS)
Is LS Minimization Possible?

• Yes!

For a Linear Equation $\hat{y}_i = b_0 + b_1 x$:

$$b_1 (\text{slope}) = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} = \frac{\sum_{i=1}^{n} x_i y_i - \frac{1}{n} \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} x_i^2 - \frac{1}{n} \left( \sum_{i=1}^{n} x_i \right)^2}$$

$$b_0 (\text{intercept}) = \bar{y} - b_1 \bar{x}$$
…a NOTE about formulae

- Hopefully you’ll be able to recognize what they’re doing.
- But don’t worry about memorizing them!

- SPSS (or whatever your choice of software) will calculate the components of the OLS Regression Line for you anyway!
Minimizing the SQUARED Errors

- Takes care of the problem of positive errors “canceling out” negative ones
- Computationally simple enough for hand or computer calculations
- Creates a unique “line of best fit”
Let’s do one by hand…

<table>
<thead>
<tr>
<th>i</th>
<th>$x_i$</th>
<th>$y_i$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>8</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
<td>8</td>
</tr>
<tr>
<td>5</td>
<td>5</td>
<td>11</td>
</tr>
<tr>
<td>6</td>
<td>6</td>
<td>13</td>
</tr>
</tbody>
</table>
Let’s do one by hand…

<table>
<thead>
<tr>
<th>i</th>
<th>$x_i$</th>
<th>$y_i$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>8</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
<td>8</td>
</tr>
<tr>
<td>5</td>
<td>5</td>
<td>11</td>
</tr>
<tr>
<td>6</td>
<td>6</td>
<td>13</td>
</tr>
</tbody>
</table>

Start with formula for $b_i$
Let’s do one by hand...

<table>
<thead>
<tr>
<th>i</th>
<th>$x_i$</th>
<th>$y_i$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>8</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
<td>8</td>
</tr>
<tr>
<td>5</td>
<td>5</td>
<td>11</td>
</tr>
<tr>
<td>6</td>
<td>6</td>
<td>13</td>
</tr>
<tr>
<td>sum</td>
<td>21</td>
<td>45</td>
</tr>
</tbody>
</table>

$$b_i = \frac{\sum_{i=1}^{n} x_i y_i - \frac{1}{n} \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} x_i^2 - \frac{1}{n} \left( \sum_{i=1}^{n} x_i \right)^2}$$

The $b_i$ formula requires $E(x)$ and $E(y)$
Let’s do one by hand…

The \( b_i \) formula requires these two columns also…

<table>
<thead>
<tr>
<th>( i )</th>
<th>( x_i )</th>
<th>( y_i )</th>
<th>( x_i y_i )</th>
<th>( x^2 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>3</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>2</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>8</td>
<td>24</td>
<td>9</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
<td>8</td>
<td>32</td>
<td>16</td>
</tr>
<tr>
<td>5</td>
<td>5</td>
<td>11</td>
<td>55</td>
<td>25</td>
</tr>
<tr>
<td>6</td>
<td>6</td>
<td>13</td>
<td>78</td>
<td>36</td>
</tr>
<tr>
<td>sum</td>
<td>21</td>
<td>45</td>
<td>196</td>
<td>91</td>
</tr>
</tbody>
</table>
Let's do one by hand...

<table>
<thead>
<tr>
<th>i</th>
<th>$x_i$</th>
<th>$y_i$</th>
<th>$x_iy_i$</th>
<th>$x^2$</th>
<th>$b_i$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>3</td>
<td>3</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>2</td>
<td>4</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>8</td>
<td>24</td>
<td>9</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>4</td>
<td>8</td>
<td>32</td>
<td>16</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>5</td>
<td>11</td>
<td>55</td>
<td>25</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>6</td>
<td>13</td>
<td>78</td>
<td>36</td>
<td></td>
</tr>
<tr>
<td>sum</td>
<td>21</td>
<td>45</td>
<td>126</td>
<td></td>
<td>$b_0 = \bar{y} - b_i \bar{x}$</td>
</tr>
</tbody>
</table>

...and $b_0$ requires means of $x$ and $y$
Using the formula...

For a Linear Equation \( \hat{y}_i = b_0 + b_1 x \):

\[
b_1 = \frac{\sum_{i=1}^{n} x_i y_i - \frac{1}{n} \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} x_i^2 - \frac{1}{n} \left( \sum_{i=1}^{n} x_i \right)^2} = \frac{196 - \frac{1}{6} (21)(45)}{91 - \frac{1}{6} (21)^2} = \frac{38.5}{17.5} = 2.2
\]

\[
b_0 = \bar{y} - b_1 \bar{x} = 7.5 - 2.2(3.5) = -0.2
\]
Resulting in one, Unique LS Regression Line:

\[ y = -0.2 + 2.2x \]

- What does this line do for us?
  - Given \( x = 15 \) .... Predict \( y \).

  \[ \hat{y} = 32.8 \]

- What can you say about the errors that you will make in your prediction?

  We’ve minimized our error in prediction.
Using LS Regression to Describe Relationships

- Can be used to establish the weight assigned to various factors that may (or may not) predict some outcome
  - Ex. Assess the value of knowing an incoming students’ SAT Verbal on their first-term GPA, based on a sample of historical student application/enrollment data

\[
GPA_{\text{first term}} = 1.2 + .312(SAT_{\text{verbal}})
\]
Using LS Regression to Describe Relationships

- Can be used to establish the weight assigned to various factors that may (or may not) predict some outcome.
  - Ex. Assess the value of knowing an incoming students’ SAT Verbal on their first-term GPA, based on a sample of historical student application/enrollment data.

$$GPA_{firstterm} = 1.2 + 0.312(SAT_{verbal})$$
Let’s run a Simple Regression using SPSS

Description of Dataset NH SAT A630

The dataset contains annual information on the set of New Hampshire students who have taken the Scholastic Aptitude Test (SAT) each year from 1976 through 1998. The variables in the dataset are defined as follows:

- **YEAR**
- **TOTAL** = Total number of SAT takers in NH
- **SATV** = Average SAT-Verbal score
- **SATM** = Average SAT-Math score
- **PCTDOCT** = Percentage of SAT takers planning on pursuing a doctorate degree
- **UNH** = Number of SAT takers sending test scores to the U New Hampshire
- **CPI** = Consumer Price Index (1983 = 100)
- **UNHPCT** = % of total SAT takers sending test scores to U New Hampshire
- **LUNHPCT** = Natural logarithm of UNHPCT (=\(\ln(UNHPCT)\))
- **Resident** = Resident tuition rate at U New Hampshire (in $1000s)
- **Nonres** = Nonresident tuition rate at U New Hampshire (in $1000s)
- **Private** = Average private tuition rate in New England (in $1000s)
- **Income** = Median family income (in $1000s)
- **Lres** = Natural logarithm of Resident (=\(\ln(Resident)\))
- **Lprivate** = Natural logarithm of Private (=\(\ln(Private)\))
- **Lincome** = Natural logarithm of Income (=\(\ln(Income)\))
- **Lnonres** = Natural logarithm of Nonres (=\(\ln(Nonres)\))
Let’s run a Simple Regression using SPSS

- Let’s predict the number of SAT scores we might expect to be sent to UNH, so our Enrollment Management Office can do some strategic planning.
  - \( Y = \text{UNH} \)

- Let’s use only 1 predictor—the total number of SAT test-takers
  - \( X_1 = \text{TOTAL} \)
    - Ok.. This is a little boring, but bare with me!
SPSS Demo

- Simple Regression—used to describe the relationship between total number of SAT test takers in NH, and the number of SAT scores sent to NH Admissions.
Moving Beyond Describing Relationships...

- Can we use Regression to “go beyond” what sample data suggest?
- Can we make *inferences* about how two variables *in the population* might be related based on observations of *sample* data?
Why would we want to do this?

- Why bother when most IR offices have access to your local “population” of students (ex. All current incoming freshmen)?
  - i.e. you don’t randomly sample your student data warehouse… you tap all data you have?!
- Because you want to make inferences about the whole incoming class based on “today’s” data
- Because you want to make inferences about next semester, next year, etc…
- Because you want to do some strategic planning that could benefit by this kind of analysis
Statistical Inferences require Assumptions about the Population

• Given Population with X and Y

\[ \mu_{y|x} = \text{conditional mean of } y, \text{ given } x, \]
\[ \mu_{y|x} = \beta_0 + \beta_1 x \text{ where } \]

\[ \beta_0 = \text{population y - intercept, and} \]
\[ \beta_1 = \text{slope of the population regression line.} \]
We assume Population X/Y Relationship is Linear
We assume Population X/Y Relationship is Linear.
RE: The Population Conditional Mean...

\[ \mu_{y|x} = \text{conditional mean of } y, \text{ given } x : \]

\[ \mu_{y|x} = 40 \]

\[ \mu_{y|x} = 30 \]

\[ \mu_{y|x} = 20 \]

For given x, y varies, with a mean and variance
RE: The Population Conditional Mean...

\[ \mu_{y|x} = \text{conditional mean of } y, \text{ given } x : \]

For given x, y varies, with a mean and variance ...leading to Prediction Error...or Disturbances
$\mu_{y|x} = \text{conditional mean of } y, \text{ given } x$:

For given $x$, $y$ varies, with a mean and variance ...leading to Prediction Error...or Disturbances...adding a term to our regression equation:
Resultant (Inferential) Regression Equation

\[ y_i = \beta_0 + \beta_1 x_i + e_i \]

- Where \( e_i \) represents the difference between TRUE \( y \) and the conditional mean of all \( y|x \)
  - Disturbances, or Error Variance
  - If our equation were perfect, \( e_i = 0 \)
- What type of relationship would have to occur for this to happen?
Leading up to some ASSUMPTIONS for Inferential Regression

1. Expected Value of disturbances $= 0$ ($E(e_i)=0$)
   a) i.e. The population regression is linear
2. Homogeneity of Variance for $e_i$
3. $e_i$ are normally distributed
4. $e_i$ are independent
Inferences About $\beta_0$ and $\beta_1$

- $b_0$ and $b_1$ are point estimates of $\beta_0$ and $\beta_1$
  - I.e. they are calculated from samples, thus they are *statistics*
  - Thus they are random variables with probability distributions (sampling distributions)
- Just as sample means are unbiased estimates as population means...
  - $b_0$ and $b_1$ are unbiased estimates of $\beta_0$ and $\beta_1$
  - We assume that their sampling distributions are normally distributed
  - …and as $n$ increases, $b_0$ and $b_1$ become closer and closer to $\beta_0$ and $\beta_1$
  - Of all possible estimators of $\beta_1$, $b_1$ has the desirable feature of having smaller sampling errors than any other unbiased estimator.
In other words...

- $b_0$ and $b_1$ are *unbiased* Estimators
  - Mean of sampling distribution = Pop mean
- $b_0$ and $b_1$ are *Consistent* Estimators
  - As $n$ increases, estimator approaches parameter
- $b_0$ and $b_1$ are a *Minimum Variance* Estimators
  - While there are other unbiased estimators “out there,” $b_0$ and $b_1$ have the smallest variance
GIVEN all of these assumptions

• We can make hypotheses about $\beta_0$ and $\beta_1$, and use $b_0$ and $b_1$ to test our hypotheses
  – like we can about means or other statistics.

• But we’re missing one remaining component of the Regression Equation.
  – Variance around the regression line, or Error Variance
  – We know our predictions aren’t perfect, but we need to quantify the “error in prediction”
Estimating Variance around the Regression Line

- The estimate of Variance @ Regression:
  \[ \sigma_e^2 = s_e^2 = \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{n-2} = \frac{SSE}{n-2} = MSE \]

- SSE = sum of squared errors, adjusted by n-2 df to incorporate sample size.
  - Df=(sample size – num of coefficients)
    - We’re estimating two coefficients, b₀ and b₁

- MSE=Mean Square Error = any SS/df

- Square Root of \( s_e^2 \) is termed “Stand Error of Regression”
Back to our SPSS Example…

• Note the SSR, SSE and MS terms in the ANOVA Summary Table

• Use ANOVA Table to Evaluate Significance of the Regression Equation
  – More important later, when we discuss *multiple* regression

• Interpret $R^2$
Hypothesis testing in Regression

- We know that the LS estimates of $b_0$, & $b_1$ are unbiased estimators of the Population coefficients $\beta_k$
- We can calculate Confidence Interval Estimates (& significance values) of $\beta_k$
- And we test hypotheses that the population beta weights ($\beta_k$) are significantly greater than some hypothesized value (usually 0)
  - Reject the notion that $\beta_k =$ hypothesized value if observed t-score probability $< \alpha$
The Theory Behind It?

- Create Null and Alternative Hypotheses:
  \[ H_0 : \beta_k = \beta_k^* \text{ (usually 0)} \]
  \[ H_a : \beta_k \neq \beta_k^* \]
- Calculate t-score(s) for coefficients
  \[ t = \frac{b_k - \beta_k^*}{s_{b_k}} = \frac{b_k}{s_{b_k}} \text{ (when } \beta_k^* = 0) \]
- Compare t-value to critical t (given alpha)
  Reject \( H_0 \) if \( t > t_{\alpha/2} \) or \( t < t_{\alpha/2} \)
  Accept \( H_0 \) if \( -t_{\alpha/2} \leq t \leq t_{\alpha/2} \)
A Closer look at t-formula

\[ H_0 : \beta_k = \beta_k^* \text{ (usually 0)} \]

\[ H_a : \beta_k \neq \beta_k^* \]

- Calculate t-score(s) for coefficients

\[ t = \frac{b_k - \beta_k}{S_{b_k}} = \frac{b_k}{S_{b_k}} \text{ (when } \beta_k^* = 0) \]

When null is true, t-should be (large or small)? Why??
When Alternative is true, t-should be (large or small)?
A Closer look at t-formula

- Calculate t-score(s) for coefficients

\[ t = \frac{b_k - \beta_k^*}{s_{b_k}} = \frac{b_k}{s_{b_k}} \text{ (when } \beta_k^* = 0) \]

**NOTE** that we are NOT comparing two sample means, even though we’re using a t-test! We are comparing our observed \( b_k \) to a constant that we choose (\( \beta_k^* \), which us usually 0).
The “Typical” Situation

• Testing Hypothesis that the coefficients are significantly higher or lower than 0

\[
t = \frac{b_k - \beta_k^*}{s_{bk}} = \frac{b_k - 0}{s_{bk}} = \frac{b_k}{s_{bk}}
\]

• If \( b_1 \) coefficient = 0, what would that mean in terms of the Regression Equation?

\[
\hat{y} = b_0 + b_1 x
\]
If Null is Accepted

- Small $t$-statistic, $p$-value is not $< \alpha$.
  - “$X$ does not appear to be linearly related to $Y$”
- i.e. $x$ doesn’t help you predict $y$
If Null is Rejected

\[ \begin{align*}
H_0 : & \beta_k = \beta_k^* \quad (\text{usually 0}) \\
H_a : & \beta_k \neq \beta_k^* 
\end{align*} \]

- Large t-statistic, p-value is < alpha.
  - “There is evidence that \( y \) and \( x_k \) are linearly related, and that \( x_k \) helps explain some of the variation in \( y \) (not accounted for by the other explanatory variables).”
  - I.e. \( x \) is helpful in predicting \( y \)
  - Parentheses used in multiple regression… coming soon!
How will we do this?

• SPSS will calculate:
  – Some diagnostics that help us evaluate our assumptions about the data
  – Estimates of the coefficients
  – SD for the estimates of coefficients
  – T-scores comparing estimate to 0
  – p-values associated with the T-test

• Humans will
  – Interpret the output in plain English!
  – Then explain it to constituents so that they can understand it too!
Back to our SPSS Example

• Find the Table of Coefficients
  – T-values, p-values
  – Can you construct the linear equation?

• Evaluate when our model is least & most accurate
  – Plot actual vs. predicted values
  – Calculate r, r^2 & Interpret
Multiple Regression Analysis
Simple vs. Multiple Regression

- “Simple” refers to predicting a single outcome (y) from a single predictor (x)

- “Multiple” refers to predicting a single outcome (y) from two or more predictors (x1, x2, x3)

  - Still assuming a linear relationship
  
  - But there are ways to “coax” linearity if it’s not already there...
Multiple Regression Examples

- Ex. Predict *Faculty Salary* from Age, Department, Years as Faculty Member and Gender
- Ex. Predict *Student Performance on GE Outcomes* from Cumulative GPA, College Major, and Gender
Multiple Regression

- We’re still talking about Linear relationships
  \[ \hat{y} = b_0 + b_1 x_1 + b_2 x_2 \ldots + b_k x_k \]
- Still using method of Least Squares to develop the equation
- Still estimating regression coefficients (betas)
Multiple Regression

• But Graphing the equation results in a plane, or more complex geometric shape, not a line, even though the relationships are still linear...
  – 3-D graphing?
  – …or 2-D graphing?
Multiple Regression Plane
(example w/2 predictors)

Predicting Sales from Price and Advertising

Linear Regression with
95.00% Mean Prediction Interval

Sales = 8.31 + -0.09 * price + 0.38 * advertising

R-Square = 0.45
Assumptions for the Population Multiple Regression

1. Expected value of disturbances is zero: \( E(e_i) = 0 \)

2. Variance of each \( e_i \) is equal to \( \sigma^2_e \)
   
   (i.e. each disturbance along the regression line has equal variance regardless of value of \( x \))

3. The \( e_i \) are normally distributed.

4. The \( e_i \) are independent. (Be careful RE longitudinal data... usually not independent.)
Assumptions for the Population Multiple Regression

1. Expected value of disturbances is zero: $E(e_i) = 0$

2. Variance of each $e_i$ is equal to $\sigma_e^2$
   (i.e. each disturbance along the regression line has equal variance regardless of value of $x$)

3. The $e_i$ are normally distributed.

4. The $e_i$ are independent. (Be careful RE locality in data... usually not independant.)

5. Predictors themselves are independant
Assessing the FIT of a Multiple Regression Line

• In Simple Regression, we mainly focused on discussing the significance of the regression coefficients.
• In Multiple Regression, we must also pay attention to the overall Regression Equation?
  – Is it any good at predicting?
  – How do we know?
Assessing the FIT of a Multiple Regression Line

• With Simple Regression, we didn’t pay much attention to this.
  – If the coefficient was significant, that implied that the equation itself was too.

• With Multiple Regression, we must first evaluate the overall equation before diving deeper.
  – Then determine which, if any coefficients are significant.
Recall Hypothesis Testing with Simple Regression…

\[ \hat{y} = b_0 + b_1 x \]

\[ H_0 : \beta_k = \beta_k^* \text{ (usually 0)} \]
\[ H_a : \beta_k \neq \beta_k^* \]

\[ t = \frac{b_k - \beta_k^*}{s_{b_k}} = \frac{b_k}{s_{b_k}} \text{ (when } \beta_k^* = 0) \]

Reject \( H_0 \) if \( t > t_{\alpha/2} \) or \( t < t_{\alpha/2} \)
Accept \( H_0 \) if \( -t_{\alpha/2} \leq t \leq t_{\alpha/2} \)
Hypothesis Testing with Multiple-Regression

\[ \hat{y} = b_0 + b_1 x_1 + b_2 x_2 ... b_k x_k \]

- Null Hypothesis: \( H_0 : \beta_1 = \beta_2 = ... = \beta_k = 0 \)
- Alternative Hypothesis: \( H_a : \) At least one coefficient is not equal to zero

<table>
<thead>
<tr>
<th>If reject</th>
<th>Reject ( H_0 ) if ( F &gt; F(\alpha; K, n - K - 1) )</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Accept ( H_0 ) if ( F \leq F(\alpha; K, n - K - 1) )</td>
</tr>
</tbody>
</table>

- Test Statistic: \( t = \frac{b_k}{s_{b_k}} \)
- Reject \( H_0 \) if \( t > t_{\alpha/2} \) or \( t < -t_{\alpha/2} \)
- Accept \( H_0 \) if \( -t_{\alpha/2} \leq t \leq t_{\alpha/2} \)
The ANOVA Summary Table

- Evaluation of the overall “Fit” of our Regression Equation
- Do all coefficients = 0 (null hyp) or is at least one of them ≠ 0 (alt hyp).
- The results of the ANOVA are found in an ANOVA Summary Table...

Reject $H_0$ if $F > F(\alpha; K, n - K - 1)$
Accept $H_0$ if $F \leq F(\alpha; K, n - K - 1)$
The ANOVA Summary Table

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>SS</th>
<th>MS</th>
<th>F</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression</td>
<td>K</td>
<td>SSR</td>
<td>SSR/K</td>
<td>MSR/MSE</td>
<td>p-value</td>
</tr>
<tr>
<td>Residual Error</td>
<td>n-k-1</td>
<td>SSE</td>
<td>SSE/(n-K-1)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>n-1</td>
<td>SST</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Sums of Squares
The F-Ratio (MSR/MSE), and the associated p-value, tell us whether or not our regression equation is predicting a “significant” amount of the variance in y from knowledge of $x_1, x_2, \ldots x_k$.

- If $p$-value < 0.05 (traditionally), equation is said to be “significant”
The $R^2$ Term
(Generated from SS Terms)

- The variation in y: $SST = SSE + SSR$
  - "Total SS=Regression SS+ Error SS"
- $R^2 = \frac{SSR}{SST}$
  - $R^2$ is the ratio of explained-to-total variance (SS) is an evaluation of the overall regression.
  - I.e. "percent of variance accounted for"
ANOVA p-value VS. Multiple-\(R^2\)

- ANOVA p-value tells us whether we can account for a **significant** proportion of variance in \(Y\), by knowledge of all of the predictors \((X_1, X_2 \ldots X_k)\).
  - \(F = \frac{MSR}{MSE}\) ... associated with a p-value

- Multiple-\(R^2\) tells is an estimate of **how much** variance we can account for in \(Y\) by knowledge of all of the predictors \((X_1, X_2 \ldots X_k)\).
  - \(R^2 = \frac{SSR}{SST}\)
The Multiple R-Square

• The “Multiple-$R^2$” value is very similar to the correlation coefficient.

\[
R^2 = \frac{SSR}{SST} = \left( 1 - \frac{SSE}{SST} \right)
\]

• But in multiple-regression it has a flaw…
  – It doesn’t decrease as new predictors are added, even if they are “useless” additions.
The Adjusted Multiple-R-Square

- We need to “adjust” the $R^2$ value to correct for the addition of more predictors.

\[ R^2 = 1 - \frac{SSE}{SST} \]

\[ R_{adj}^2 = 1 - \frac{SSE/(n-K-1)}{SST/(n-1)} \]

- Note how the SS in numerator and denominator are adjusted for their df?
The Adjusted Multiple-R-Square

- This “adjustment” results in an adjusted R-square value that compensates for the number of predictors in the model.
  - No longer represents “the percent of variance accounted for”
  - But CAN BE used to compare different multiple-regression models
    - More on this later…
Let’s Give it a Try, eh?

• SPSS Example of Multiple Regression
  – One Outcome Variable (number SAT scores sent to NH)
  – Several Predictors (Total SAT Takers, other predictors)
• What’s the $R^2$? Adjusted $R^2$?
• Is the equation itself any good? (i.e. can it account for sig. prop. of $Y$ variance?)
• Which, if any, of the predictors are useful?
  – Interpret
Steps We’ll Take...

1. F-test for overall fit of regression
2. If F-test is significant, examine the t-tests for each of the coefficients.
3. Report the total percent variation in y explained by the x predictors
4. Examine the Adjusted R2
SPSS Example

- Multiple Regression predicting Number of SAT scores sent to UNH, by
  - Total number of test takers
  - Average SAT Verbal Score
  - Average SAT Math Score
Statistical Variable Selection Techniques
Strategies when theory cannot guide you...

• Thus far, the theory has been our guide on choosing predictors to consider
  – Theory is always the best strategy!!

• Sometimes you may be on a “data mining” mission…

• There are techniques that can help you
  – With Rob’s strong dose of caution!
Statistical Strategies

- **Selection algorithms**: rules for deciding when to drop or add variables
  1. Backwards Elimination
  2. Forward Selection
  3. Stepwise Regression
  4. Run All Possible Models
Words of Caution

- None guarantee you get the right model because they do not check assumptions or search for omitted factors like curvature.
- None have the ability to use a researcher's knowledge about the situation being analyzed.
- Many among the scientific community do not respect statistical selection strategies like these because they are not grounded in theory, and they capitalize on sample variance relationships that may not exist in the population...
Backwards Elimination

• Start with all variables in the equation.
• Examine the variables in the model for significance and identify the least significant one.
• Remove this variable if it does not meet some minimum significance level.
• Run a new regression and repeat until all remaining variables are significant.
Forward Selection

- At each stage, it looks at the $x$ variables not in the current equation and tests to see if they will be significant if they are added. (I.e. a significant partial-F statistics would result.)
- In the first stage, the $x$ with the highest correlation with $y$ is added.
- At later stages it is much harder to see how the next $x$ is selected.
Stepwise Regression

- A limitation with the backwards procedure is that a variable that gets eliminated is never considered again.

- With forward selection, variables entering stay in, even if they lose significance.

- Stepwise regression corrects these flaws. A variable entering can later leave. A variable eliminated can later go back in.
Stepwise...

• Begins like Forward (Chooses best predictor, adds it, tests for sig Partial-F, Keeps if pass criteria)
• Next look at remaining predictors, choose “best” one (Highest Partial-F), and includes it.
• Then behaves like Backwards… Potentially REMOVING one of the variables already included if it’s not necessary.
• Then adds a new one…
• Then tests to remove any of those included…
• Until finished.
Stepwise...

- Ultimately, all variables in the equation are adding significantly, and none of the ones eliminated would (according to criteria we establish ahead of time)
- It is possible to add a variable, remove it later, then add again at a later step!
Example of using Stepwise with SPSS

- Consider parameters for adding/removing variables
  - "Alpha to Remove"
    - maximum p-value a variable can have and stay in the equation
  - "Alpha to Enter"
    - minimum p-value a variable needs to enter the equation

- Often we use values like .15 or .20 because this encourages the procedures to look at models with more variables.
SPSS Example

• Predict the number of SAT scores sent to NH
  – Using Stepwise Selection Technique
    • Including all (untransformed) predictors in the data set
All Possible Models?

• If reasonable, this is likely a better solution than Stepwise, but...
  – Some software (SPSS) cannot easily accommodate
  – Can be unreasonable if there are many potential predictors
  – Still not as good as theory
    • ex. what if a non-linear transformation is really the driver?

• Model selection usually based on Adjusted $R^2$ for OLS regression
Related Advanced Topics

- Regression with Bivariate or Multinominal Outcomes
  - Logit Procedure
    - Quite different in output and interpretation
  - Logistic Regression
    - Useful for >2 categories to the outcome variable
- Regression for Ordinal outcomes
  - Poisson, Negative Binomial, Others
- Hierarchical (nested) Regressive models
  - “Block” Regression in SPSS
  - Used to compare models with increasing complexity…
New Advances

- Hierarchical Linear Modeling (HLM)
  - A.k.a. MLM, Mixed Modeling
- Allows for nesting to be considered
- Allows both fixed and random effects
- Allows for time-dependent covariates
- Allows for group-level effects modeling
Time Series Regression

Modeling Known (or easily observable) Patterns in Time-Series Data
Time Series Regression

- Some data are linear in their relationship, but have "cycles" that we'd like to capture also
  - Ex. New Admits over time
    - Where predictable cycles exist among Fall, Spring, Summer terms
Consider this time series plot:
Consider this time series plot:

20 Year’s of Data!
Number of applications rec’d since 1985, where each observation is a semester.
Consider this time series plot:

Beginning with Fall, 1985
Consider this time series plot:

Then Spring, 1986
Consider this time series plot:

Followed by Summer, 1986....

And so on.
We could fit a linear regression to the data... $R^2 = .71$
But if we could “capture” the cycle of FA/SP/SU wouldn’t that be better?

Obvious cycles in the data, with FA/SP/SU repeating the pattern.
How?

- Include “dummy” predictors indicating whether data is from a Fall, Spring, or Summer term
  - Use 2 of the 3, leaving one as “reference”
- Run Multiple Regression:
Comparison of the two models:

### Simple Regression Model

#### Model Summary

<table>
<thead>
<tr>
<th>Model</th>
<th>R</th>
<th>R Square</th>
<th>Adjusted R Square</th>
<th>Std. Error of the Estimate</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>.841</td>
<td>.707</td>
<td>.702</td>
<td>2161.135</td>
</tr>
</tbody>
</table>

- a. Predictors: (Constant), Observation
- b. Dependent Variable: Number of Applications Received

#### ANOVA

<table>
<thead>
<tr>
<th>Model</th>
<th>Sum of Squares</th>
<th>df</th>
<th>Mean Square</th>
<th>F</th>
<th>Sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression</td>
<td>6.8E+008</td>
<td>1</td>
<td>677290217.4</td>
<td>145.014</td>
<td>.000</td>
</tr>
<tr>
<td>Residual</td>
<td>2.8E+008</td>
<td>60</td>
<td>4670502.918</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>9.6E+008</td>
<td>61</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- a. Predictors: (Constant), Observation
- b. Dependent Variable: Number of Applications Received

#### Coefficients

<table>
<thead>
<tr>
<th>Model</th>
<th>Unstandardized Coefficients</th>
<th>Standardized Coefficients</th>
<th>t</th>
<th>Sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 (Constant)</td>
<td>781.925</td>
<td>556.637</td>
<td>1.371</td>
<td>.175</td>
</tr>
<tr>
<td>Observation</td>
<td>184.682</td>
<td>15.337</td>
<td>.841</td>
<td>.000</td>
</tr>
</tbody>
</table>

- a. Dependent Variable: Number of Applications Received

### Time Series Regression with Seasonal Predictors

#### Model Summary

<table>
<thead>
<tr>
<th>Model</th>
<th>R</th>
<th>R Square</th>
<th>Adjusted R Square</th>
<th>Std. Error of the Estimate</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>.964</td>
<td>.930</td>
<td>.926</td>
<td>1075.387</td>
</tr>
</tbody>
</table>

- a. Predictors: (Constant), Spring, Observation, Fall
- b. Dependent Variable: Number of Applications Received

#### ANOVA

<table>
<thead>
<tr>
<th>Model</th>
<th>Sum of Squares</th>
<th>df</th>
<th>Mean Square</th>
<th>F</th>
<th>Sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression</td>
<td>8.9E+008</td>
<td>3</td>
<td>296815296.0</td>
<td>256.659</td>
<td>.000</td>
</tr>
<tr>
<td>Residual</td>
<td>67074505.0</td>
<td>58</td>
<td>1156456.975</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>9.6E+008</td>
<td>61</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- a. Predictors: (Constant), Spring, Observation, Fall
- b. Dependent Variable: Number of Applications Received

#### Coefficients

<table>
<thead>
<tr>
<th>Model</th>
<th>Unstandardized Coefficients</th>
<th>Standardized Coefficients</th>
<th>t</th>
<th>Sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 (Constant)</td>
<td>-1352.942</td>
<td>340.067</td>
<td>-3.978</td>
<td>.000</td>
</tr>
<tr>
<td>Observation</td>
<td>186.214</td>
<td>7.634</td>
<td>.848</td>
<td>24.393</td>
</tr>
<tr>
<td>Fall</td>
<td>4490.993</td>
<td>336.016</td>
<td>.541</td>
<td>13.365</td>
</tr>
<tr>
<td>Spring</td>
<td>1611.276</td>
<td>336.016</td>
<td>.194</td>
<td>4.795</td>
</tr>
</tbody>
</table>

- a. Dependent Variable: Number of Applications Received
Is there anything else we could use in our data???
Is there anything else we could use in our data???

Fall, 1998: New Admit Standards
Modeling Policy Changes…

• Simply add a dummy predictor that captures the policy change!
  
  – …Though it may be tempting, please make no assumptions about the fact that “policy change” and “dummy” are in the same sentence. 😊

<table>
<thead>
<tr>
<th>Model</th>
<th>Variables Entered</th>
<th>Variables Removed</th>
<th>Method</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>New Admission Standards (FA98), Spring, Fall, Observation</td>
<td>.</td>
<td>Enter</td>
</tr>
</tbody>
</table>

a. All requested variables entered.

b. Dependent Variable: Number of Applications Recd
Results?

In this case, the policy change did not significantly impact admits after modeling time and cycles of semesters...
Results?

And the adjusted R2 is similar to the simpler model that did not include the policy change predictors.
Exploratory Factor Analysis
Principal –Axis Exploratory Factor Analysis

- Common in survey research
- Useful for “discovery” of underlying constructs
- Useful as a strategy for condensing data
- Useful as a strategy to approximate continuous data from ordinal data elements
  - Combining several Likert-Scaled items into one construct score that behaves “normally”
What EFA can do?

• Purpose is to discover simple patterns among variables
• If patterns are found, we call them “factors,” or “constructs”—hence the name
• EX: Is intelligence uni-dimensional, or multi-dimensional?
  - Hint: Consider College Board Exams…
    • Verbal, Math, Logic
EFA Graphic (after solution has been rendered)

Item 1
Item 5
Item 13
Item 7

Item 2
Item 3
Item 11
Item 6
Item 9

Item 4
Item 8
Item 10
Item 12

Factor 1
Factor 2
Factor 3
Goals of EFA

- To better understand the underlying constructs.
  - Thus inferences are made about the constructs, not individual items
How it works

- Begins with a simple correlation matrix
  - In fact, you don’t need raw data in some cases (rotation)
- EFA attempts to “categorize” variables according to how similar/dissimilar they are to other variables
  - By calculating factor loadings....
- Goal is to produce the minimum number of factors that adequately explains the data
Some Details

• As always, there are options w/how to run EFA, and there are detailed references available

• “Rotation” options to simplify our understanding of factor structure
  – Orthogonal (more complex structure, but more independence of factors)
  – Oblique (simpler structure, but factors may correlate more)
Rotation

• Procedure (ex. Varimax) that searches for linear combinations (i.e. rotations) of original factors so that the variance of the loadings is maximized
Varimax Rotation

- Varimax Rotation is probably the most popular choice for EFA (Kaiser, 1958)
  - Each factor should have small number of items loading heavily on it
    - Each variable should load mostly onto only one factor
    - Thus simplifying our understanding of underlying constructs
Wine Tasting Example

- Factor Rotations in Factor Analyses
  - Herve Abdi, University of Texas at Dallas
- Five Wines are Rated by Seven Questions

Table 1: An (artificial) example for PCA and rotation. Five wines are described by seven variables.

<table>
<thead>
<tr>
<th></th>
<th>Hedonic</th>
<th>For meat</th>
<th>For dessert</th>
<th>Price</th>
<th>Sugar</th>
<th>Alcohol</th>
<th>Acidity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wine 1</td>
<td>14</td>
<td>7</td>
<td>8</td>
<td>7</td>
<td>7</td>
<td>13</td>
<td>7</td>
</tr>
<tr>
<td>Wine 2</td>
<td>10</td>
<td>7</td>
<td>6</td>
<td>4</td>
<td>3</td>
<td>14</td>
<td>7</td>
</tr>
<tr>
<td>Wine 3</td>
<td>8</td>
<td>5</td>
<td>5</td>
<td>10</td>
<td>5</td>
<td>12</td>
<td>5</td>
</tr>
<tr>
<td>Wine 4</td>
<td>2</td>
<td>4</td>
<td>7</td>
<td>16</td>
<td>7</td>
<td>11</td>
<td>3</td>
</tr>
<tr>
<td>Wine 5</td>
<td>6</td>
<td>2</td>
<td>4</td>
<td>13</td>
<td>3</td>
<td>10</td>
<td>3</td>
</tr>
</tbody>
</table>
Unrotated Two-Factor Solution: See any patterns?

Table 2: Wine example: Original loadings of the seven variables on the first two components.

<table>
<thead>
<tr>
<th></th>
<th>Hedonic</th>
<th>For meat</th>
<th>For dessert</th>
<th>Price</th>
<th>Sugar</th>
<th>Alcohol</th>
<th>Acidity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Factor 1</td>
<td>-0.3965</td>
<td>-0.4454</td>
<td>-0.2646</td>
<td>0.4160</td>
<td>-0.0485</td>
<td>-0.4385</td>
<td>-0.4547</td>
</tr>
<tr>
<td>Factor 2</td>
<td>0.1149</td>
<td>-0.1090</td>
<td>-0.5854</td>
<td>-0.3111</td>
<td>-0.7245</td>
<td>0.0555</td>
<td>0.0865</td>
</tr>
</tbody>
</table>
Varimax Rotated Solution: See any patterns?

Table 3: Wine example: Loadings, after varimax rotation, of the seven variables on the first two components.

<table>
<thead>
<tr>
<th>Variable</th>
<th>For meat</th>
<th>For dessert</th>
<th>Price</th>
<th>Sugar</th>
<th>Alcohol</th>
<th>Acidity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Factor 1</td>
<td>-0.4125</td>
<td>-0.4057</td>
<td>0.4790</td>
<td>0.1286</td>
<td>-0.4389</td>
<td>-0.4620</td>
</tr>
<tr>
<td>Factor 2</td>
<td>0.0153</td>
<td>-0.2138</td>
<td>-0.6321</td>
<td>-0.2010</td>
<td>-0.7146</td>
<td>-0.0525</td>
</tr>
</tbody>
</table>

- Seems likely that there are two dimensions
  - One factor of sweetness
  - The other linked to price and complex taste qualities
Next Example using SPSS for EFA

- Data set provided by Mary Ann Coughlin
  - Editor & Co-Author of AIR Intermediate/Advanced Statistics in IR Monograph
  - Twenty-Seven item questionnaire of graduates from a small liberal arts college.
  - Asking about things they gained from their education
Next Example using SPSS for EFA

Twenty-Seven Questions that tap into perceived benefits of an education...

Ripe for EFA!
How would you summarize the data?

- Table of Means?
- Histograms?
### Table of Descriptive Statistics:

<table>
<thead>
<tr>
<th>Description</th>
<th>N</th>
<th>Range</th>
<th>Minimum</th>
<th>Maximum</th>
<th>Mean</th>
<th>Std. Error</th>
<th>Std. Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>Write effectively</td>
<td>621</td>
<td>3</td>
<td>0</td>
<td>3</td>
<td>2.46</td>
<td>0.026</td>
<td>0.648</td>
</tr>
<tr>
<td>Communicate well orally</td>
<td>618</td>
<td>3</td>
<td>0</td>
<td>3</td>
<td>2.21</td>
<td>0.031</td>
<td>0.779</td>
</tr>
<tr>
<td>Acquire new skills and knowledge on my own</td>
<td>614</td>
<td>3</td>
<td>0</td>
<td>3</td>
<td>2.45</td>
<td>0.029</td>
<td>0.714</td>
</tr>
<tr>
<td>Think analytically and logically</td>
<td>616</td>
<td>3</td>
<td>0</td>
<td>3</td>
<td>2.40</td>
<td>0.029</td>
<td>0.716</td>
</tr>
<tr>
<td>Formulate creative / original ideas and solutions</td>
<td>611</td>
<td>3</td>
<td>0</td>
<td>3</td>
<td>2.15</td>
<td>0.032</td>
<td>0.803</td>
</tr>
<tr>
<td>Evaluate and choose between alternative courses</td>
<td>608</td>
<td>3</td>
<td>0</td>
<td>3</td>
<td>2.01</td>
<td>0.033</td>
<td>0.816</td>
</tr>
<tr>
<td>Lead and supervise tasks and groups of people</td>
<td>611</td>
<td>3</td>
<td>0</td>
<td>3</td>
<td>1.89</td>
<td>0.038</td>
<td>0.949</td>
</tr>
<tr>
<td>Relate well to people of different races, nations</td>
<td>615</td>
<td>3</td>
<td>0</td>
<td>3</td>
<td>2.23</td>
<td>0.035</td>
<td>0.873</td>
</tr>
<tr>
<td>Function effectively as a member of a team</td>
<td>610</td>
<td>3</td>
<td>0</td>
<td>3</td>
<td>1.91</td>
<td>0.035</td>
<td>0.860</td>
</tr>
<tr>
<td>Use computers for basic tasks (word processing)</td>
<td>611</td>
<td>3</td>
<td>0</td>
<td>3</td>
<td>2.02</td>
<td>0.042</td>
<td>1.045</td>
</tr>
<tr>
<td>Use computers for complex tasks (graphing)</td>
<td>610</td>
<td>3</td>
<td>0</td>
<td>3</td>
<td>1.00</td>
<td>0.045</td>
<td>1.115</td>
</tr>
<tr>
<td>Place current problems in historical prospective</td>
<td>609</td>
<td>3</td>
<td>0</td>
<td>3</td>
<td>2.23</td>
<td>0.034</td>
<td>0.835</td>
</tr>
<tr>
<td>Identify moral and ethical issues</td>
<td>612</td>
<td>3</td>
<td>0</td>
<td>3</td>
<td>2.09</td>
<td>0.034</td>
<td>0.842</td>
</tr>
<tr>
<td>Understand myself, my abilities, interests</td>
<td>614</td>
<td>3</td>
<td>0</td>
<td>3</td>
<td>2.41</td>
<td>0.032</td>
<td>0.800</td>
</tr>
<tr>
<td>Function independently without supervision</td>
<td>612</td>
<td>3</td>
<td>0</td>
<td>3</td>
<td>2.26</td>
<td>0.035</td>
<td>0.870</td>
</tr>
<tr>
<td>Gain in-depth knowledge of a field</td>
<td>614</td>
<td>3</td>
<td>0</td>
<td>3</td>
<td>2.38</td>
<td>0.030</td>
<td>0.744</td>
</tr>
<tr>
<td>Plan and execute complex projects</td>
<td>604</td>
<td>3</td>
<td>0</td>
<td>3</td>
<td>2.11</td>
<td>0.033</td>
<td>0.809</td>
</tr>
<tr>
<td>Read or speak a foreign language</td>
<td>609</td>
<td>3</td>
<td>0</td>
<td>3</td>
<td>1.38</td>
<td>0.048</td>
<td>1.196</td>
</tr>
<tr>
<td>Appreciate art, literature, music, drama</td>
<td>611</td>
<td>3</td>
<td>0</td>
<td>3</td>
<td>2.12</td>
<td>0.036</td>
<td>0.899</td>
</tr>
<tr>
<td>Acquire broad knowledge in the Arts and Sciences</td>
<td>614</td>
<td>3</td>
<td>0</td>
<td>3</td>
<td>2.19</td>
<td>0.032</td>
<td>0.796</td>
</tr>
<tr>
<td>Develop feminist awareness</td>
<td>613</td>
<td>3</td>
<td>0</td>
<td>3</td>
<td>2.56</td>
<td>0.029</td>
<td>0.710</td>
</tr>
<tr>
<td>Develop awareness of social problems</td>
<td>610</td>
<td>3</td>
<td>0</td>
<td>3</td>
<td>2.36</td>
<td>0.030</td>
<td>0.747</td>
</tr>
<tr>
<td>Develop self-esteem /self-confidence</td>
<td>614</td>
<td>3</td>
<td>0</td>
<td>3</td>
<td>2.35</td>
<td>0.035</td>
<td>0.859</td>
</tr>
<tr>
<td>Form close friendships</td>
<td>617</td>
<td>3</td>
<td>0</td>
<td>3</td>
<td>2.46</td>
<td>0.032</td>
<td>0.803</td>
</tr>
<tr>
<td>Establish a course of action to accomplish goals</td>
<td>614</td>
<td>3</td>
<td>0</td>
<td>3</td>
<td>2.13</td>
<td>0.031</td>
<td>0.780</td>
</tr>
<tr>
<td>Synthesize and integrate ideas and information</td>
<td>612</td>
<td>3</td>
<td>0</td>
<td>3</td>
<td>2.30</td>
<td>0.029</td>
<td>0.720</td>
</tr>
<tr>
<td>Understand the role of science and technology</td>
<td>614</td>
<td>3</td>
<td>0</td>
<td>3</td>
<td>1.55</td>
<td>0.039</td>
<td>0.956</td>
</tr>
<tr>
<td>Valid N (listwise)</td>
<td>537</td>
<td>3</td>
<td>0</td>
<td>3</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Bucket-o-Histograms

Etc.....
What’s the Big Picture?

- Simple to get the “take-home” message from our graduates?
  - Generally happy
  - Simple Descriptives can tell us that.
- But what’s the “big picture” of the benefits of a college education from our college, in the eyes of our graduates?
  - This kind of question is ripe for EFA
SPSS Case Study

• Run Exploratory Factor Analysis
  – Principal-Axis Method
  – With Varimax Rotation

• Interpret the Results!
Six Underlying Constructs Resulting from our Rotated Factor Structure

**Rotated Factor Matrix**

<table>
<thead>
<tr>
<th>Construct</th>
<th>Factor 1</th>
<th>Factor 2</th>
<th>Factor 3</th>
<th>Factor 4</th>
<th>Factor 5</th>
<th>Factor 6</th>
</tr>
</thead>
<tbody>
<tr>
<td>Think analytically and logically</td>
<td>.643</td>
<td>.150</td>
<td>.091</td>
<td>.006</td>
<td>.118</td>
<td>-.026</td>
</tr>
<tr>
<td>Formulate creative/original ideas and solutions</td>
<td>.625</td>
<td>.146</td>
<td>.145</td>
<td>.199</td>
<td>.031</td>
<td>.124</td>
</tr>
<tr>
<td>Synthesize and integrate ideas and information</td>
<td>.587</td>
<td>.202</td>
<td>.217</td>
<td>.113</td>
<td>.208</td>
<td>.072</td>
</tr>
<tr>
<td>Acquire new skills and knowledge on my own</td>
<td>.581</td>
<td>.070</td>
<td>.141</td>
<td>.090</td>
<td>.095</td>
<td>.146</td>
</tr>
<tr>
<td>Plan and execute complex projects</td>
<td>.537</td>
<td>.042</td>
<td>.127</td>
<td>.148</td>
<td>.230</td>
<td>.168</td>
</tr>
<tr>
<td>Write effectively</td>
<td>.512</td>
<td>.263</td>
<td>.094</td>
<td>.027</td>
<td>-.007</td>
<td>.142</td>
</tr>
<tr>
<td>Establish a course of action to accomplish goals</td>
<td>.496</td>
<td>.195</td>
<td>.386</td>
<td>.223</td>
<td>.204</td>
<td>.112</td>
</tr>
<tr>
<td>Evaluate and choose between alternative courses</td>
<td>.485</td>
<td>.175</td>
<td>.176</td>
<td>.397</td>
<td>.087</td>
<td>.067</td>
</tr>
<tr>
<td>Communicate well orally</td>
<td>.460</td>
<td>.119</td>
<td>.265</td>
<td>.207</td>
<td>.037</td>
<td>.168</td>
</tr>
<tr>
<td>Gain in-depth knowledge of a field</td>
<td>.442</td>
<td>.074</td>
<td>.080</td>
<td>.039</td>
<td>.160</td>
<td>.117</td>
</tr>
<tr>
<td>Develop awareness of social problems</td>
<td>.158</td>
<td>.732</td>
<td>.233</td>
<td>.091</td>
<td>.102</td>
<td>.099</td>
</tr>
<tr>
<td>Develop feminist awareness</td>
<td>.082</td>
<td>.560</td>
<td>.148</td>
<td>.025</td>
<td>-.037</td>
<td>.158</td>
</tr>
<tr>
<td>Identify moral and ethical issues</td>
<td>.288</td>
<td>.542</td>
<td>.168</td>
<td>.185</td>
<td>.098</td>
<td>.126</td>
</tr>
<tr>
<td>Place current problems in historical prospective</td>
<td>.303</td>
<td>.433</td>
<td>.040</td>
<td>.166</td>
<td>-.063</td>
<td>.198</td>
</tr>
<tr>
<td>Form close friendships</td>
<td>.122</td>
<td>.163</td>
<td>.569</td>
<td>.143</td>
<td>.077</td>
<td>.084</td>
</tr>
<tr>
<td>Understand myself, my abilities, interests</td>
<td>.328</td>
<td>.275</td>
<td>.547</td>
<td>.159</td>
<td>-.056</td>
<td>.156</td>
</tr>
<tr>
<td>Develop self-esteem/self-confidence</td>
<td>.370</td>
<td>.313</td>
<td>.543</td>
<td>.165</td>
<td>.096</td>
<td>.162</td>
</tr>
<tr>
<td>Function independently without supervision</td>
<td>.372</td>
<td>.100</td>
<td>.455</td>
<td>.170</td>
<td>.165</td>
<td>.189</td>
</tr>
<tr>
<td>Relate well to people of different races, nations</td>
<td>.125</td>
<td>.335</td>
<td>.337</td>
<td>.307</td>
<td>.165</td>
<td>.070</td>
</tr>
<tr>
<td>Lead and supervise tasks and groups of people</td>
<td>.178</td>
<td>.089</td>
<td>.189</td>
<td>.724</td>
<td>.149</td>
<td>.048</td>
</tr>
<tr>
<td>Function effectively as a member of a team</td>
<td>.160</td>
<td>.190</td>
<td>.208</td>
<td>.567</td>
<td>.240</td>
<td>.038</td>
</tr>
<tr>
<td>Use computers for complex tasks (graphing)</td>
<td>.082</td>
<td>-.034</td>
<td>-.024</td>
<td>.112</td>
<td>.601</td>
<td>-.030</td>
</tr>
<tr>
<td>Understand the role of science and technology</td>
<td>.380</td>
<td>-.013</td>
<td>.089</td>
<td>.138</td>
<td>.571</td>
<td>.164</td>
</tr>
<tr>
<td>Use computers for basic tasks (word processing)</td>
<td>.122</td>
<td>.118</td>
<td>.247</td>
<td>.108</td>
<td>.402</td>
<td>.053</td>
</tr>
<tr>
<td>Appreciate art, literature, music, drama</td>
<td>.178</td>
<td>.248</td>
<td>.141</td>
<td>.088</td>
<td>-.054</td>
<td>.650</td>
</tr>
<tr>
<td>Acquire broad knowledge in the Arts and Sciences</td>
<td>.269</td>
<td>.124</td>
<td>.064</td>
<td>.050</td>
<td>.277</td>
<td>.413</td>
</tr>
<tr>
<td>Read or speak a foreign language</td>
<td>.076</td>
<td>.079</td>
<td>.078</td>
<td>.003</td>
<td>.038</td>
<td>.402</td>
</tr>
</tbody>
</table>

Extraction Method: Principal Axis Factoring.
Rotation Method: Varimax with Kaiser Normalization.

a. Rotation converged in 7 iterations.
Six Underlying Constructs Resulting from our Rotated Factor Structure

<table>
<thead>
<tr>
<th>Rotated Factor Matrix</th>
<th>Factor</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
</tr>
<tr>
<td>Think analytically and logically</td>
<td>0.643</td>
</tr>
<tr>
<td>Formulate creative / original ideas and solutions</td>
<td>0.625</td>
</tr>
<tr>
<td>Synthesize and integrate ideas and information</td>
<td>0.587</td>
</tr>
<tr>
<td>Acquire new skills and knowledge on my own</td>
<td>0.581</td>
</tr>
<tr>
<td>Plan and execute complex projects</td>
<td>0.537</td>
</tr>
<tr>
<td>Write effectively</td>
<td>0.512</td>
</tr>
<tr>
<td>Establish a course of action to accomplish goals</td>
<td>0.496</td>
</tr>
<tr>
<td>Evaluate and choose between alternative courses</td>
<td>0.485</td>
</tr>
<tr>
<td>Communicate well orally</td>
<td>0.460</td>
</tr>
<tr>
<td>Gain in-depth knowledge of a field</td>
<td>0.442</td>
</tr>
<tr>
<td>Develop awareness of social problems</td>
<td>0.158</td>
</tr>
<tr>
<td>Develop feminist awareness</td>
<td>0.082</td>
</tr>
<tr>
<td>Identify moral and ethical issues</td>
<td>0.288</td>
</tr>
<tr>
<td>Place current problems in historical perspective</td>
<td>0.303</td>
</tr>
<tr>
<td>Form close friendships</td>
<td>0.122</td>
</tr>
<tr>
<td>Understand myself, my abilities, interests</td>
<td>0.328</td>
</tr>
<tr>
<td>Develop self-esteem / self-confidence</td>
<td>0.370</td>
</tr>
<tr>
<td>Function independently without supervision</td>
<td>0.372</td>
</tr>
<tr>
<td>Relate well to people of different races, nations</td>
<td>0.125</td>
</tr>
<tr>
<td>Lead and supervise tasks and groups of people</td>
<td>0.178</td>
</tr>
<tr>
<td>Function effectively as a member of a team</td>
<td>0.160</td>
</tr>
<tr>
<td>Use computers for complex tasks (graphing)</td>
<td>0.082</td>
</tr>
<tr>
<td>Understand the role of science and technology</td>
<td>0.380</td>
</tr>
<tr>
<td>Use computers for basic tasks (word processing)</td>
<td>0.122</td>
</tr>
<tr>
<td>Appreciate art, literature, music, drama</td>
<td>0.178</td>
</tr>
<tr>
<td>Acquire broad knowledge in the Arts and Sciences</td>
<td>0.269</td>
</tr>
<tr>
<td>Read or speak a foreign language</td>
<td>0.076</td>
</tr>
</tbody>
</table>

Extraction Method: Principal Axis Factoring.
Rotation Method: Varimax with Kaiser Normalization.

a. Rotation converged in 7 iterations.
Six Underlying Constructs Resulting from our Rotated Factor Structure

<table>
<thead>
<tr>
<th>Rotated Factor Matrix</th>
<th>Knowledge Gains</th>
<th>Social/Moral Reasoning</th>
<th>Self-Awareness/Independence</th>
<th>Leadership/Teamwork</th>
<th>Technology</th>
<th>Arts &amp; Humanities</th>
</tr>
</thead>
<tbody>
<tr>
<td>Think analytically and logically</td>
<td>0.643</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Formulate creative / original ideas and solutions</td>
<td>0.625</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Synthesize and integrate ideas and information</td>
<td>0.587</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Acquire new skills and knowledge on my own</td>
<td>0.581</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Plan and execute complex projects</td>
<td>0.537</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Write effectively</td>
<td>0.512</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Establish a course of action to accomplish goals</td>
<td>0.496</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Evaluate and choose between alternative courses</td>
<td>0.485</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Communicate well orally</td>
<td>0.460</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gain in-depth knowledge of a field</td>
<td>0.442</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Develop awareness of social problems</td>
<td>0.158</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Develop feminist awareness</td>
<td>0.082</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Identify moral and ethical issues</td>
<td>0.288</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Place current problems in historical prospective</td>
<td>0.303</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Form close friendships</td>
<td>0.122</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Understand myself, my abilities, interests</td>
<td>0.328</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Develop self-esteem /self-confidence</td>
<td>0.370</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Function independently without supervision</td>
<td>0.372</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Relate well to people of different races, nati</td>
<td>0.307</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lead and supervise tasks and groups of people</td>
<td>0.160</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Function effectively as a member of a team</td>
<td>0.160</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Use computers for complex tasks (graphing)</td>
<td>0.082</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Understand the role of science and technology</td>
<td>0.380</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Use computers for basic tasks (word processing)</td>
<td>0.122</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Appreciate art, literature, music, drama</td>
<td>0.178</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Acquire broad knowledge in the Arts and Sciences</td>
<td>0.269</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Read or speak a foreign language</td>
<td>0.269</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Extraction Method: Principal Axis Factoring.
Rotation Method: Varimax with Kaiser Normalization.

a. Rotation converged in 7 iterations.
Advanced Topics Related to EFA

• **Confirmatory Factory Analysis**
  – Testing hypotheses about factor structure

• **Structural Equation Modeling**
  – Testing structure on the “predictor” and “outcome” side of the equations
  – Path-Analysis-Like model fitting
  – Mediator/Moderator effects testing
  – Just to name a few examples!
Questions? Comments? Course Evaluations Please!!!

Robert Ploutz-Snyder, Ph.D.
Biostatistician NASA JSC
USRA / Division of Space Life Sciences,
Research Associate Professor of Medicine
SUNY Upstate Medical University

What topics needed more info?
Less info? If more, what should I eliminate??