SMU Data Science Program

Experimental Statistics II

Modeling Information and Guide

Section 1 Univariate Linear Modeling

1. Multiple Linear Regression / Model Selection Approaches
2. Two Way ANOVA
3. Time Series
4. Repeated Measures Analysis

Section 2 Multivariate Techniques

1. Multivariate Two Sample Testing / Multivariate Analysis of Variance (MANOVA)
2. Linear Discriminate Analysis
3. Principle Components
4. Canonical Correlations

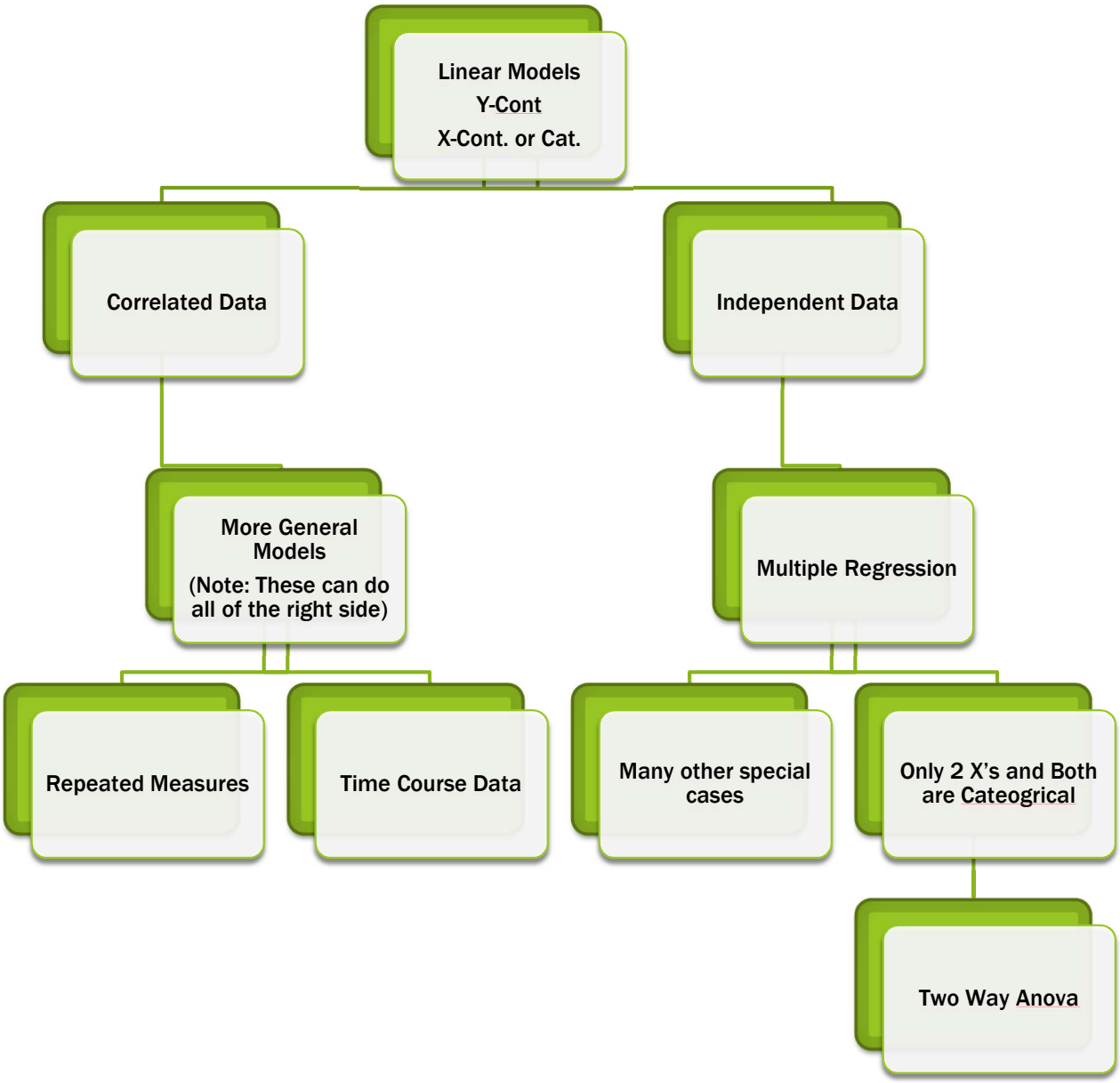
Section 3 Analysis of Counts and Binary Outcomes

1. Comparison of Proportions or Odds
2. Additional Tools for Count Data
3. Logistic Regression

Section 4 Special Topics

1. Unsupervised Machine Learning
2. The Bootstrap

Section 1 Univariate Linear Modeling



Section 1 Univariate Linear Modeling -Topic 1: Multiple Linear Regression

1. Purpose/Questions of Interest/Research Questions

Used for 2 main purposes:

1) Want to develop a statistical model to predict an outcome
2) Want to know the association between the outcome and an explanatory variable(covariate) while possible adjusting for other variables.

2. Assumptions / Structure of the Data

Used in situations with 1 defined dependent variable that is continuous and 1 or more independent variables that can be continuous or categorical

Linearity - The relationship between the dependent variable and the continuous independent variables must be linear

Normality – Residuals of the linear model is assumed to be normally distributed

Equal Variance – The variance of the residuals is constant for every combination of independent variables and thus constant across all of the predicted values

Independence – Observations are identically and independently distributed (i.i.d.)

3. Sampling Method

Random sample of population of interest, with fixed X's.

Can be observational or experimental study designs.

Can only be cross sectional (observations at a single point in time or space, no repeated measures)

4. Special Descriptive Statistics and/or Graphics

Continuous variables – Using 5 number summary, histograms, box plots, scatter plots

Categorical variables – usual count tables/percents. Also look at summary statistics of dependent variable by levels of the categorical variable. Bar charts, Pie graphs, etc.

Scatterplot matrix/ Proc Corr - Examine relationships between the dependent and the independent variables. Also examine for possibility of multicollinearity.

Diagnostics Statistics & Plots

Residuals

Raw= Observed – Predicted

Standardized= Z score, look for values more extreme than +/- 2 or 3

Studentized= t score that takes into leverage into account, look for values more extreme that +/- 2 or 3

Cook's D=Uses raw residuals and leverage to see how coefficient estimates are affected without the current observation. Look for values greater than 1.

Leverage= how far away an observation is relative to the center of all of the explanatory variables

Graphs for all these are generated in SAS.

Multicollinearity

Variance Inflation Factor (VIF) – look out for values above 10

Scatterplot matrices and correlation values

5. What can we estimate?

Regression coefficients are estimated by Least Squares

LASSO,LARS, Elastic Net, and Ridge Regression are all penalized version of Least Squares

Confidence intervals

Can be obtained for each regression coefficient as well as new predicted values

Coefficient interpretation – Continuous, for every one unit increment of X, Y increases by coefficient value, while holding the other explanatory values fixed. Categorical, adds an additional value to the intercept.

6. Hypothesis Testing

High Level:

Overall significance of model: Null: All B's=0, Alternative: At least one is not 0 (Ftest)

Lower Level:

If overall test is significant, we want to know which ones are not 0.

Null: intercept or coefficient being tested is 0, Alternative: intercept of coefficient being testing is not 0 (known as partial F-tests, but we typically use the T-test equivalent)

Testing is only valid when assumptions are met (See #2 above)

Note: Contrasts can be written to test for differences between coefficient estimates or more generally, any linear combination of the coefficients.

Other types of partial F-tests can be conducted by fitting a full model versus a reduced model and conducting an F-test

7. Relationship between other techniques / Other Info

Multiple linear regression is the most general technique for independent data under the assumption of normality, constant variance. Two way ANOVA along with many other ANOVA type designs are special cases. The general concept of multiple regression can be extended to handle response variables who are not continuous such as counts or proportions (logistic regression, generalized linear models). For correlated data, we can extend the multiple regression frame work to handle time series or repeated measures.

8. General Analysis Flow

1. Identify the question of interest (See #1)
2. Exploratory analysis

Descriptive statistics and scatterplots

Assess potential outliers that may be errors in recording

Remove any redundant variables that will create problems with multicollinearity

Assess linearity of variables and conduct appropriate transformations

Finalize the full model in which to conduct analysis (this can be done manually or for many variables a model selection technique could help to whittle things down)

3. Analysis

Fit full model and assess model assumptions through residual diagnostics.

Conduct overall F-test for significance.

If significant, perform individual t-test for regression coefficients or other testing of interest to answer the question

Any insignificant factors can be removed and the analysis can be rerun. Likewise for observations that are outliers and it makes sense to remove them.

If prediction is the key goal and data is large enough. Assess how well the data set performs on an independent data set.

4. Reporting

Provide the final regression model equation.

Provide appropriate interpretation to regression coefficients that are significant and you wish to discuss.
For prediction, provide predicted values as well as 95% prediction intervals.
Optional: Conduct secondary analysis comparing different model selection techniques to see if the story changes much. In large number variables it likely will, but is important none the less to see that other predictors can do just as good of a job as the ones you picked.

9. SAS Computer Programs

```
Proc reg data=      ;
Model Y=X's / p vif;
Output out=out p=p cookd=cookd r=rawres rstudent=jackres student=standres ucl=ucl lcl=lcl h=lev;
Run;quit;
*SAS gives you many diagnostic plots.;
*Below is some code to do similar things.;

Proc gchart data=out;
Vbar rawres standres jackres lev;
Run;
Proc gplot data=out;
 Plot p*Y (rawres stanres jackres)*(p X's)  cookd*(Y p);
Run;quit;
```

10. Limitations of the technique / Things to look out for

All the assumptions listed in 2.
Analysis only generalizes to range of observations in the explanatory variables, excluding outliers (NO extrapolation)
Model is obviously data driven and final model will vary between analysts
Association does not imply causation
Multicollinearity and confounding need to be assessed well

Model Selection

FORWARD

Adds the covariates to the model one at a time in the order presented in the model statement. If the variable is statistically significant at the specified alpha then the covariate stays in the model and the next covariate is entered. Once a variable is "included" it cannot be dropped.

BACKWARD

STEPWISE

LASSO

Uses a penalized least squares approach that squeezes the regression coefficients to 0 when the penalty is large. The algorithm starts with a large penalty and gradually relaxes the penalty to allow for a single variable to be added into the model (the coefficient is no longer 0). At each step, a model selection criterion such as AIC, SBC, AICc, etc can be used to obtain an optimal model. Additionally, the user can specify cross validation techinques to obtain an optimal model as well.

LARS

Similar to the approach of LASSO but formulated slightly different. LARS can produce the LASSO solutions in a more efficient way.

ELASTIC NET

Procedure identical to LASSO however the penalty is different. Elastic net uses a combination of both the LASSO penalty as well as the RIDGE regression penalty.

Section 1 Univariate Linear Modeling -Topic 2: Two Way ANOVA

1. Purpose/Questions of Interest/Research Questions

Two Way ANOVA:

Used for simplifying multiple regression analysis when there are two categorical explanatory factors (the block and the treatment). The focus here is typically not prediction, but rather understanding what factor level combinations can change the mean of the response variable. Each factor can have multiple levels. Particularly helpful to determine if the model is additive or non-additive, if significant interaction(non-additive) is present in the data, or using linear combinations to contrast multiple groups in a single test.

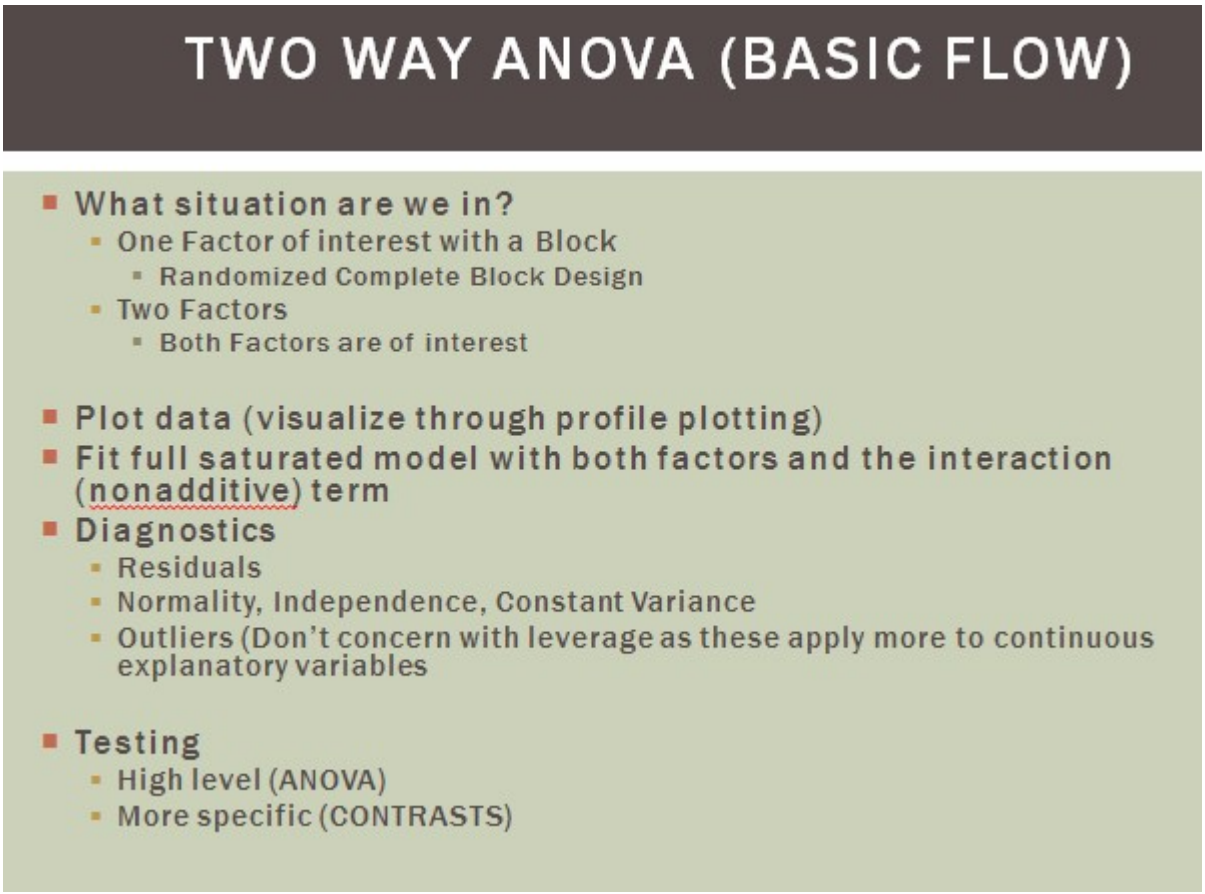
More General Definition of Multiway ANOVA:

Also known as Factorial ANOVA, due to the use of a factorial design where two (or more)independent variables are manipulated in a single study so that the treatment conditions represent all possible combinations of the various levels of the independent variables.

Double click this icon to activate embedded slide deck on this subject:

Session5PPT-2WayANOVA.pptx

There are two types of two way ANOVA – those with significant interactions (aka NonAdditive) and those without (additive). They require slightly different approaches as outlined below.



2. Assumptions / Structure of the Data

Because it is multiple regression, it has the same underlying assumptions, with the following modifications to the structure of the data:

- **Level of Measurement:** The criterion variable should be assessed on an interval or ratio level of measurement (i.e. continuous). Both predictor variables should be nominal-level variables (i.e. categorical variables).
- **Independent observations (No repeated measures)**
- **Random sampling**
- **Normal distributions**: Each cell (factor level combination) should be drawn from a normally distributed population. If each cell contains more than **30 participants**, the test is robust against moderate departures from normality (CLT kicks in…)
- **Homogeneity of variance**: Cell populations (factor level combinations) should have equal variances. If the number of participants in the largest cell is **no more than 1.5 times greater** than the number of participants in the smallest cell, then the test is robust against violations of the homogeneity assumption.

The assumptions of normality and equal variances are typically explored through examination of the residuals (exactly like multiple regression). It is beneficial for the sample sizes for each factor combination to be the same as it is the most powerful setup, but is not required.

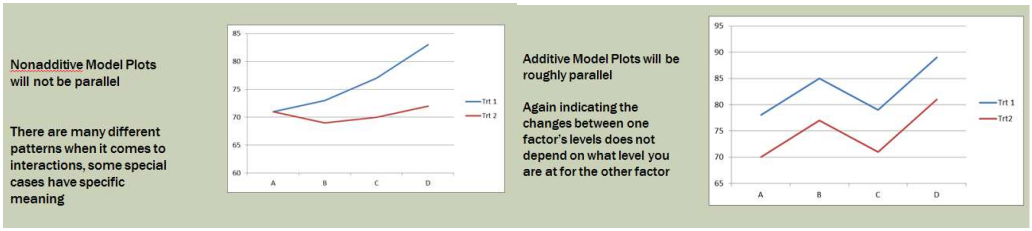
3. Sampling Method

Random sample from the population of interest for each of the factor level combinations.
In experimental settings, the experimental unit is randomly assigned to a factor level combination and the response is then measured.
Can be observational or experimental study designs.
Special case of two way anova:

1. Randomized Complete Block Design
   Only one factor is of interest while the other factor serves as a block to control variability across a potential known confounder

4. Special Descriptive Statistics and/or Graphics

a. Residual Diagnostics as seen in Regression
b. Summary statistics, histograms, boxplots, by factor levels and combinations
c. Coded scatterplots to help assess whether an interaction might be present



5. What can we estimate?

As with multiple regression, the regression coefficients, and the common variance parameter are estimated with least squares. Raw regression coefficients are a little bit confusing to interpret for two way anova, the power of PROC GLM allows us to obtain the LSmeans, which is simply the least squares estimates for the means of each factor level combination, something we can readily understand. With these estimates, we can calculate addition estimates regarding any individual factor's means (when no interaction). For all the means we can also provide estimates of the standard errors to conduct tests and create confidence intervals.

In a nut shell:

a. Mean estimate and standard error for the levels of just predictor A
b. Mean estimate and standard error for the levels of just predictor B
c. Mean estimate and standard error for the level combinations of predictor A and B

6. Hypothesis Testing

High Level:
There are potentially three overall F-tests that can be conducted. The order in which they are conducted matters. The full (saturated) model is used to conduct the analysis (both factors and their interaction term)

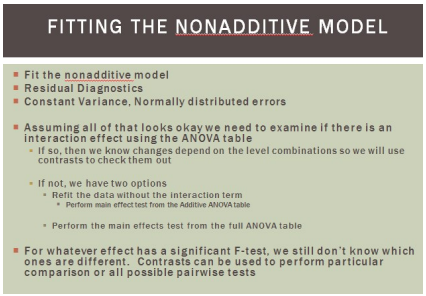
First test for significant interaction between the main effects. If significant, then jump to lower level contrasts.

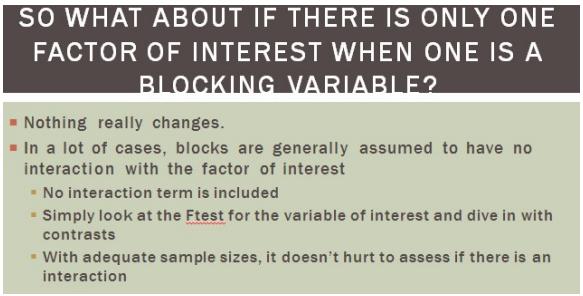
Null: The changes in means within predictor A does not depend on the level of predictor B.
More simply: "There is no interaction between the two factors".

If interaction is not significant, then test for simple effects one main effect at a time.
Null: There is no difference in means between the different levels within the factor (main effects)

Lower Level (Contrasts):

For any of the above overall F-tests that are significant, additional comparisons can be made between specific means from the different factors and their levels through contrast. For any unplanned comparisons, remember to adjust for multiple tests using Bonferroni or Tukey type procedures.





7. Relationship between other techniques / Other Info

An extension to the one way ANOVA as well as a special case of multiple regression. It is also the most simplest case of providing a way to analyze a completely randomized block design when there is only one explanatory variable of interest and it is categorical (CRBD). GLM is the preferred PROC as it handles unbalanced sample sizes appropriately.

8. Analysis Overview

- Statement of the problem
- Nature of the variables
- Summary stats (PROC MEANS) and mean plots (PROC GPLOT)
- Fit Full model with interaction term, Residual Diagnostics for assumptions
- Statistical Tests
- Clearly define Null Hypothesis/Alternative Hypothesis
  1. ANOVA summary table
  2. Interaction F Test First
  3. Main Effect Tests only if Interaction is not significant
- Overall High Level Conclusion of High Level Hypothesis Test

- Magnitude of treatment effect (is it practically significant?)
  1. Obtained by conducting contrasts either specifically or all pairwise. Unplanned tests must be multiple test corrected.
  2. Clearly defining the null/alternative hypothesis here and providing conclusions of the tests.
- Additional figures representing results

9. SAS Computer Programs

```
PROC GLM DATA = dataset-name;
        CLASS    predictor predictorB;
        MODEL criterion-variable =        predictorA predictorB predictorA*predictorB;
        LSMEANS predictorA predictorB predictorA*predictorB;
        LSMEANS predictorA predictorB predictor*predictorB/ PDIFF  ADJUST=TUKEY
ALPHA=alpha-level;
        CONTRAST 'High vs Low Adoptive (F-test)'   predictorA 1 -1 ;
        ESTIMATE   'High vs Low Biological (t-test)'    predictorB 1 -1;
Run;
```

10. Limitations of the technique / Things to look out for

Unbalanced designs require the use of LSMeans statement. The SAS code has used that here as the LSMeans statement will not significantly alter the results in the event of a balanced design, so there is no reason to not use as the base case. Also, if the factorial design uses continuous variables instead of categorical, will need to go to ANCOVA.

Unplanned comparisons must be adjusted for multiple comparisons. Factor levels can sometimes have quite drastic variances (not equal variances). This was not discussed but a two way anova can be run in PROC mixed allowing for the variances to be unequal. See the SAS help for the appropriate option.

Section 1 Univariate Linear Modeling -Topic 3: Time Series

1. Purpose/Questions of Interest/Research Questions

Conduct multiple linear regression analysis when the data are collected over equally spaced time points.
Used for 2 main purposes:

1) Want to develop a statistical model to predict an outcome (we didn't discuss how to perform prediction in class, see SAS help for some initial guides, there is some complexity here)
2) Want to know the association between the outcome and an explanatory variable (covariate) while possible adjusting for other variables.

2. Assumptions / Structure of the Data

Used in situations with 1 defined dependent variable that is continuous and 1 or more independent variables that can be continuous or categorical. For time series we need at least 100 observations to effectively assess the correlation properties of the data.

Double click this icon to activate embedded slide deck on this subject:

Session4PPT.pptx

Linearity - The relationship between the dependent variable and the continuous independent variables must be linear
Normality – Residuals of the linear model is assumed to be normally distributed
Equal Variance – The variance of the residuals is constant for every combination of independent variables and thus constant across all of the predicted values
Independence – Not needed. The errors must be identically distributed but not independent. An assumption will need to be made on what type of correlation structure the data has (Possibly autoregressive, moving average, combination of the two, ARIMA, etc.)
Time Series related assumptions – We did not go in too much detail here, but the main assumption here for most correlation structures is a property in time series known as a stationary process.

3. Sampling Method

Response and explanatory variables are observed on the same experimental unit over equally spaced time points.
Can be observational or experimental study designs.

4. Special Descriptive Statistics and/or Graphics

Simple plots of data over time, Scatter plots of response vs X's, descriptive statistics, see (MLR section)

Residual Diagnostics
Autocorrelation function = Plots the estimated correlation of the data as a function of lag (how far apart they are in time) Patterns in the ACF give insight to what correlation structure is there.
Partial autocorrelation function = Plots the partial correlation function as a function of lag. Partial autocorrelation plots can tell you what autoregressive process you have by observing where the partial correlation quickly drop to 0. For example, if the partial autocorrelations are high for lag 1,2, and 3, and then we see a significant drop from 4 and on that hover around 0, this is an indicator that we have an AR(3) process. Similar patterns can be seen for Moving Average type processes.

5. What can we estimate?

Once a specific time series process is specified (our strategy was to let a model selection criterion such as AIC help us decide), PROC AUTOREG appropriately estimates the regression coefficients in the model, as well as the corresponding coefficients that control the time series process.

Confidence intervals

Confidence intervals for the regression coefficients can be obtained and uses an appropriate estimate of the standard error. Prediction intervals can also be obtained but we did not cover this.

Coefficient interpretation: Same as MLR

6. Hypothesis Testing

High Level:

Overall significance of model: Null: All B's=0, Alternative: At least one is not 0 (Ftest)

Lower Level:

If overall test is significant, we want to know which ones are not 0.

Null: intercept or coefficient being tested is 0, Alternative: intercept of coefficient being testing is not 0 (known as partial F-tests, but we typically use the T-test equivalent)

Testing is only valid when assumptions are met (See #2 above)

7. Relationship between other techniques / Other Info

The main focus on doing multiple regression is still here for this type of analysis, we are just adding the appropriate modeling changes and analysis when the data is collected over equally spaced time points. The main rule of thumb here is that you really need at least 100 observations to adequately assess what type of correlation structure is present in the data.

8. General Analysis Flow

1. Fit a multiple regression model using standard techniques such as PROC REG and obtain the residuals
2. Use time series diagnostics to assess the time series correlation structure. Make a decision on what structure the time series is (autoregressive, moving average, etc).
3. Other than the rule of thumb, ACF and PACF are hard to use sometimes to determine an appropriate model when time series have combinations of cyclical patterns, moving averages, autoregressive,etc. (See more in depth texts)
4. If you feel the correlation is an AR process, go ahead and use proc autoreg to fit one. Maybe fit multiple AR(1) , AR(2), or AR(3), and use AIC, BIC to pick the best fit. (MODEL selection techniques can still apply)
5. Conducting Overall F-tests, if significant conduct individual t-test on the regression coefficients

9. SAS Computer Programs

```
Proc autoreg data=          ;
Model Y=X's  / nlag=1;  *nlag for autoregressive process;
Run;
```

For Moving average or combinations of AR and MA see PROC ARIMA which was not covered in the course.

See PROC Timeseries for tools to get ACF, PACF, and other time series related plots and tests.

10. Limitations of the technique / Things to look out for

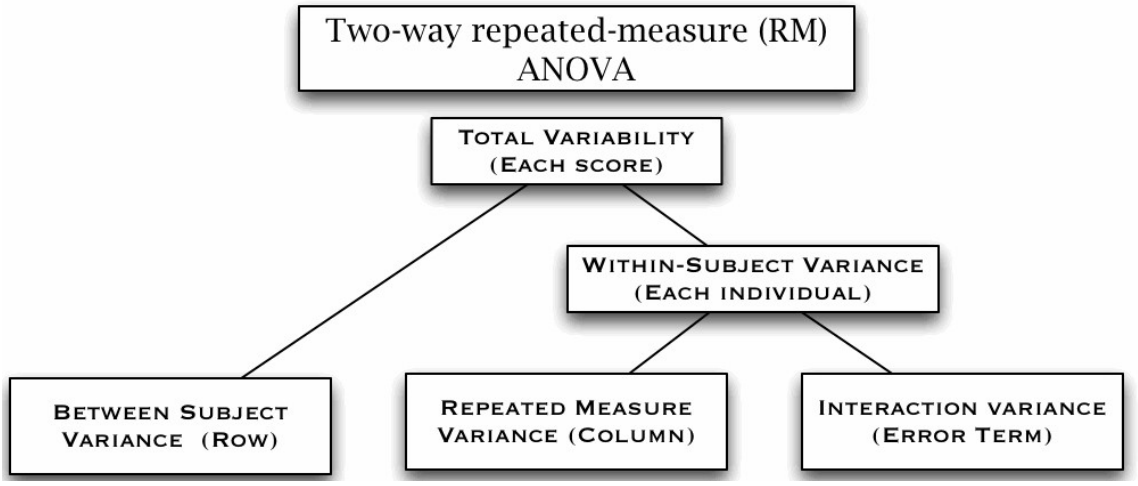
Assessing the appropriate time series process to include in a time series regression analysis can be a bit tricky. There are some general guidelines to try to determine what process to use as a model using patterns seen through the ACF and PACF. These can sometimes be misleading. There are more advanced ways to investigate appropriate models not covered in this course. PROC AUTOREG has a model selection criterion to help determine the appropriate AR(x) process but venturing outside of AR(x) processes should be approached with caution and handled with care.

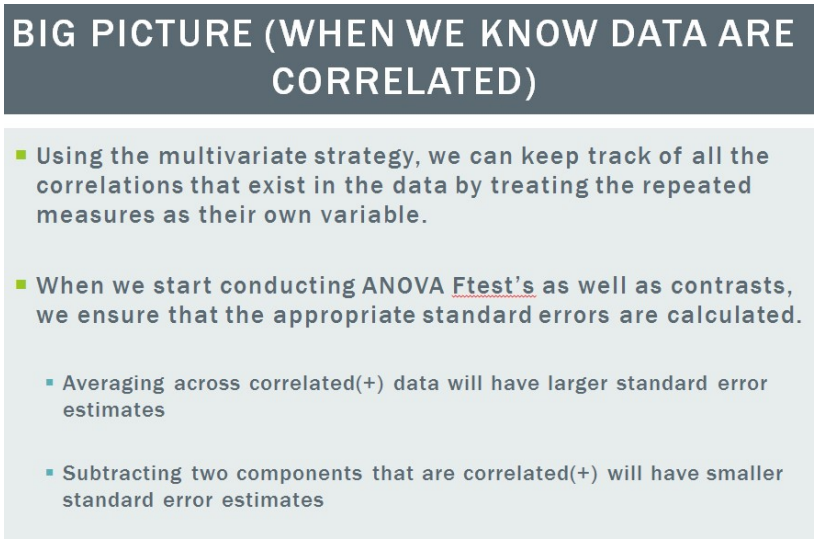
Section 1 Univariate Linear Modeling -Topic 4: Repeated Measures (RM ANOVA)

1. Purpose/Questions of Interest/Research Questions

A special kind of multivariate response where the same variable is measured on each subject at multiple times or locations. So there is technically just one response variable. There are several experimental designs frequently used in repeated measure analysis such as longitudinal studies, cross over experiments, and split plot experiments.

Note, that in class we explored repeated measures in a two way anova setting that fits the split plot experiment scenario which is a very common study design used in practice. Below is a simple diagram illustrating how subject variability is taken into account using "Score" as the response variable name (this is under the sphericity assumption). Repeated measures however can be conjured up in any type of general multiple regression set up in which observations are observed on the same subject or item





2. Assumptions / Structure of the Data

All repeated measure analyses involve data that have multiple measurements for each subjects at different intervals of time. These analyses are inherently multivariate, meaning there are more than one response variable being examined with the explanatory variables.

Double click this icon to activate embedded slide deck on this subject:

Session6PPT.pptx

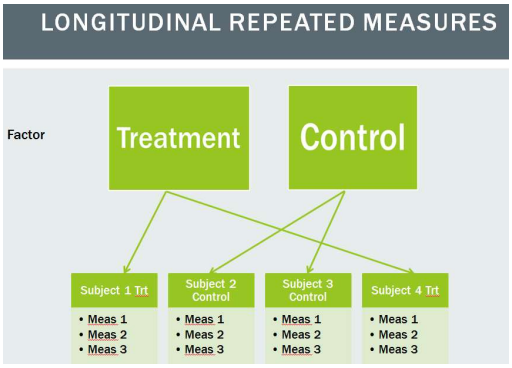
Similar to ANOVA, assumptions are as follows:

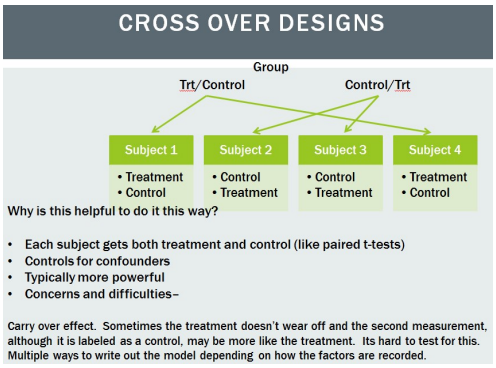
- **Level of Measurement:** The criterion variable should be assessed on an interval or ratio level of measurement (i.e. continuous). Both predictor variables should be nominal-level variables (i.e. categorical variables).
- **Independent observations** - although observations on the same subject are obviously not independent of each other, observations between subjects must be considered independent.
- **Random sampling**
- **Multivariate Normal distributions**: there should not be substantive evidence against normality within each repeated measure component for each of the factor level combinations.
- **Symmetry Condition: Sphericity (Univariate)**: The Univariate model of repeated measures assumes sphericity, which assumes that the correlation between the repeated measures is the same throughout. Proc GLM has a measurement that shows whether this assumption is met (Mauchly's test). If the sphericity assumption is not met, then the researcher should move to either a full blown multivariate assumption with the repeated measures, or consider modeling the repeated measures with various other correlation structures like autoregressive AR(1)**.**
- **Homogeneity of variance:** Variance of the residuals should be constant as in regression and two way anova.

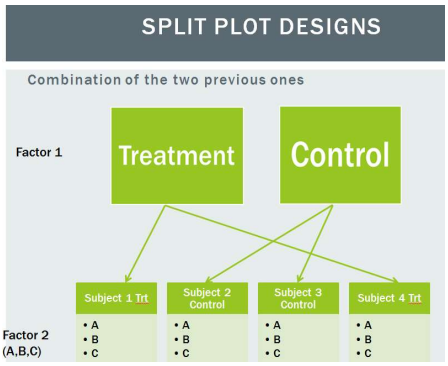
3. Sampling Method

Sampling methods can vary depending on the study design. In general, independent subjects are randomized into groups. Those individuals then have multiple measurements taken.

Longitudinal studies - Repeated measurements on the same variable for the same subjects over time.
Crossover experiments – each subject receives more than 1 treatment, usually with a recovery period in between.
Split Plot Experiments – subjects are randomized to allocate experimental units to levels of the first treatment, then the order of presentation for a second treatment is randomized. All subjects receive all levels of the second treatment. Can be down over time or at several locations.

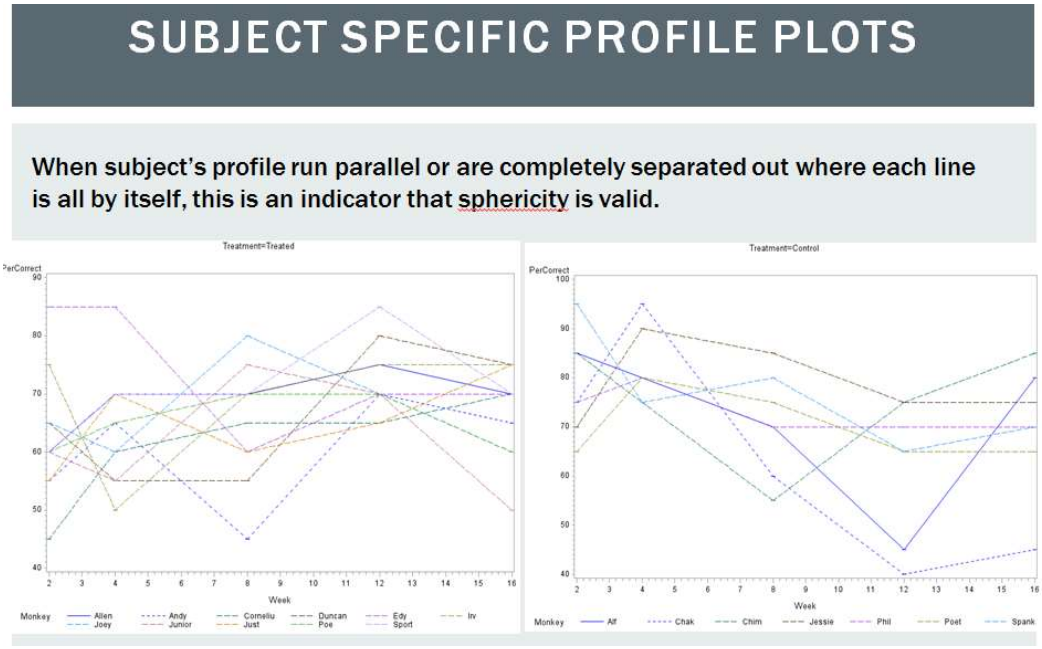






4. Special Descriptive Statistics and/or Graphics

For repeated measures anova, the fundamental graphics and statistics are the same for two way anova. We can include additional subject specific profile plots to help see if sphericity is an adequate assumption. The below graph shows a very weak form of sphericity as the between subject variability is not very high.



5. What can we estimate?

A repeated measure anova allows us to obtain the LSmeans estimates, which is simply the least squares estimates for the means of each factor level combination, exactly as with the two way ANOVA. With these estimates, we can calculate addition estimates regarding any individual factor's means (when no interaction). For all the means we can also provide estimates of the standard errors to conduct tests and create confidence intervals.

In addition, we are estimating the correlation values (Covariance Matrix) between the repeated measures.

The take home difference between repeated measures anova and the two way anova approach is that the standard error estimates for the LSMEANs are accounting for the repeated measures correlations. This ensures that any tests from contrasts that we would like to conduct are the most appropriate.

6. Hypothesis Testing

The testing procedure and general framework is the same for repeated measures as the two way anova. However, before any testing can be performed, an additional decision must be made by the statistician on how the correlation of repeated measures will be modeled. Typical choices include full blown multivariate (SAS PROC MIXED type=UN), sphericity (SAS type=CS), or others such as AR(1) (SAS type=AR(1)). When sample sizes are large, it is best to leave it unstructured as it is the most robust.

The statistician can conduct Mauchly's test for sphericity to help make a decision.
Null: The correlation between the repeated measurements is the same (constant)

An alternative form of making a decision is in using PROC mixed. The procedure is to fit all the scenarios and choose the model with the best model selection criterion (AIC, BIC, etc.)

7. Relationship between other techniques / Other Info

In our case and in our discussion, we are dealing with 2 categorical factors like a Two Way ANOVA but dealing with nonindependent data through repeated measures. The techniques discussed for the repeated measures anova scenario translates directly to more general regression problems with continuous explanatory variables. The analysis plan is the same as multiple regression, just making sure the repeated measures are modeled as well as possible.

8. General Analysis Overview

- Statement of the problem
- Nature of the variables
- Summary stats (PROC MEANS) and mean plots (PROC GPLOT)
- Fit Full model with interaction term, Residual Diagnostics for assumptions
- Use model selection and/or tests for sphericity to determine the best way to handle the repeated measure correlation

- Statistical Tests
- Clearly define Null Hypothesis/Alternative Hypothesis
  1. ANOVA summary table
  2. Interaction F Test First
  3. Main Effect Tests only if Interaction is not significant
- Overall High Level Conclusion of High Level Hypothesis Test
- Magnitude of treatment effect (is it practically significant?)
  1. Obtained by conducting contrasts either specifically or all pairwise. Unplanned tests must be multiple test corrected.
  2. Clearly defining the null/alternative hypothesis here and providing conclusions of the tests.
- Additional figures representing results

9. SAS Computer Programs

The instructor is partial to proc mixed as the SAS coding is much more similar to the other procs. REG, GLM, ANOVA, etc. PROC MIXED is also more flexible in handling more study designs such as the monkey experiment discussed in class.

```
proc mixed data=monkey;
class Treatment Memory Monkey Week;
model PerCorrect=Treatment Week Treatment*Week;
repeated Week/ type=CS subject=Monkey;
lsmeans Treatment*Week / pdiff tdiff;
estimate 'w2 treated vs control' Treatment -1 1 Treatment*Week -1 0 0 0
0 1 0 0 0 0;
estimate 'What is this contrast?' Treatment -1 1 Treatment*Week -.5 -.5
0 0 0 .5 .5 0 0 0;
estimate 'w2 vs w12' Treatment -.5 .5 Treatment*Week 0 0 0 0 0 .5 0 0 -
.5 0;
run;
```

Note the following SAS options in regards to the model correlation structure:

- repeated Week/ type=CS subject=Monkey; *Sphericity;
- repeated Week/ type=UN subject=Monkey; *Unstructured Full Multivariate;
- repeated Week/ type=AR(1) subject=Monkey; *Autoregressive;
- There are half a dozen other ones we can try (see help);
- Each time you run proc mixed, the top of the result file will give the models AIC BIC etc.
  - Note: You can only compare the models when the model statement is exactly the same (Model PerCorrect=Treatment Week Treatment*Week;)

Multivariate Version in GLM (Data set must be structured with the repeated measures in separate columns. GLM will throw out observations with missing data.)

```
data simpleRM;
input subject time1 time2 time3;
datalines;
1 30 28 34
```

```
2 14 18 22
3 24 20 30
4 38 34 44
5 26 28 30
;
run;
/* Repeated Measures ANOVA Code */
PROC GLM data=simpleRM; /*data option not necessary, but it's good practice */
  model time1 time2 time3 = / nouni;
 /*nouni = no univariate tests. If not specified, SAS will run three */
  /* univariate ANOVA, one for each level of time */
  repeated time 3 / printe;
  /* repeated tells SAS that the variables on the LHS of the model statement */
  /* are repeated measures and not separate variables */
```

10. Limitations of the technique / Things to look out for

Multivariate techniques are really helpful in making sure we are conducting the most appropriate analysis. The downside is, in any multivariate setting, the more variables you have the higher the sample size is required to actually conduct the analysis. For smaller sample sizes we have to rely on assumptions, like sphericity, in order to simplify the problem.

One possible confusing aspect to repeated measures is that we are using multivariate techniques to solve a univariate problem. Here we classify RM ANOVA as a univariate technique as the goal is to analyze a single response variable. However, multivariate analysis is needed to handle the repeated measures so it can also be classified as a multivariate technique.

Section 2 Multivariate Techniques -Topic 1: MANOVA

1. Purpose/Questions of Interest/Research Questions

MANOVA is similar to ANOVA in that it tests for significant differences between two or more groups of participants. The important difference is that ANOVA involves just ONE response (dependent) variable whereas MANOVA is appropriate when there is MORE THAN ONE response variable. The point is that you can perform a **single computation** that **determines** whether there is a **significant difference** between treatment groups **when compared simultaneously**. This is important as the method accounts for the correlation that possibly exists between the responses. The overall MANOVA test, appropriately controls type-I errors compared to conducting multiple ANOVA analysis for each variable. (See section 7 for a list of multivariate advantages)

Other needs: Sample size needs to be sufficient, samples within each factor level need to be larger than the number of levels.

Remember the extension from one variable to many. With one variable, we just have a variance to worry about, multiple variables we need a matrix (Covariance matrix) to book keep all the variances as well as the covariances for each pair of response variables that measure how the variables depend on one another.

2. Assumptions / Structure of the Data

Double click this icon to activate embedded slide deck on this subject:

Session7PPT.pptx

Similar to ANOVA but extended to a multivariate parameterization, assumptions are as follows:

- **Level of Measurement:** The dependent variables should be assessed on an interval or ratio level of measurement (i.e. continuous). Both predictor variables should be nominal-level variables (i.e. categorical variables).
- **Independent observations**
- **Random sampling**
- **Multivariate Normal distributions**: At each level of the factor (predictor), it is assumed that the response vector is Multivariate normal. Each level can have its own mean vector.
- **Homogeneity of covariance matrices:** Multivariate extension to equal variance. Although the mean vectors can be different across the levels, it is assumed that the covariance matrix is the same across the factor levels.
- **Sample Size:** : Sample size needs to be sufficient, samples within each factor level need to be larger than the number of levels
- **No Outliers**

3. Sampling Method

Generally, same as ANOVA with the following caveat:

- Sample size needs to be sufficient, samples within each factor level need to be larger than the number of levels.

4. Special Descriptive Statistics and/or Graphics

When there are only 2 or 3 response variables, scatterplots and 3d scatterplots can be used to view the data (color code by factor level)

Basic descriptive statistics and graphs used in ANOVA can be applied here as well one variable at a time to assess for univariate normality, constant variance, and outlier screening.

5. What can we estimate?

Generally the same as ANOVA. For univariate ANOVA, estimates for the mean of each factor level can be obtained. The same is true for MANOVA except the estimates for the means are obtainable across all the response variables. So, technically speaking, we can estimate the mean vector for each of the factor levels of the multivariate response.

In addition, using the assumption of equal covariance matrices, we can estimate the individual variances for each response as well as their dependencies through covariance estimates.

6. Hypothesis Testing

OVERALL MANOVA TEST (HIGH LEVEL)

Null: population means for entire set of response variables do not differ across treatments (factor levels) 1, 2, ..., *k*
Note: This does not mean the means within one response variable should be the same as means

$$H_0: \begin{bmatrix} \mu_{11} \\ \mu_{12} \\ . \\ . \\ . \\ \mu_{1p} \end{bmatrix} = \begin{bmatrix} \mu_{21} \\ \mu_{22} \\ . \\ . \\ . \\ \mu_{2p} \end{bmatrix} = \ldots = \begin{bmatrix} \mu_{g1} \\ \mu_{g2} \\ . \\ . \\ . \\ \mu_{kp} \end{bmatrix}$$

in another response variable.

Alternative: Populations means for an entire set of response variables differs in at least one the treatments (factor levels).

Test is conducted by using one of the following test statistics

- Wilks Lambda
- Pillai's Trace
- Hotelling Lawley Trace
- Roys Greatest Root

Difference in the test statistics is simply different approaches of reducing a ratio of two covariance matrices (ratio of Mean Squares for univariate) down to a single value (See ppt slides). Note: The F-tests that SAS conducts are exact for large sample sizes. When sample sizes are small SAS can conduct more exact tests, but the accuracy depends on the situation.

Univariate ANOVA testing (Lower Level)

If the overall MANOVA test is rejected, we resort back to univariate anova modeling approaches to determine which response variables contain the differences. Multiple testing corrections must be applied.

7. Relationship between other techniques / Other Info

General Rundown for the advantage of multivariate analysis

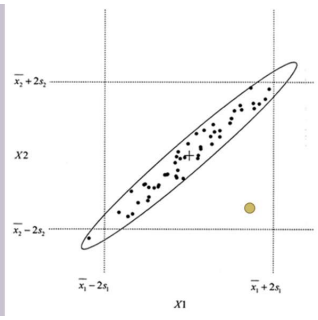
- Type-I error: Probability(reject the Null when the Null is true) aka False Positive
- Power: Probability(fail to reject the Null when the Null is FALSE) aka reject when you should reject – false negative
- Analyzing multiple response variables one at a time is another way of multiple testing, just like in our one variable analysis when we do different types of comparisons. We could do a single comparison but multiple times across different response variables.

- The idea is that correlated data tend to yield similar results so when you commit a type-I error for variable 1 you are now setting yourself up for another type=I error for the correlated variable 2.
- Also correlated variables are more consistent and its easier to identify scenarios that are not null and thus reject more often (increasing power)

Suppose our null hypothesis for the mean of height and weight is given by the yellow dot

One at a time confidence intervals say for both variables we will not reject the null hypothesis

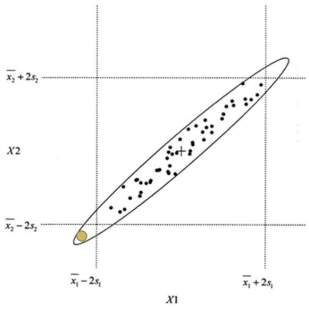
Joint confidence ellipse (taking into account correlation) says the hypothesized value is outside of the confidence ellipse indicating that we should reject the null.



Suppose are null hypothesis for the mean of height and weight is given by the yellow dot

One at a time confidence intervals say for both variables we will reject the null

Joint confidence ellipse (taking into account correlation) says the hypothesized value is inside of the confidence ellipse indicating that we should fail to reject



8. General Analysis Overview

Generally speaking, the first steps are the same as for ANOVA:

a. Make sure everything looks correct and there is no missing data, check for outliers:
   Proc Means data = test;
   Run;
b. Identify the multiple responses and the single categorical factor
c. Plot DATA as much as you can, Two variable scatterplots, three variable 3d plots, etc. Assumption is the errors should be multivariate normal. So plots of the raw data pairwise, should have linear elliptical type relationships with possible different locations, but the shapes of the ellipses for each group should be the same (Equal covariance matrices).
d. Fit MANOVA along with univariate ANOVA models
   i. It wouldn't hurt to examine, univariately, the residuals and make sure that each individual is behaving like it should (constant variance, normality)
   ii. Some software have a test to test for equal covariance matrices between the groups, much like the F-test for equal variances in the two sample t-test problem (PROC DISCRIM)
   iii. Screen for outliers all throughout the processes above (both bullet points)
e. If everything looks good, exam the overall MANOVA F-tests
   i. If not significant then stop and we are done (Fail to reject)
      1. Wilks Lambda
      2. Pillai's Trace
      3. Hotelling Lawley Trace
      4. Roys Greatest Root
   ii. It is possible for some of the univariate analysis to have significant F-tests but we don't care because the overall tests protects us from making a type-I error

f. If the overall F-tests is significant then the fun begins
    i. Most straight forward thing to do is then to jump in and do one at a time analysis ANOVA F-tests and contrasts using our univariate approach we discussed in class
    ii. Penalize yourself for multiple tests

9. SAS Computer Programs

```
Proc GLM Data=Hemophil plot=diagnostics;
 Class Group;
 Model Activity Antigen = Group;
 Output Out=Errs R=Eact Eant;
 Means Group / CLM T Alpha=0.025;
 Means Group / CLDIFF T Alpha=0.025;
 Contrast 'Carrier vs Noncarrier' group 1-1;
 Manova H=_All_ / PrintE PrintH Canonical;
 ODS Output Cancoef=CanonicalCoefficients;
Run;
Quit;
```

10. Limitations of the technique / Things to look out for

Multiple test corrections:

- There are a bunch of ways you can do this. Simple may be the most preferred but may be too conservative depending on the number of factor levels and number of response variables
- For example suppose you had 4 response variables and the factor has 4 levels. For each response there are 6 different ways to compare the 4 levels
- 6comparisons times 4 response variables = 24 total comparisons
- P-value bonferronni cutoff is .05/24 = .0002
- The other option would be to conduct anova F tests for each response variable (bonf correction for 4 tests / or don't correct at all here) and for any response variable that is significant conduct the pairwise tests (bonf correction for 6 tests)
- At the end it all boils down to how much you want to penalize yourself.

In general, the more variables you have to analyze the messier MANOVA gets in terms of keeping everything straight and conducting straight forward analysis. Its easy to get lost in all of it so organization and a clear thought out plan is key. Examining everything one variable time

goes a long way in checking assumptions for multivariate normality, as if there are any issues for one variable there is an inherent issue with the assumption at the multivariate level.

- More than two levels it gets really busy
- More than three response variables and we can't even plot them together (Can't see higher than 3 dimensions)
- Again the idea is try to see if the multivariate normal assumptions with constant covariance matrix is appropriate. We should see linear relationships that follow ellipse patterns in 2d and ellipsoids in 3d
- Any nonlinear type relationships, transform to linear, if not, maybe consider throwing out the variables in question

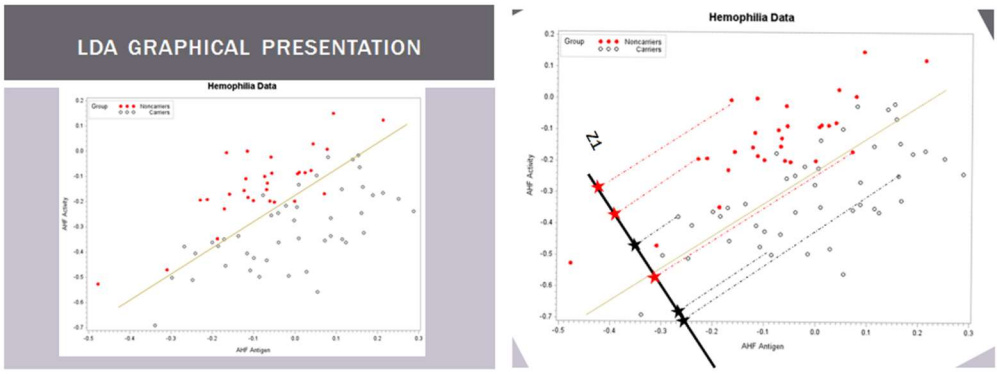
Section 2 Multivariate Techniques -Topic 2: LDA

1. Purpose/Questions of Interest/Research Questions

Used to classify an observation (multiple response values) into one of two or more groups/categories. Used for both prediction and as an indicator of the groups association with the response variables (aka the same type of question MANOVA is answering).

When we only have two groups, we can also use logistic regression to predict. When to use LDA or QDA or logistic regression?
All response variables are continuous as well as multivariate normal. If covariance matrices are assumed equal then LDA, if not, then QDA. Otherwise use logisitic regression.

For the two group case, the idea is to create a new variable Z, which is a linear combination of the original variables, that will separate out the groups well and can be used as a classification rule. Example: If Z>2 its classified as group A. If Z<2 its classified as group B.



2. Assumptions / Structure of the Data

See embedded slide deck from MANOVA as LDA is there as well.

The structure of the data and assumptions is the same as that of MANOVA, but we switch the rolls of the response and explanatory variable to answer the new question. So the factor levels (groups/categories) is considered the response variable, and the multiple variables that we observe measurements on (continuous measurements) are now considered the explanatory/predictor variables.

One difference in the assumptions is that MANOVA requires the assumption of equal covariance matrices among the groups. This is required for LDA, if this assumption is not appropriate, then QDA should be performed. SAS performs a test and does this automatically based on the outcome of the test.

3. Sampling Method

Random samples from the populations (categories) to be discriminated
Same as MANOVA.

4. Special Descriptive Statistics and/or Graphics

The main components of MANOVA still apply for model checking

5. What can we estimate?

Technically speaking, the estimation here is the coefficients for the linear combination of the new discriminating variable Z or multiple Z's if we have more than two groups.

6. Hypothesis Testing

We did not discuss in too much detail about hypothesis testing here, but PROC DISCRIM does provide the test for equal covariance matrices across the groups that could be helpful in the MANOVA analysis in the previous section. It is used in PROC DISCRIM to determine if LDA or QDA should be performed.

Null: The Covariance Matrix are the same for each factor level/category/group

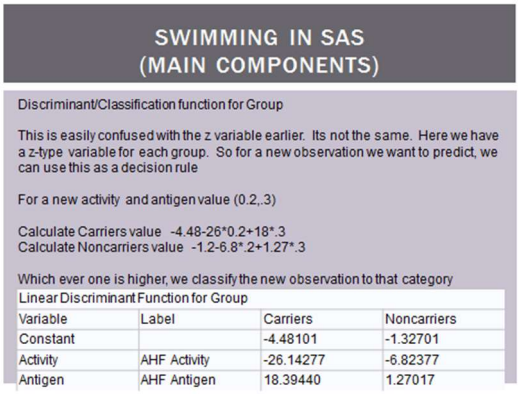
There are other tests, that SAS provide that test if the new discriminating variables are actually significant in terms of their ability to discriminate. This test turns out to be the same overall test for MANOVA when there are just two groups.

7. Relationship between other techniques / Other Info

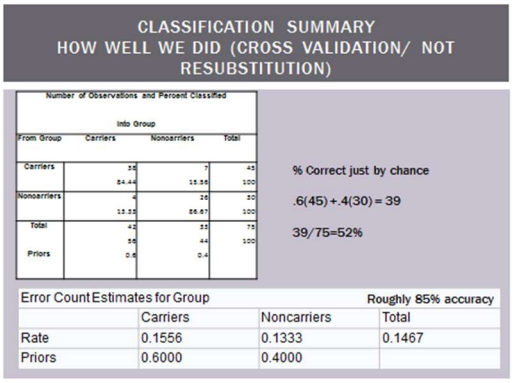
LDA is closely related to MANOVA, as the assumptions are the same. QDA should be used if the assumption of equal covariance matrices between the groups is not appropriate.

8. Analysis Overview

- Summary stats and graphics (scatterplots and 3d scatterplots) color coded by groups
- Plots should exhibit linear trends and have elliptical shape patterns to satisfy the multivariate normal assumption
- Examining residuals from univariate ANOVA models are another possible way to assess assumptions of normality
- Conduct test for equal covariance matrices across the groups to detmine if LDA or QDA should be used to determine the classification rules
- If Fail to reject: LDA If Reject: QDA
- Determine if prior probabilities will be used or that the groups are equally likely
- SAS provides the classification functions that can be used to predict a new set of values of the variable to then predict what group it belongs too. The user simply needs to calculate the linear combination using the coefficients provided for each group. The group with the largest value is determined as classification group for that observation.



- Report the performance of the classification procedure. Using the prior probabilities we can also calculate what we expect the performance would be just by random guessing



9. SAS Computer Programs

```
proc discrim data=Hemophil pool=test crossvalidate canonical testdata=test testout=a ;
class group;
var Activity Antigen;
priors "Carriers"=.6 "Noncarriers"=.4;
run;
```

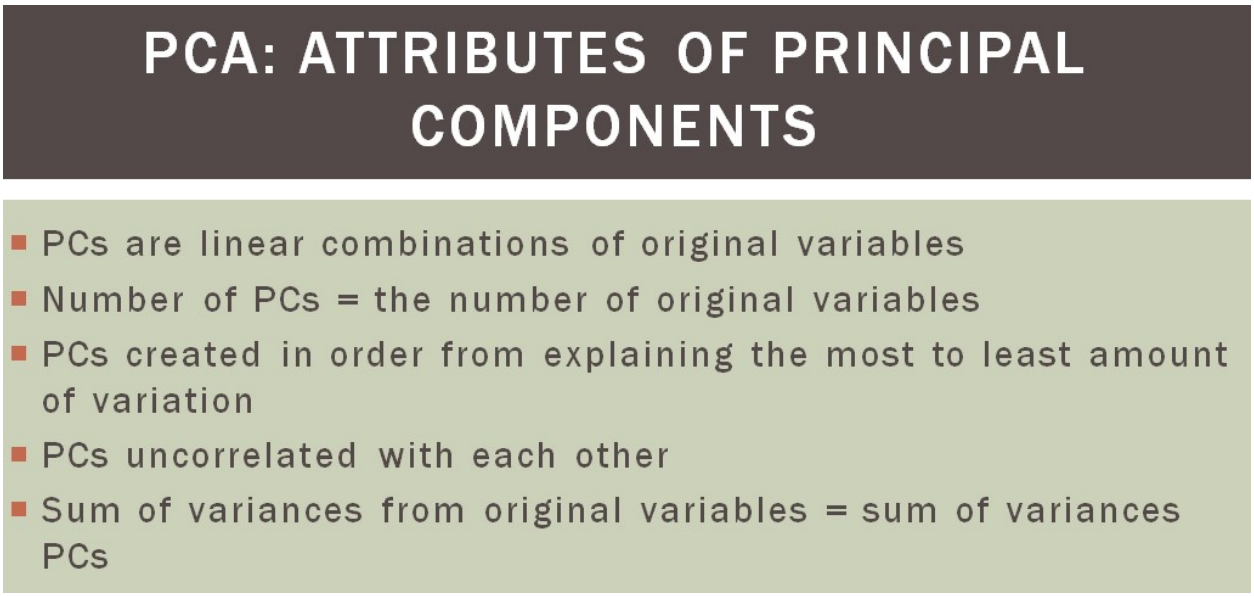
10. Limitations of the technique / Things to look out for

LDA and QDA only allow for continuous predictors and that predictors are assumed to be multivariate normal. We know from multiple regression that many predictor variables will often times be skewed (nonnormal) and will include categorical factors. When the assumptions of LDA and QDA are not met we can turn to other, more general, linear model type approaches such as logistic regression to handle scenarios that are outside the assumptions of LDA and QDA.

Section 2 Multivariate Techniques -Topic 3: PCA

1. Purpose/Questions of Interest/Research Questions

- PCA is typically used as an exploratory analysis and data reduction technique.
- Purpose: From a group of variables, PCA creates new uncorrelated variables. Information of these new variables can be used:
  - to understand the relationship among the original variables;
  - for other analyses such as multiple linear regression;
  - for exploratory analyses to graphically investigate clusters of observations or potential associations with other variables.
- Data Type: Continuous variables
- Sampling: Independent random sample



2. Assumptions / Structure of the Data

Double click this icon to activate embedded slide deck on this subject:

Session8PPT.PPTX

- **Level measurement**: all variables should be assessed on an interval or ratio level of measurement (continuous).
- **Random Sampling**
- **Linearity**: the relationship between all observed variables should be linear.
- **Bivariate normal distribution**: Each pair of observed variables should display a bivariate normal distribution (i.e. they should form an elliptical shape when plotted). Holistically, all of the variables must follow a multivariate normal distribution.

- **Sufficient sample size**: to obtain reliable results, the minimal number of participants (sample size) providing useable data should be 10 times the number of variables being analyzed.

3. More PCA information

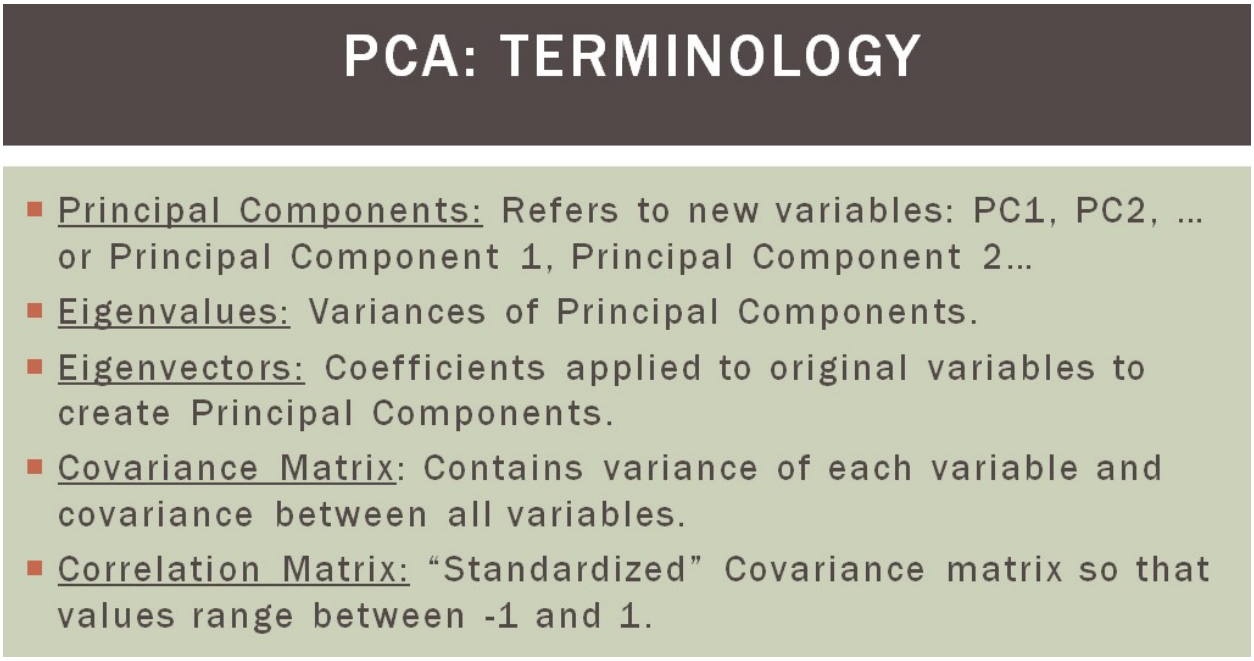
The first component extracted in a PCA accounts for a maximal amount of total variance in the observed variables. The second component will have two important characteristics; First, it will account for the maximal variance in the dataset that was **not accounted for in the first principal** component and second, it will be **uncorrelated with the first principal component**.

A principal component analysis proceeds in this manner with each new component accounting for progressively smaller amounts of variance. This is why only the first few components are usually retained and interpreted.

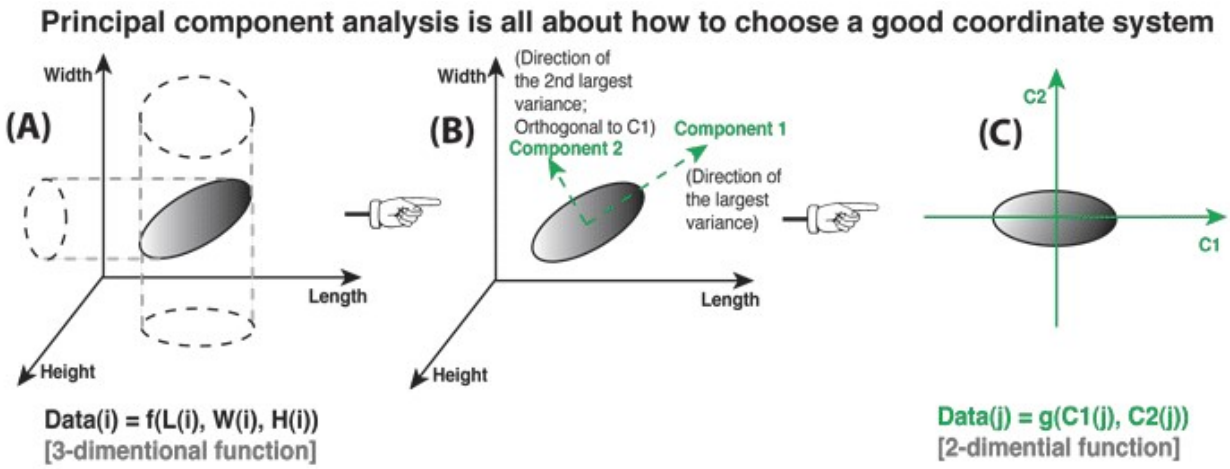
When the analysis is complete, the resulting components display varying degrees of correlation with the observed variable, but are completely uncorrelated with each other.

Another fun fact for PCA, is that when dealing with a large number of variables, PC's tend to be multivariate normally distributed even if the underlying variables deviate from normality. Central limit theorem comes into play here.

4. Special Descriptive Statistics and/or Graphics



5. What can we estimate?



6. General Analysis Overview

a. Identify reasons for conducting PCA
b. Summary statistics of original variables
c. Correlation analysis of original variables
d. Decide criteria for number of PCs to be chosen
e. Decide to run PCA on correlation or covariance matrix
f. Run analysis
g. Interpret results

Interpret Results

- Variance and proportion of variation for each PC

Eigenvalues of the Correlation Matrix

	Eigenvalue	Difference	Proportion	Cumulative
1	2.75861596	1.80893352	0.5517	0.5517
2	0.94968244	0.19597232	0.1899	0.7417
3	0.75371012	0.36713571	0.1507	0.8924
4	0.38657441	0.23515734	0.0773	0.9697
5	0.15141707		0.0303	1.0000

- Conclusion: The first principal component accounts for 55% of the zip code features' variability and the first three principal components account for almost 90%.

Interpret Results

- Coefficients (eigenvectors) of original variables.

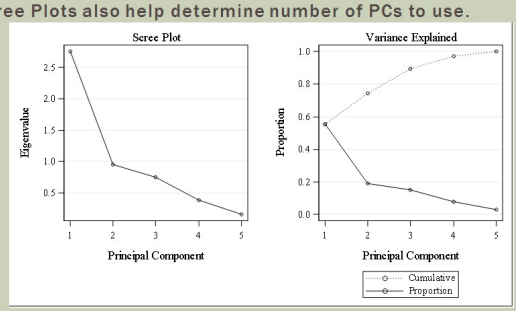
Eigenvectors

	Prin1	Prin2	Prin3	Prin4	Prin5
Fire	0.504823	0.030903	0.189071	0.829630	0.142006
Theft	0.309078	0.832612	0.240781	-.218340	-.324932
Age	0.398535	0.192335	-.816327	-.123544	0.350034
Income	-.505043	0.457245	0.142614	0.137146	0.704777
Race	0.485517	-.244411	0.468559	-.479552	0.504994

- The larger the absolute value the more a variable impacts a PC
- PC1: Fire, Income, Race
- PC2: Theft and Income

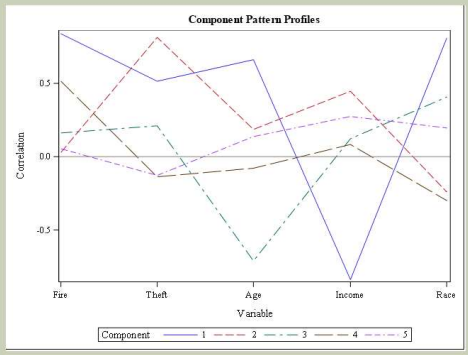
Interpret Results

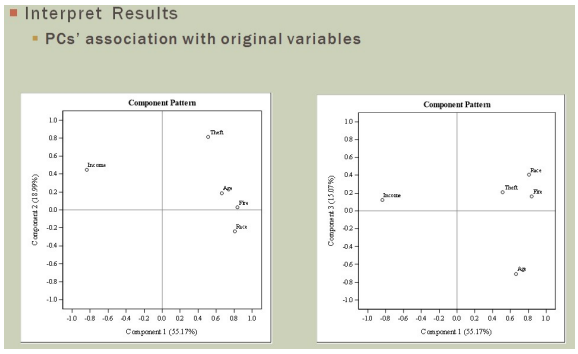
- Variance and proportion of variation for each PC
- Scree Plots also help determine number of PCs to use.

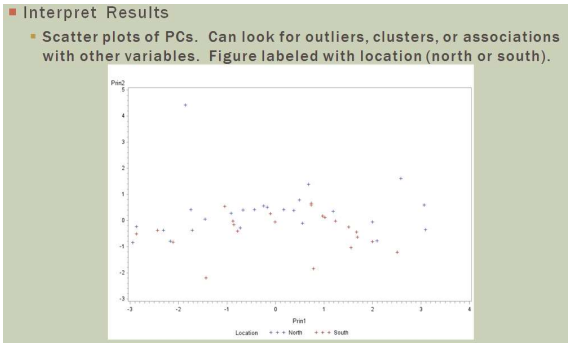


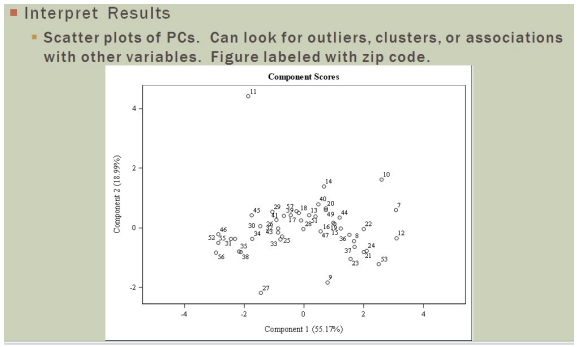
Interpret Results

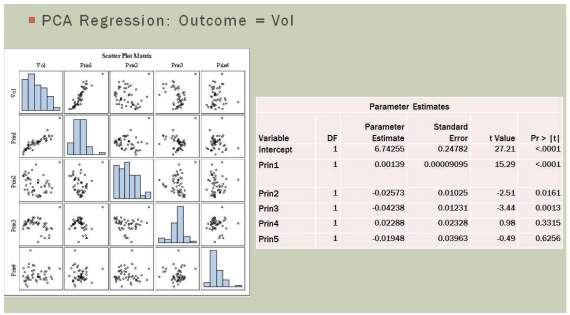
- PCs' association with original variables











7. Relationship between other techniques / Other Info

NOTE: **PCA is not factor analysis.** Factor analysis assumes that an underlying causal structure exists in the data. PCA makes no such assumption.

8. SAS Computer Programs

```
ods graphics on;
proc princomp plots=all data=q out=pca;
var Fire Theft Age Income Race ;
id Zip ;
run;
ods graphics off;
```

9. Limitations of the technique / Things to look out for

- Centering (covariance matrix) and Standardized (correlation matrix) variables: Done automatically in proc princomp but other software may require you to do so.
- Missing data: Observations not included in analysis.
- Outliers: Heavily influences results.
- Data reduction: Still need all original variables.
- PCA Regression: PCs with low variability can still be associated with outcome of interest. PCs can have non-linear associations with outcome and interactions with each other.
- Categorical variables: Can try creating binary variable (0 and 1s).
- Sample Size: Rule of thumb is to have 10 observations / 1 variable. (So for 3 variables, sample size should be at least 30)
- Normal distribution: Transforming original data may improve results.

- PCA most useful when have correlated data and large number of variables.