**SMU Data Science Program**

**Experimental Statistics II**

**Final Exam**

**To Be Completed Throughout the Course**

**By Benjamin Goodwin**

**Section 1 Modeling Continuous Responses**

1. Multiple Linear Regression
2. Feature Selection
3. Two Way ANOVA
4. Time Series
5. Repeated Measures Analysis
6. MANOVA (not required for final)

**Section 2 Modeling Categorical Responses**

1. Linear Discriminant Analysis
2. Logistic Regression
3. Classification Trees / Random Forest (not required for final)

**Section 3 Unsupervised Tools**

1. Principle Component Analysis
2. Clustering and Heatmap Visuals

Section 1: Modeling a Continuous Response

**Section 1: Modeling Continuous Responses**

**Topic 1: Multiple Linear Regression**

1. **Main Goal of the topic**

Used for 2 main purposes:

* Want to develop a statistical model to predict an outcome
* Want to know and describe the relationship between the outcome and an explanatory variable while possibly adjusting for other variables.

1. **Assumptions / Structure of the Data**

Data type the explanatory variables can take:

* Continuous
* Categorical

Key Assumptions (if applicable):

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

1. **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.
* ASE type plots

Diagnostic Statistics & Plots

* Raw Residuals = Observed – Predicted
* Standardized residuals= Z score, look for values more extreme than +/- 2 or 3
* Studentized resioduals = 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 software

Multicollinearity

* Variance Inflation Factor (VIF) – look out for values above 10
* Scatterplot matrices and correlation values. Heatmaps are good.

1. **Hypothesis Testing**

High Level F-test:

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

Lower Level t-test:

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).

1. **What are the pros and cons of this tool in regards to it ultimately achieving its main objective?**

CONS

* MLR (without any interaction terms) assumes a strict linear (addative) relationship between the response and explanatories. If the true relationships are more complicated, it is up to the modeler to add the complexity into the model through transformations, polynomials, interactions, etc.
* All the assumptions listed in 2 are required for the hypothesis tests to be valid. Fixing these issues can lead to less than optimal interpretations outside of log transformations.
* Can suffer from overfitting if too complex a model is proposed, or underfitting if not complex enough. Feature selection is helpful to assess this.

PROS

* Multicollinearity is not an issue for prediction
* Method will outperform other methods when assumptions are true and trend is appropriately modeled.
* Go-to method if interpretation is a key component of the research question.

1. **General Analysis Flow (For Completionists)**
2. Identify the question of interest (See #1)
3. Exploratory analysis (EDA)

* 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)

1. Analysis

* Fit full model and assess model assumptions through residual diagnostics. Fix if necessary.
* 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.

1. Reporting

* Provide the final regression model equation.
* Provide appropriate interpretation to regression coefficients that are significant and those you wish to discuss to answer the researchers questions.
* 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.

**Section 1: Modeling Continuous Responses**

**Topic 2: Feature selection tools**

1. **Main Goal of the topic**

**These tools are typically implemented in predictive models to help determine a candidate model with good bias/variance trade off. Most of the tools listed can be applied in logistic regression setting as well as multiple linear. Some tools will be mentioned later.**

**In statistics, stepwise regression includes regression models in which the choice of predictive variables are carried out by an automatic procedure.**

1. **Summary of feature selection procedures**

**Feature selection procedures include regression models in which the choice of predictive variables are carried out by an automatic procedure. In each step, a variable is considered for addition or subtraction from the set of explanatory variables based on some prespecified criterion.**

**In essence, feature selection procedures are a way to build a model by adding or removing predictor variables, usually a series of F-Tests or T-Tests. The variables to be added or removed are chosen based on the test statistics of the estimated coefficients. A key point to note on feature selection procedures is to keep a good birds eye view of whether the model makes sense or not.**

**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**

**In the backward selection procedure, you fit a full model containing all p predictors and then iteratively removes the least useful predictor, one-at-a-time. It is important to note that, to perform backward selection, the situation must be where there are more observations than variables because least squares regression can be performed when n is greater than p.**

**The basic flow of backward selection is as follows:**

* **Begin with all variables in the model**
* **Remove the variable with the largest p-value**
* **The new (p-1) variable model is t, and the variable with the largest p-value is removed**
* **Continue until a stopping rule is reached.**
* **This can typically happen when all remaining variables are significant.**

**STEPWISE**

**Stepwise regression is the collective term for backward and forward selection. To recap again:**

**Backward:**

* **Start with all variables in the model**
* **Remove the variable with the largest p-value**
* **The new (p-1) variable model is t, and the variable with the largest p-value is removed**
* **Continue until a stopping rule is reached.**
* **This can typically happen when all remaining variables are significant.**

**Forward:**

* **Start with no variables in the model**
* **Add one variable at a time as the model progresses**
* **If there are many predictors, use the “F-to-add” statistic is created using the same steps above. The variable with the highest “F-to-add” statistic is added to the model.**

**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.

**Variable Importance Ranking and “mtry”**

Metric derived from bagging and random forrest models (see topics later). Often used to provide a reduced set of features into other models outside of its original intended use.

**Section 1: Modeling Continuous Responses**

**Topic 3: TWO WAY ANOVA**

1. **Main Goal of the topic**

At a high level, the main goal of two way ANOVA is to estimate how the mean of a quantitative variable changes according to the levels of two categorical variables. We will use two-way ANOVA when we want to know how to independent variables, in combination, affect a dependent variable.

We use two-way ANOVA when we have collected data on a quantitative dependent variable at multiple levels of two categorical independent variables (the number two is important). The two-way ANOVA will use the F-test, which will compare the variance in each group to the overall variance in the dependent variable. It is also important to note that interactions are important here and should one be present, we will make inference on the interaction.

1. **Assumptions / Structure of the Data**

Like many other statistical tools, there are assumptions and it is crucial to meet these assumptions. A two-way ANOVA uses all of the normal assumptions for parametric tests of difference:

* Homogeneity of variance
  + The variation around the mean for each group being compared should be similar among all groups. If the data in question doesn’t meet this assumption, take a look a non-parametric alternatives, like the Kruskal-Wallis test.
* Independence of observations
  + The independent variables should not be dependent on one another. However, with categorical variables, this task becomes difficult, so typically you have to rely on good experimental design.
  + If the data fails this assumption, a strategy should include using a blocking variable or using a repeated-measures ANOVA.
* Normally-distributed dependent variable
  + The dependent variable should exhibit normally distributed behavior.
  + Should the assumption fail, try a transformation.
* No significant outliers
* The two independent variables should each consist of two or more categorical independent groups.

1. **Special Descriptive Statistics, performance metrics, and/or Graphics**

With two-way ANOVA, important to identify analysis situation.

* One factor of interest within a block
  + Randomized complete block design
* Two Factors
  + Both factors are of interest

Metric: F-tests are the deciding metrics for two-way ANOVA. There are a few notes on the F-test in this context.

* Overall F-test can be used to see if there are any differences.
* Additional F-tests can tell us something about each factor and if a nonadditive model is supported
* However! These tests don’t tell us which specific groups or combination of groups are different from each other. We will need to take the next steps and use different tests.

Metric: Contrasts. Once an overall F-test is conducted and found significant and if the factor of interest has more than two levels, we would like to know which factor levels are different.

Remember that once you start conducting multiple tests, adjust p-values using various procedures including:

* Bonferroni
* Tukey
* Dunnett

Plots: Mean profile plotting

Good for visually assessing if an interaction exists

* Plot standard errors (Line plot)
  + Objective: Gives us a sense of uncertainty in the mean estimates for each group
  + Cannot assess constant variance unless sample sizes are equal
* Plot standard error (Error bars)
  + Gives us a sense of the population variability
  + Can asses constant variance

Modeling notes: Fit a full saturated model with both factors and the interaction (nonadditive)

Diagnostic plots

* Residuals
  + If all of residuals are 0, and the interaction F-test is missing. This is due to not having any replicate observations are the interaction levels.
* Normality, Independence, Constant Variance
* Outliers (Don’t concern with leverage as these apply to more continuous explanatory variables)

Interactions are an important item of interest and have unique rules with two-way ANOVA.

If the interaction F-test is significant, we know the difference between levels of 1 factor, depends on the levels of the other factor.

* Note! In this situation, even if the “Main effects” are significant we do not need to explore them because of the interaction effect

Performance Metric: AIC calculates the best-fit model by finding the model that explains the largest amount of variation in the response variable while using the fewest parameters. Look for lowest AIC.

1. **Hypothesis Testing**

Steps for conducting a two-way ANOVA hypothesis test

Step 1: Hypotheses (three are tested at same time)

There is no difference in group means at any level of the first independent variable

There is no difference in group means at any level of the second independent variable.

The effect of one independent variable does not depend on the effect of the other independent variable (No interaction effect)

Note: Two-way ANOVA without interaction (additive model) tests only the first two of these hypotheses.

Step 2: Set significance level: Typically alpha=0.05

Step 3: Calculate F-statistic from table

Step 4: Determine P-value

Step 5: Reach a conclusion based on p-value and significance level

Step 6: Make a decision (Reject, or fail to reject the null hypothesis)

Step 7: Reiterate statistical conclusion, comment on scope, report p-value, confidence interval and interpretation and next steps.

1. **What are the pros and cons of this tool in regards to it ultimately achieving its main objective?**

CONS: We are somewhat restricted by the “two-way” nature of two-way ANOVA. It is worth mentioning that with enough sample size and all of the factor level combinations, we could include all of the two way ANOVA and three way interaction and conduct a similar analysis for three way ANOVA, similar to two-way ANOVA. The issue is that we can’t always observe every combination due to the study design and the logistics.

Another con is that there is the requirement of post-ANOVA testing to determine which groups are actually different, which is different from the ANOVA conclusion of “yes/no there is a difference.”

PROS: In general, it is an improved technique over the t and z tests. We can also conduct analysis of various factors at a time. It is also an economical method of parametric testing. To me, a huge pro is that the flow and procedure is straight forward and the decisions are easy to make regarding conclusions.

1. **General Analysis Flow (For Completionists. Good idea for yourself, not required for Final)**

* **Determine the applicable situation**
  + **One Factor of interest with a block**
    - **Randomized complete block design**
  + **Two Factors**
    - **Both Factors are of interest**
* **Plot data (visualize through mean profile plotting)**
* **Fit full saturated model with both factors and the interaction (nonadditive) term**
* **Diagnostics:**
  + **Residuals**
  + **Normality, Independence, Constant Variance**
  + **Outliers (Don’t concern with leverage as these apply to more continuous explanatory variables)**
* **Testing**
  + **High level (ANOVA)**
  + **Contrasts**
    - **If nonadditive (interaction) is significant then we report difference of groups by factor levels**
    - **If not significant, we compare contrasts of the individual factor levels (very simple -> very complex)**
    - **Both situations call for multiple testing corrections**

**Section 1: Modeling Continuous Responses**

**Topic 4: Time Series**

1. **Main Goal of the topic**

When data is collected over time, the independence assumption of regression is violated, this is problematic because we can no longer draw inference on a model generated from these data. This results in estimates and standard errors being biased when modeling and conducting hypothesis tests. Furthermore, general methods treat data as if there is more information than is actually present. This can sometimes be remedied by adjusting standard errors, transforming the data so that it behaves like independent data (read: regression), or finally generalized least squares combined with proper estimation.

Seems like a lot just to remedy an assumption violation, hence time series.

A time series is a series of values of a quantity obtained at successive times, often with equal intervals between them.

We can also use time series to evaluate if the mean of the time series has changed after some point in time.

Additionally, the tool can determine if explanatory variables are contributing to what’s going on with the dependent variable.

1. **Assumptions / Structure of the Data**

Key assumptions:

* Normality of residuals
* Constant variance

Correlated observations are tougher on hypothesis testing, it is however very useful to predict future valued based on previous information.

Time series model assumes you have data that is equally spaced in time.

Different model types; ARIMA, AR, MA

* Assumptions are: Constant mean, constant variance, constant autocorrelations
* Assumptions must be met for ARIMA, AR,MA

Structure of stationary data:

* Lag1 (1st difference) – Response is now the change from time x to its previous time point- wandering behavior
* Sometimes you need to take the difference again
* Seasonal differencing
* Log transform
* Model deterministic behavior out
* Regress explanatory variables

ARIMA Models

* AR (Autoregressive) integrated MA (Moving Average)
  + A representation of a type of random process
    - Used to describe certain time-varying processes in nature, economics, etc.
* MA (Moving Average)
  + MA models have different properties and exhibit different types of time series behavior, autocorrelations, etc. It is related to the idea that a random error that happens at a given point in time remains (proportion of it) on a finite number of subsequent time points.
* Integrated stands for differencing the data at first to generate a stationary process.

Always examine the residuals from the model to see if any autocorrelation still exists!

Autocorrelation plots, Partial ACP, Tests involving AR1 (Runs test or Durbin Watson Test)

Rules of thumb to help identify the features

1. **Special Descriptive Statistics, performance metrics, and/or Graphics**

* The autocorrelation plot
* ACF/PACF plots
* Residuals to check for autocorrelation
* Stationary data
  + Constant mean- the time series may show cyclical behavior or general up and down behavior but it clearly is centered around a mean value.
    - Subpopulations of X\_t have the same mean for each t.
  + Constant variance-typically violated when low observations of the time series are less variable than higher observations
    - Subpopulations of X for a given time have a finite and constant variance for all of t.
  + Constant autocorrelations- (correlations of lags don’t depend on position in time series)
    - This allows the estimation of the ACF
    - Wavelets and G-stationary models are alternatives when not true
    - The covariance of X\_t1 and X\_t2 depends only on t\_2-t\_1. That is, the covariance between data points is dependent only on how far apart they are, not where they are.
* Rules of thumb for ACF and PACF plots

|  |  |  |
| --- | --- | --- |
| Conditional Mean Model | ACF | PACF |
| AR(p) | Tails off gradually | Cuts off after p lags |
| MA(q) | Cuts off after q lags | Tails off gradually |
| ARMA(p,q) | Tails off gradually | Tails off gradually |

Trends must be handled appropriately

* Determine if data is stationary or not
  + If yes, build model (ARR,MA, ARMA, ARIMA)
  + If no, model needs to be made stationary
    - Take a first difference ARIMA
    - Include time as a covariate

1. **Hypothesis Testing**

Summary: Time series hypothesis testing talks about how we identify whether different time periods have significantly different observations. The difference between a time series hypothesis test and a regular t-test has to do with serial dependency (not to mention trend and seasonality), so we can’t separate the observations into groups.

We are left with the residuals of the time series, and if the residuals fulfill the normality assumption, we are free to do the hypothesis test. We can create a null and alternative hypothesis as we usually do.

Step 1:

H0: There is no difference between the time periods (all periods are equal)

HA: There is a difference between the time periods. (at least one period differs)

Step 2:

Set significance level: Typically alpha =0.05

Step 3:

Calculate test statistic

We can then use a two sample t-test to know whether the mean of each group is significantly than the other group(s). The main question here is, “Is the signal stronger than the noise?” This is accomplished by calculating the mean difference in groups and then the pooled standard deviation.

Step 4:

Obtain P-value

Step 5: Reach a conclusion based on p-value and significance level

Step 6: Make a decision (Reject, or fail to reject the null hypothesis)

Step 7: Reiterate statistical conclusion, comment on scope, report p-value, confidence interval and interpretation and next steps.

1. **What are the pros and cons of this tool in regards to it ultimately achieving its main objective?**

CONS: Time series analysis takes thought and time to learn and can be slightly difficult to interpret. There are numerous factors to consider when interpreting a time series, such as autocorrelation patterns, seasonality, and stationarity. Because of these items a number of models were created to describe these series all require training to interpret and make inference on.

PROS: Time series is a beneficial tool if you have data that does not meet the independence assumption of linear regression.

The most major benefit of time series analysis is that it can be the basis to forecast data. By the very definition of time series it can uncover patterns in the data to be used to predict future data points.

Time series can also help an analyst better understand data, when the person doing an analysis understands the type of data they are dealing with, time series can be super beneficial for the right dataset.

1. **General Analysis Flow (For Completionists. Good idea for yourself, not required for Final)**

**General Approach to Time Series**

* **Regression model:**
  + **If you have explanatory variables, fit the data using regression techniques and obtain the residuals (observed-predicted)**
  + **Assess if there is any serial correlation**
    - **Use Autocorrelation plots, Partial ACP, Tests involving AR1** 
      * **Runs test or Durbin-Watson test**
    - **Rules of thumb to help identify the features**
* **If sufficient evidence exists, we refit the model with a time series model using autoregressive processes (auto reg w/ nlag option) or possibly multiple models from AR to MA and ARIMA.**
* **Re-examine the residuals to see if serial correlation is removed by looking at the new ACF and PACF plots (Don’t forget normality)**
  + **If everything looks good, we can’t trust the t-test and p-values provided for specific hypothesis test**
  + **Forecasts can be generated to illustrate future predictions**
  + **The last two steps can be automated similar to feature selection in regression.**

**General Workflow of ARIMA**

* **Plot the series and explanatory variables**
* **Make a decision on if it is stationary or not**
  + **If not:**
    - **Try differencing the data**
    - **Obtain seasonally adjusted data**
    - **Model other deterministic behavior through explanatory variables via OLS**
* **Examine the residuals from this model to see if any autocorrelation still exists**
  + **If so, use rule of thumbs or AIC selection procedure to pick an appropriate ARIMA model**
* **After accounting for the serial correlation, re-examine residuals to make sure they are behaving uncorrelated and are normally distributed**
* **Repeat steps above as necessary**
* **Proceed to tests and forecasting**

**Section 1: Modeling Continuous Responses**

**Topic 5: Repeated Measures**

1. **Main Goal of the topic**

In general a repeated-measure design is a research design in which subjects are measured two or more times on the dependent variable. Instead of using different subjects for each level of the treatment, the subjects are given more than one treatment and are measured after each.

This definition means that each subject will be its own control. In repeated-measures analysis, scores for the same subject are dependent, so the score for different subjects are independent.

1. **Assumptions / Structure of the Data**

* The dependent variable is measured on interval or ratio scale (dependent variable is continuous)
* The sample was randomly selected from the population. The cases represent a random sample from the population, and there is no dependency in the scores between participants.
* The dependent variable is normally distributed in the population for each level of the within-subjects factor
* The population variances for the test occasions are equal. The population correlation coefficients between pairs of test occasion scores are equal. The population variance of difference scores computed between any two levels of a within-subjects factor is the same value regardless of which two levels are chosen. This assumption is sometimes referred to as the sphericity assumption or as the homogeneity-of-variance-of-differences assumption. The sphericity assumption is meaningful only if there are more than two levels of a within-subjects factor.

There are many different types of repeated measures analysis:

* Longitudinal studies
  + Repeated measurements on the same variable for the same subjects at different times
* Crossover Experiments
  + Each subject receives more than one treatment
    - First measure is response after one treatment
    - Switched (“crossed over”) to the second treatment
  + Subjects randomized to first treatment then switched to second
    - Random assignment to one of several treatment orders is ideal
    - Order can affect response
* Split-Plot Experiments over Time
  + Includes features of both longitudinal and crossover experiments
  + First randomization allocated experimental units to one of several treatments
  + Second randomization determines presentation order of second treatments
    - All subjects receive all levels of second treatment
    - Subjects receive only one level of first treatment

1. **Special Descriptive Statistics, performance metrics, and/or Graphics**

Occasionally in a repeated measures experiment, a multivariate response will occur. This includes more than one response variable for each experimental unit (or sampling unit)

Typically, multivariate methods are conceptually easier and more powerful for analyzing repeated measure over time.

Univariate Analysis:

* Summarize the multivariate repeated measure
  + Typical summaries include
    - Profile average, final response, maximum, minimum, treatment difference, time at which maximum occurs, estimated slope of a regression of the response on time

Multivariate Analysis:

* Provides a way of summarizing measures of uncertainty from several univariate analyses
* Adjusts for covariates
* Uses techniques that parallel single-response techniques
* There are extensions of:
  + One-sample and two-sample t-tests
  + Analysis of variance
  + Regression

One-Way ANOVA vs. Repeated Measures ANOVA are different entirely.

If we have repeated measure, we need to analyze data as repeated measures, not treating the repeated measures as if they are separate levels of a factor.

1. **Hypothesis Testing**

Note: Repeated Measures studies are multivariate! We will test H0 by determining whether the means on all k-1 of these difference variables are simultaneously zero

Step 1:

H0: There is no difference between any/all groups

HA: There is a difference between any/all groups (at least one differs)

Step 2:

Set significance level: Typically alpha =0.05

Step 3:

Compute covariance and T^2 and convert to an F-statistic with (k-1) and (n-k+1) DOF.

Step 4:

Obtain P-value from F-statistic

Step 5: Reach a conclusion based on p-value and significance level

Step 6: Make a decision (Reject, or fail to reject the null hypothesis)

Step 7: Reiterate statistical conclusion, comment on scope, report p-value, confidence interval and interpretation and next steps.

1. **What are the pros and cons of this tool in regards to it ultimately achieving its main objective?**

CONS: Since in repeated-measure designs the same group of participants are measured on multiple occasions, therefore sometimes the order of measurements have a differential effect on participants responses as some experiments involve repeated exposure to the same task.

* Practice effort or learning effect-as participants complete the measures after each condition, they may get better practice, or they could get bored or tired. As a result, the participants change as they are repeatedly tested.
* Latency effect refers to a situation in which the effect of a treatment is not evident until a subsequent level of treatment is introduced. A latency effect may predispose a researcher to erroneously contend that the administered treatment had little to no effect on the monitored behavior when, in reality, the effect of the treatment was not evidenced until an additional condition had been implemented.
* Carry-over effect refers to the influence of a previous level of treatment on the observed behavior in a subsequent level of the same treatment condition.

Combined these effects tend to skew results by influencing the responses of participants.

PROS: Elegant solution for specific types of study designs. Can treat differences between adjacent repeated measure as response variables. Can also estimate covariance matrix from differences. Multivariate repeated measures are usually preferable because sphericity assumption is difficult to test and to hold

1. **General Analysis Flow (For Completionists. Good idea for yourself, not required for Final)**

Section 2: Modeling a Categorical Response

**Section 2: Modeling Continuous Responses**

**Topic 1: Logistic Regression**

1. **Main Goal of the topic**

I think it is best to begin the topic of logistic regression by explaining why MLR isn’t a catch all for statistics questions. If you think of a categorical variable in MLR, you must convert the categories to numeric values, and that order matters and coding of categories encoded to 0,1,2 is different from 0,2,5 and a MLR model will predict probabilities outside of a binary response range. Enter logistic regression, and assuming our response is binary we can use a general regression approach on a transformed version of the predicted probabilities.

In summary: Logistic regression is used to describe data and to explain the relationship between one dependent binary variable and one or more nominal, ordinal, interval or ratio-level independent variables.

1. **Assumptions / Structure of the Data**

It is interesting to note that logistic regression does not require a linear relationship between the dependent and independent variables and differs from MLR in that the residuals do not need to be normally distributed and that homoscedasticity is not required, and finally that the dependent variable in logistic regression is not measured on an interval or ratio scale

Assumptions:

* Binary logistic regression requires the dependent variable to be binary and ordinal logistic regression requires the dependent variable to be ordinal.
* Logistic regression requires that the observations be independent of each other.
* Logistic regression assumes linearity of independent variables and log odds.
  + It requires that the independent variables are linearly related to the log odds.
* And a general rule of thumb states that logistic regression requires fairly large data sets.

Logistic regression can be binomial, ordinal or multinomial. Binomial or binary logistic regression deals with situations in which the observed outcome for a dependent variable can have only two possible types, “0” and “1” (which could for example represent “pass” or “fail”). Multinomial logistic regression deals with situations where the outcome can have three or more possible types, and ordinal logistic regression deals with dependent variables that are ordered.

Data may be continuous or categorical.

1. **Special Descriptive Statistics, performance metrics, and/or Graphics**

Unlike regular old regression, logistic regression is used for predicting dependent variables that take membership in one of a limited number of categories rather than a continuous outcome.

Logistic regression calculates the odds of the event happening for different levels of each independent variable, and then takes its logarithm to create a continuous criterion as a transformed version of the dependent variable. The logarithm of the odds is the logit of the probability.

Logistic regression measures the relationship between the categorical dependent variable and one or more independent variables by estimating probabilities using a logistic function, which is the cumulative distribution function of logistic distribution.

* Key formula: Log(odds) =
  + This formula can be used to interpret how the explanatory variables are associated with the response.

Odds ratio is a key statistic

Try to avoid interactions in logistic regression

Performance metrics:

* + If you have categorical predictors
    - Deviance and Pearson Goodness of Fit Tests
    - Comparing model to a more complex one (forward selection)
    - LASSO
    - Hosmer-Lemeshow test
  + If you have at least one or more continuous predictors
    - Comparing model to a more complex one (Forward selection)
    - LASSO
    - Hosmer-Lemeshow test
  + Tests datasets and Cross Validation

1. **Hypothesis Testing**

Step 1:

H0: B\_0=B\_1=B\_2….=0

HA: At least one B\_i differs

Step 2:

Set significance level: Typically alpha =0.05

Step 3:

Compute chi-square test statistic

Step 4:

Obtain P-value from chi-square

Step 5: Reach a conclusion based on p-value and significance level

Step 6: Make a decision (Reject, or fail to reject the null hypothesis)

Step 7: Reiterate statistical conclusion, comment on scope, report p-value, confidence interval and interpretation and next steps.

1. **What are the pros and cons of this tool in regards to it ultimately achieving its main objective?**

CONS: If the number of observations is lesser than the number of features, Logistic Regression should not be used. Otherwise, it may overfit. Logistic Regression constructs linear boundaries. A major limitation of Logistic Regression is the assumption of linearity between the dependent variable and the independent variables. It can only be used to predict discrete functions, so it is bound to the discrete number set.

Non-linear problems can’t be solved with logistic regression.

Multicollinearity can be an issue.

PROS: Logistic regression is fairly easy to implement and interpret (aside from a few corner cases). It makes no assumptions about the distribution of classes in feature space. It can easily extend to multiple classes, and it can quickly classify unknown data.

Logistic regression also had good accuracy for many simple data sets and it performs well when the dataset is linearly separable

1. **General Analysis Flow (For Completionists. Good idea for yourself, not required for Final)**

**Logistic regression is multiple regression for a dichotomous outcome**

* **Start by doing some EDA**
  + **Summary tables of the response versus categories**
  + **Summarize continuous variables by response status**
  + **If multiple continuous variables exist, pairwise scatter plots but color coded or labeled by response status are helpful**
* **Model Fitting**
  + **Same approach as MLR**
    - **Feature selection**
    - **Common sense/Manual iteration**
    - **Consider adding complexity through interactions, transformations, etc.**
  + **Assess goodness of fit / Test Lack of Fit / Assessing if the model is decent**
    - **Regardless of predictors: Residual diagnostics**
* **Provide estimates of the coefficients in the most interpretable way highlighting the statistically significant ones and their interpretation if understanding the explanatory variables relationship is important.**
* **Provide prediction performance through ROC curves and confusion matrices if prediction is important**

**Section 2: Modeling Continuous Responses**

**Topic 2: Linear and Quadratic Discriminant Analysis**

1. **Main Goal of the topic**

LDA/QDA have a similar setup as MANOVA but answer a different question. The main goals of LDA/QDA are:

* Prediction/Classification
  + Given a new observation, can I predict based on that new observation
  + Response variable is now the factor levels
* Understanding
  + Is there a simpler way to understand what variables contribute to the differences between groups.
  + LDA can help get information to answer this question.
* Between the two techniques LDA is used when a linear boundary is required between classifiers and QDA is used to find a non-linear boundary between classifiers.
* These are data reduction techniques

1. **Assumptions / Structure of the Data**

Assumptions of LDA:

* Everything from MANOVA plus:
* LDA assumes a normally distributed data
* Class specific mean vector
* LDA assumes a common covariance matrix. So, a covariance matrix that is common to all classes in a data set.

Assumptions of QDA:

* Everything from LDA plus:
* Observations of each class is drawn from a normal distribution
* QDA assumes that each class has its own covariance matrix

When the number of predictors is large the number of parameters we have to estimate with QDA becomes very large because we have to estimate a sperate covariance matrix for each class.

The higher the dimension of the data set the more parameters we have to estimate. This can lead to high variance and so we have to be careful when using QDA.

LDA is less flexible than QDA because there are fewer parameters to estimate.

1. **Special Descriptive Statistics, performance metrics, and/or Graphics**

Multivariate analysis gets you decent type-I error control and power.

* Type-I error: Probability of false positive
* Power: Probability rejecting when you should FTR

Its difficult to talk about LDA and QDA without mentioning MANOVA. In MANOVA we are interested in F-Statistics and the follow are MANOVA derived F-Tests

* Wilks Lambda
* Pillai’s Trace
* Hotelling Lawley Trace
* Roys Greatest Root

Each of these approaches reduce the ratio of two covariance matrices down to a single value.

Don’t forget to penalize for multiple testing

Additionally don’t forget that MANOVA cannot be applied in all settings.

A good thing to also note is that the discriminant scores are determined by the eigen values and eigenvectors of the covariance matrices from the MANOVA model (between and within, H and E, model and residual)

The scores themselves attempt to maximize the separation between groups, while minimizing within groups

For 2 group classification. We can use the discriminant score to come up with a decision rule to classify a new observation.

Take a look a frequency tables.

1. **Hypothesis Testing**

Step 1:

H0: Mu\_0=Mu\_1=Mu\_i

HA: At least one M\_i differs

Step 2:

Set significance level: Typically alpha =0.05

Step 3:

Compute f-test statistics

Step 4:

Obtain P-value from f-test

Step 5: Reach a conclusion based on p-value and significance level

Step 6: Make a decision (Reject, or fail to reject the null hypothesis)

Step 7: Reiterate statistical conclusion, comment on scope, report p-value, confidence interval and interpretation and next steps.

1. **What are the pros and cons of this tool in regards to it ultimately achieving its main objective?**

CONS: LDA requires normally distributed data on predictors and is sometimes not a great method when there are few categorical variables. QDA features a more flexible decision boundary, can quickly classify, and is easy to implement and can sometimes outperform LDA.

PROS: It is a simple, fast and portable modeling tool. The training time of both LDA and QDA can be lengthy and the required matrix operations are heavy on the computational side.

**Section 2: Modeling Continuous Responses**

**Topic 3: Decision Trees and Random Forest**

1. **Main Goal of the topic**
2. **Assumptions / Structure of the Data**
3. **Special Descriptive Statistics, performance metrics, and/or Graphics**
4. **Hypothesis Testing**
5. **What are the pros and cons of this tool in regards to it ultimately achieving its main objective?**

CONS:

PROS:

1. **General Analysis Flow (For Completionists. Good idea for yourself, not required for Final)**

Section 3: Unsupervised Tools

**Section 3: Unsupervised Tools**

**Topic 1: Principle Component Analysis**

1. **Main Goal of the topic**

* Perform data reduction in order to describe relationships among variables
* PCA is typically used an exploratory analysis and data reduction technique (aka an unsupervised 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 and classification

1. **Assumptions / Structure of the Data**

* Continuous data is required to perform PCA
* The data must also come from a independent random sample
* This is loosely enforced in practice
* PCs are linear combinations of original variables
* Number of new PC variables = the number of original
* PCs created in order from explaining the most to least amount of variability
* PCs are uncorrelated with each other
* Sum of variances from original variables = sum of the variances of the PCs

1. **Special Descriptive Statistics and/or Graphics**

* Scree plots are useful for determining the number of PCs to choose
  + Pay attention to amount of variability explained
* Biplots are useful for observing the direction and magnitude of PCs against data
* Need to choose whether to run PCA on correlation or covariance matrix
  + - You should use the correlation matrix when variables do not have the same units and would be heavily influenced by variables with large variance
* Useful terminology for PCA
  + Principal components: Refers to new variables: PC1, PC2, … or Principal Component 1, Principal Component 2
  + Eigenvalues: Variances of Principal Components
  + Eigenvectors: Coefficients applied to original variables to create Principal Components.
  + Covariance Matrix: Contains variance of each variable and covariance between all variables.
  + Correlation Matrix: “Standardized” covariance matrix so that values range between -1 and 1.

1. **How can PCA be used inside of a predictive model setting?**

If you are in a classification setting (predicting a categorical response) and there is a decent set of continuous predictors to work with then you can apply the following steps to use PCA in a predictive modeling setting.

* Conduct the PCA on the continuous set
* Plot the first few PC’s (or a number based on the scree plot) but color code the points by the response variable you hope to classify
* If the colors separate out, then the predictive model you build later (LDA, logistic, random forest) typically will perform well and vice versa.

A few notes on PCA with LDA

* Sometimes LDA can’t run well (or at all) if the sample size is small relative to the number of variables (The so called, “small n, large: situation)
* One solution is that we can run PCA first to reduce the number of variable down so that our “n” is much bigger than our “p”
* Again just like PCA regression, we can decide to use the first “x” number of principal components and use them to classify/predict a categorical response
  + This solution can work really well in some cases and not so well in others if the sample size is not an issues, then running LDA on the original variables is preferred as the, PC’s are being transformed with no classification rules in mind.

1. **What are the pro’s and con’s of this tool in regards to it ultimately achieving its main objective?**

CONS: Providing interpretation on PC’s that can be hard to explain. No guarantee that the PC’s will correlate better than the original predictors. Like a multivariate techniques, sample size needs to scale well with respect to the number of variables PCA is applied to

PROS: PCA can be used to explore the relationship among the original variables, and it can also be used to simplify a regression or prediction problem by reducing the number of predictors to work with up front and multicollinearity issues are resolved. PCA can also be used to determine if your predictors will actually be effective in predicting the categories up front.

**Section 3: Unsupervised Tools**

**Topic 2: Clustering and heatmap visuals**

1. **Main Goal of the topic**

* Goal: Obtain clusters with smaller within-cluster variability than between-cluster variability
* Clustering methods do no use any knowledge about how the observations are categorized at all (“completely unsupervised”)
* Two general strategies
  + Hierarchical clustering
    - Agglomerative clustering
      * Each feature beings in its own cluster
      * “Closest” features are merged until suitable number of clusters results.
      * After each agglomeration, all distances are recalculated
      * Algorithms include: Complete linkage, single linkage, average linkage, centroid clustering, Ward’s clustering
    - Divisive clustering
      * All features begin in a single cluster
      * At subsequent steps, “loosest” cluster is split in two
      * Huge number of ways to split initial cluster:
        + (2^(g-1)-1)
  + Partitioning
    - Data split into a specified number of nonoverlapping clusters
      * K-means (medioids)
        + Find K mean points in p dimensions
        + Take each of observed points and measure distance of observed points to its mean points
        + Move points around cluster centers until it reaches minimum variability of each observed points to cluster centers
      * Self-organizing maps (SOM)
      * Partitioning around medioids (PAM)
* Heatmaps are a plot of raw or standardized continuous variables where (each row is an observation or variables) each column is the other (variables/observations)
  + Plot is a grid of square or rectangles color coded by the value of the underlying dataset
  + Hierarchical clustering can be applied separately to both the rows and columns to see how the data clusters naturally

1. **Assumptions / Structure of the Data**

* Picking number of clusters is an important step
  + Typically done by plotting one or several of the internal values across a wide range of possible cluster numbers
    - There is never a hard and fast rule for picking the number
* No training set
* No knowledge of cluster type or (often) number of clusters
* No guarantee that optimal solution has been reached
* Prerequisites:
  + Selection of samples
  + Selection of features to use in clustering
  + Choice of algorithm
* K-means clustering typically produces tighter clusters than does hierarchical clustering
* Notions of distance are important
  + Pairwise
    - Features are viewed as two k- or p-vectors in some space
    - Compute distance within samples (pairwise)
  + Sampling
    - Feature profiles are two different samples generated from underlying probability distribution functions for measure of interest
  + Combination

1. **Special Descriptive Statistics and/or Graphics**

* Clusters tend to be associated with known groups
* Practicalities
  + Choice of scale is important
  + Choice of distance measures has less effect on result than choice of algorithm
  + There are measures of cluster validity to determine which clustering method gives best results according to a given criterion.
* Coefficients
  + Divisive coefficient
    - D(i) = diameter of the last cluster to which observation I belongs, divided by the diameter of the whole data set
    - The dc = average of all 1-d(i)
    - Should not be used to compare data sets of very different sizes
  + Agglomerative coefficient
    - M(i) = the dissimilarity of each cluster I to the first cluster it is merged with, divided by the dissimilarity of the merger in the final step of the algorithm
    - The ac = average of all 1-m(i)
* Heatmaps are an obvious visual extension of clustering
  + Heatmaps can help identify variability in data whether its technical artifacts (blocking type variables) or its real biological/meaningful variability

1. **Hypothesis Testing**

In terms of using statistical significance tests to validate cluster analysis results is almost always going to result in the null being rejected. We cannot test for difference in distributions for groups that were defined using the same data. This could be considered selective testing, and this applies for testing the entire model, however listed below are different methods for assessing cluster validity.

We have different methods of assessing cluster validity

Internal vs External

* External
  + Known class or groups
  + Can use Chi-square tests
* Internal
  + RMSSTD-small
  + SPRSQ-high
  + Cubic Clustering Criterion
  + Pseudo t2
  + Silhouette Statistic
  + Dunn’s Index
* In summary for assessing cluster validity
  + Some criteria suitable for partitioning and hierarchical methods
  + Pseudo t2 index only suitable for hierarchical methods
  + Different methods for ascertaining cluster validity
    - Should make sense statistically and in context of data

1. **What are the pros and cons of this tool in regards to it ultimately achieving its main objective?**

CONS: Hierarchical clustering may produce potentially unbalanced trees and/or clusters consisting of single observations. Hierarchical clustering may result in pattern of big clusters with most of data and smaller clusters surrounding it.

PROS: Clustering can help big data through heat map visualizations. Hierarchical clustering can be applied separately to both the rows and the columns to visualize how the data clusters naturally. Clustering helps identify variability in your data whether its technical artifacts (blocking type variables) or its real biological/meaningful variability. Heatmaps can also help identify outliers. They can also help identify if you are going to find statistical differences between groups and can help identify variables that are highly correlated that may be of meaningful importance.