**SMU Data Science Program**

**Experimental Statistics II**

**Final Exam**

**To Be Completed Throughout the Course**

**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.**

1. **Summary of feature selection procedures**

**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.

**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**
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 1: Modeling Continuous Responses**

**Topic 4: Time Series**

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 1: Modeling Continuous Responses**

**Topic 5: Repeated Measures**

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 2: Modeling a Categorical Response

**Section 2: Modeling Continuous Responses**

**Topic 1: Logistic Regression**

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 2: Modeling Continuous Responses**

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

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 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**
2. **Assumptions / Structure of the Data**
3. **Special Descriptive Statistics and/or Graphics**
4. **How can PCA be used inside of a predictive model setting?**
5. **What are the pro’s and con’s of this tool in regards to it ultimately achieving its main objective?**

CONS:

PROS:

**Section 3: Unsupervised Tools**

**Topic 2: Clustering and heatmap visuals**

1. **Main Goal of the topic**
2. **Assumptions / Structure of the Data**
3. **Special Descriptive Statistics 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: