**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

**Section 2 Modeling Categorical Responses**

1. Linear Discriminant Analysis
2. Logistic Regression

**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 residuals = 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.

**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**

Starts with an initial model that has all possible predictors. With each backward step, the variable with that is the least significant is removed. Backward steps are completed until the specific criteria, like adjusted R-square, can no longer be improved.

**STEPWISE**

This selection procedure starts with no predicters in the current model. Each step performs one forward and one backward step. The step wise completes when either another variable cannot be added during the forward step, or another variable cannot be removed during the backward step.

**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.

**Topic 3: TWO WAY ANOVA**

1. **Main Goal of the topic**

This topic’s main goal is the same as multiple linear regression:

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

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

Structure of the data: There are two predictors and each predictor is categorical

Assumptions (Exact same as MLR):

* Normality: Residuals have an appearance of being normally distributed
* Constant Variance: Residuals have an appearance of a random cloud around the 0 line.
* Independence: The observations are independent of each other.

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

* Summary statistics: Understand count or proportions by level of the categorical variable. Review a table of one categorical variable vs the other, with the cells populated by the response mean.

Diagnostic Statistics & Plots

* Raw Residuals of observed minus predicted
* QQ norm plot
* Studentized residuals
* Means plots using standard deviation bars provide a good visual to answer the constant variance assumption and give insight into whether the model is additive or not.

1. **Hypothesis Testing**

High Level F-test

Identical to MLR. Overall significance test for the model. Ho = All betas = 0. Ha =one beta is not 0.

Type 3 F-test

Provides significance of each factor. Also provides input into whether model is additive of non-additive. If interaction variable is significant, the model is non-additive.

Least Square Means for Effect Background t-test

The test compares the means of each level of each predictor. The purpose is to determine which means differences are significant. These also requires an adjustment for the multiple unplanned tests. These adjustments can be accomplished using Bonferroni, Tukey or Dunnett.

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

CONS: Can be difficult to interpret when the interaction term is significant.

PROS: Quick visuals to understand a two categorical predictor model with quick descriptive visuals and comparisons of means.

1. **General Analysis Flow (For Completionists. Good idea for yourself, not required for Final)**
   1. Be aware of how many factors of interest exist. One factor with a block or two factors.
   2. Visual the data through a table and a subsequent means plot mentioned above.
   3. Run a full model with both factors and the interaction term
   4. Check the diagnostics
      1. Residuals – normality, constant variance, independence
   5. Conduct overall F-test for model significance
   6. Perform type 3 f-test for factor/interaction significance
      1. If interaction is significant, report differences in means by group of the factor levels
      2. If interaction is not significant, you can report differences by individual factor levels
   7. Remember to perform a correction for multiple testing. In this class we used Bonferroni during the homework.

**Topic 4: Time Series**

1. **Main Goal of the topic**

* Forecasting future values based on previous observations.
* Determine if the mean of the time series has shifted after a point in time. Example: Did revenue increase after a specific marketing campaign.
* Determine if there are other predictors contributing to the response variable.

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

* Structure of the Data: collected over time. The observations are equally spaced over time.
* Assumptions
  + Stationary
    - Constant Mean - regardless of up/down behavior, it is centered around a specific mean value.
    - Constant variance – this is similar to MLR, the difference here is that it applies to every time, t, in the series.
    - Constant autocorrelation – simply put, the correlation between points is only dependent on how far apart they are, not where they are in the series. Essentially, you should be able to take subsets of your series and get the same acf plot, regardless of the location of the range.
* Non-Stationary
  + If the assumption are invalidated, we need to add complexity to get stationary.
  + A few ways to get to stationary:
    - Differencing – Change response to the difference between a point in time, t, and the previous time point, t-1.
    - Log Transformation is an option in time series
    - Add predictors to the model (this appears to be the most common).

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

Autocorrelation plots: These plots are effectively residual plots calculated for each log and then graphed. Both Auto Correlation Function (ACF) and Partial Auto Correlation Function(PACF) provide insight into which model to run. The table below provides the rules of thumb.

|  |  |  |
| --- | --- | --- |
| **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 |

Akaike Information Criterion(AIC): Statistic helps decide which ARIMA model works best. This allows you to run a models on either side of the rule of thumb suggestion to find the optimal model.

1. **Hypothesis Testing**

Durbin Watson test: This is a hypothesis test to determine if data has serial correlation. The statistic is a number between 0 and 4. A value of 2, denotes no serial correlation. Above and below 2, denotes a positive or negative autocorrelation.

Lower-level t-test: When adding predictors to the time series, you can use a t-test to determine significance of the predictors.

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

CONS:

* For stationary models the predictions ultimately regress toward the mean. Keep this in mind when forecasting/predicting as the visual of the regression towards the mean is visually obvious.
* Non-stationary models do not regress toward a mean, but do regress toward a regression line.

PROS:

* Time series questions are everywhere, this provides the methods to answer those questions.
* The time series also accounts for the independence assumption being violated and provide residuals

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

* Plot the time series data
* Determine stationarity, if needed act on non-stationarity based on suggestions above.
* Use rule of thumb to determine starting model and then use AIC to determine best ARIMA model
* Check the residuals of the final model to ensure ACF and PACF align with rules of thumb
* Proceed to answer key question. This was mostly prediction in class.

**Topic 5: Repeated Measures**

1. **Main Goal of the topic**

* Measures the same variable on each subject several times
* Use to answer a question of interests that involves several facets, leading to a multivariate response. If the question was answered using separate univariate analysis, the result could be misleading. Effectively, a parameter could be determined statistically significant when it is not.

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

* Repeated measures builds on other models, be it regression or analysis of variance. The assumptions and structure of the data relies on the underlying model being used.
* Whatever the assumptions are for the given problem, independence is off the table since it is repeated measures.

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

* Plot the data with associated standard error. The standard error helps to understand the covariance that is occurring.
* Residuals plots separated by subject over time provides a good view of correlation
* Residuals in general:
  + Normality and constant variance are still needed
* AIC and BIC used for model selection

1. **Hypothesis Testing**

Type 3 F-test: Tests for the significance of predictors and interaction.

Further hypothesis testing is dependent on the underlying model.

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

CONS:

* Small sample sizes can be problematic

PROS:

* Many kinds of studies lead to repeated measures. This tool gives the analyst an opportunity to address the correlation between data points, which is usually manifested in a subject being measured at multiple points in time.

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

* Plot the data, analyze summary stats
* Run a full model using repeated measures
* Check the residuals:
  + Normality and Constant variance
  + GLM will provide a test for sphericity
  + Proc Mixed will use AIC/BIC model selection
* Choose a final model
* Conduct testing based on the model, MLR, Anova, etc.

Section 2: Modeling a Categorical Response

**Topic 1: Logistic Regression**

1. **Main Goal of the topic**

* Provide a likelihood of how predictors influence a response occurring
* Can be used to determine if specific predictors can predict an event or not

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

* The structure of the data can be both categorical and numeric
* Assumptions: This breaks away from MLR as the residual diagnostics focus more on identifying leverage points than assessing normality, constant variance and independence.
* Observations are independent, explanatory variables are linearly related to the response through the log(odds) of the primary even occurring.

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

* Summary stats for continuous variables and proportion tables for categorical tables.
* Pairwise scatter plots (ggpairs) coded by response level help to see correlation.
* Influence diagnostic plots provide view into residuals, leverage and Cook’s D.
* Effects plots to display model output can help visually explain the interpretation
* If using the model to predict:
  + Use a confusion matrix to understand accuracy, specificity and sensitivity
  + Thinking about changing the prediction boundary to assist with the metrics of the confusion matrix.
  + ROC curves – provides a visual and Area-under-the-curve (AUC) metric to determine optimal prediction boundary based on the model. The goal of the ROC curve is to not be on the 45 degree line.

1. **Hypothesis Testing**

Hosemer and Lemeshow Goodness-of-Fit: This is a chi-square test. Ho = model is not a good fit. Ha= model is a good fit.

High Level Chi-Square tests: The Likelihood Ratio, Score and Wald tests can all be run at the same time. Ho = betas are 0. Ha = betas are not 0.

Type 3 Analysis: This chi-square test can give you an indication of which predictors are significant

Low level chi-square tests: This looks at each variable in the regression and provides insight whether it is significant or not. Could be a double click from the Type 3 analysis, as a certain level of a categorical could be significant while the rest are not. This could warrant additional investigation.

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

CONS:

* Interactions can be extremely difficult to interpret for an explanatory model. May be okay for a two predictor predictive model though.
* If the response has more than two levels, the tools change. ROC curves may not be available.

PROS:

* When dealing with a categorical response, this provides a way to understand the probability of the response happening
* Confusion matrix is extremely helpful when dealing with a predictive model.

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

* Gather summary tables, proportion tables by response level and ggpairs plots colored by response level.
* A few options to approaching model fit
  + Feature selection
  + Do one variable at a time
  + Put thought into adding interactions up front to see if it helps answer the question of interest.
* Assess the fit. Know that the goodness of fit and hosmer-lemeshow tests are from the days of small sample sizes. Cross validation is a good assessment of fit for today’s larger data sets.
* Check the residual diagnostics for points that have significant leverage or Cook’s D.
* Finalize the model and provide the beta estimates and interpretation.
  + Interpretation is on the odds ratio scale when you exponentiate the regression coefficients – exp (beta)
* Conduct prediction cross validation and confusion matrix. Iterate through with ROC curves to increase prediction accuracy, sensitivity and specificity.

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

1. **Main Goal of the topic**

* Used for Prediction/Classification, given a new observation into the data set.
* Is there a easy way to understand what variables contribute to difference in the predictor levels. Essentially, can we streamline the variables from MANOVA?

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

* Numerical data
  + The role of the response and predictors is reversed for LDA. So if residuals (related to the response is multivariate normal, then the predictors are in LDA)
* Assumptions:
  + Normality: The residuals are multi-variate normal. If residuals are non-normal, need to look at alternatives, logistic regression being one of them.
  + Constant covariance matrix: If done with a scatterplot color based on levels of the predictor. Inspection should reveal an elliptical pattern with a straight-line relationship. Can also complete an equal covariance matrices test in software.
  + If equal covariance is not present. Will need to use Quadratic Discriminant Analysis.
* My statement:   PREDICTORS for EACH category of the response (Y/N) must follow a multivariate normal distribution with the same covariance matrix.  Observations are independent.
  + Basically, replace residuals with predictors and you are essentially there.

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

* Univariate residual plots for each variable will provide understanding of variance and normality.
* Scatterplot coded by level of predictor to complete visual checks.
* Use frequency and proportion tables to understand how to structure the test set
* Discriminant function for Group: Used for new observations to predict, calculate each level based on the observation. The highest value is tagged as the classification.
* Classification Summary: Provides summary of error rates.

1. **Hypothesis Testing**

Test for homogeneity of covariance matrices: This is a chi-square test to check for equal covariance matrices. Ho= matrices not equal. Ha = matrices are equal

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

CONS:

* Can only be used with numeric variables.
* Need to be careful not to overfit the model.

PROS:

* Good tool for prediction involving multivariate analysis.

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

* Run through assumptions to determine if you use LDA or QDA.
* Conduct EDA, using a Frequency /Proportion Table. This could be helpful in identifying priors for running the code
* Construct the LDA model
* Shift the prediction boundary based on performance and classification table

Section 3: Unsupervised Tools

**Topic 1: Principle Component Analysis**

1. **Main Goal of the topic**

* Reduces the number of variables to be used in a model, be it explanatory or prediction.
* Provides insight into how much specific predictors account for the variability

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

* Data needs to be numerical.
* Assumptions:
  + No significant assumptions

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

* Summary statistics for all predictors
* Investigate correlation between the predictors
* Scree plots – Provide a visual to determine the number of Principle Components that should be used.
* Component Pattern graphs – provides insight into how the original predictors are associated with the Principal Components. Can be a line chart that graphs the eigenvectors by original predictor separated by Principal Component. Can also create a two by comparing Principle Components with the original predictors plotted on the chart.

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

PCA can provide a sanity check of which predictors will perform well in the predictive model built. Sometimes you can even use the Principle Components within the model. To conduct this sanity check, conduct the PCA on the continuous predictors. Plot the desired number of Principle Components and color code the points based on the response levels that you want to predict. If there is separation, move forward with testing whether using the original variables in the Principle Components or the Principle Components themselves in the predictive model.

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

CONS:

* Interpretation of the Principle Components can be hard for an explanatory model
* No guarantee Principle Components will be better.
* Sample size needs to grow as the number of predictors grow. Need to have 10 observations / predictor.

PROS:

* Reduces the number of variables to a more manageable number
* Removes any multicollinearity
* Provides understanding of how the predictors contribute to the variability

**Topic 2: Clustering and heatmap visuals**

1. **Main Goal of the topic**

* Helps identify variability in the data
* Helps to identify outliers
* Assist in identifying if the groups have statistical differeces
* Identify correlation between predictors

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

* Data is numerical in nature.
* No core assumptions in this technique.

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

* Root mean square standard deviation – This is calculated for each cluster. This should be as small as possible
* Semipartial R-square – measures homogeneity of merged clusters. Identifies loss of homogeneity due to combining two clusters, means you actually don’t want to merge the clusters.
* R-square -> Desire is to be large
* Pseudo F-statistic: Also referred to as the CH Criterion. Optimizes for the number of clusters.
* Pseudo t-statistic: Can only be applied in hierarchical methods. Optimizes number of clusters to smallest number such that the pseudo t squared is lower than a critical value.
* Cubic Clustering Criterion Index: Maximum value is the correct number of clusters
* Dunn index: Ration of the smallest distance between observations not in the same cluster to the largest intra-cluster distance.

1. **Hypothesis Testing**

* Low level Chi-square: Determine if there is a difference in numbers among clusters for a specific predictor

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

CONS:

* Predictors need to be standardized prior to clustering

PROS:

* Clustering paired with heatmaps provides great insight into the data being explored.