**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.
3. **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.)

1. **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)

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.

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

1. **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.

1. **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

1. **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.

1. **General Analysis Flow**
2. Identify the question of interest (See #1)
3. 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)

1. 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.

1. 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.

1. **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;

1. **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**

**Left for Student**

**STEPWISE**

**Left for Student**

**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**

**Left for Student**

1. **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.

1. **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

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

**Left for Student**

1. **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:

* 1. Mean estimate and standard error for the levels of just predictor A
  2. Mean estimate and standard error for the levels of just predictor B
  3. Mean estimate and standard error for the level combinations of predictor A and B

1. **Hypothesis Testing**

**Left for Student**

**High Level:**

**Lower Level (Contrasts):**

1. **Relationship between other techniques / Other Info**

**Left for Student**

1. **Analysis Overview**

**Left for Student**

1. **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;

1. **Limitations of the technique / Things to look out for**

**Left for Student**