Chris Horras, Megan Dibble, Camden Murtagh

**Introduction**

Every year hundreds of thousands of students apply to colleges all over the country looking to obtain a higher education in order to set up a long lasting, rewarding career. When considering which schools should top their application list, each potential student has different factors they consider and every student weighs these factors differently. Some may look for certain majors or programs while others filter the school by size. One factor that is almost always a determinate is graduation rate, that is the percentage of students that will graduate from that school within 6 years. Although these rates are always changing for every school, some schools are more successful at graduating students than others. This analysis looks to find an effective model to explain the graduation rate of an incoming freshman class at a given U.S. college.

**Initial Model**

In our initial model (A.1) we included all the predictor variables and specially coded each state into the regions that are used by the US Census Bureau. Using 0,1 coding for both the region and funding categorical variables, our baseline for the model was publicly funded schools in the west region. In this model we found eight significant predictor variables as well as a significant intercept. The normality assumption (A.2.1) and linearity assumptions were not violated, however in some cases the equal variance (A.2.2) was violated. Also, there was evidence of multicollinearity seen (A.1)

**Intermediate Steps and Models**

Our initial model showed some violations of the assumptions as well as showed evidence of multicollinearity (A.1). These equal variance issues were mainly seen in the predictors of number of applications (A.2.2.1), number of applicants accepted (A.2.2.2), and number of new students enrolled (A.2.2.3). We fixed these issues by creating some new predictors. We added Accepted/Apps to represent percentage of applications that are accepted for a school and we added Enrolled/Accepted to represent the percentage of students who enroll when accepted for a school. We then removed all three of the involved predictors because the full time undergraduate predictor still accounts for overall school size. After testing this new model we found remaining multicollinearity in the tuition predictor. So we added Spending/Tuition to represent the percentage of expenditure on the student out of what the student pays. We removed the spending predictor but left in the tuition variable to ensure we didn’t miss the possible significant effect of total cost. This final form of the model (B.1) still showed a high VIF for the Tuition predictor but was removed after running best subsets. Also, after checking the residual vs. each of the new predictors plots (B.2) we find that there are no obvious linearity or equal variance violations. Since we effectively removed multicollinearity, fixed all of the violated assumptions, and there are no influential observations, we now move on to running the best subsets regression.

**Best Subsets & Interactions**

We ran the best subsets regression and had JMP display the best 2 regressions for each value of p (C.1). There were two clear choices for best models due to low Cp values. While the model with 11 predictor variables had a marginally lower Cp, we chose the model with 10 predictors because it was simpler. Also, the p-value for the 11th predictor (FTUG) that when added to the regression was 0.146, making it insignificant.

After arriving at our model with all significant predictors, we tried adding interactions. We added interactions between the quantitative variables and the categorical variables, and none of these terms were significant. When looking at quantitative\*quantitative variables, we chose to test interactions that made sense and were interpretable. Again, none of the interactions were significant, so we stuck with our model chosen from the best subsets output.

To confirm our model choice, we computed a partial F-test using a full model including all predictors that we ran best subsets on and a reduced model which contained our selected 10 predictors (C.2). The test statistic was below the critical value, so we failed to reject that the slopes of the predictors that we excluded were insignificant. It should be noted that in order to accurately compare the full and reduced model, we excluded the 254 rows of data that had missing values for any of the predictor variables. (The original data set contained 936 rows). Without doing this, the reduced model would contain more observations than the full, so their total degrees of freedom would not match. Thus, the analysis of the final model has these 254 rows excluded as well.

**Final Model**

After arriving at the final model, we checked to make sure there were no significant issues. Since none of the VIFs were above 4 (C.3), the model was sufficiently free of multicollinearity. When checking the regression assumptions, we found that that the normality assumption looked ok (C.4.1). The studentized residual vs. predicted and predictors plots in C.4.2 all appeared to be not have any curves or fan shapes, which confirmed that the linearity and equal variance assumptions were not violated. Any slight fan in residuals was due to a concentration in values on one end of the graph. There were only 7 outliers in this regression, which can be found in C.5.2. Additionally, there were 3 high leverage observations, compared to the 26 found in the initial model (C.5.3). Finally, there were no influential observations.

**Interpretations of Results**

We can be 95% confident in each of the conclusions stated below, and the numbers stated correspond to when all of the other variables are average. Also, the model can only explain graduation rates for all U.S. colleges between the years of 1994 and 1995 since that is the population from which this data was sampled.

Colleges that receive private funding are associated with a 5.370% higher mean graduation rate than colleges that receive public funding. This relationship exists likely because private funding equates to more funding, which can be invested in the school and in students. Funding can allow for a higher quality of education and college experience, and in turn a higher graduation rate.

When the percent of new students from the top 10% of their high school class increases by 10% , there is an associated increase of 1.69% in the mean graduation rate. This increase in graduation rate is to be expected, since high achieving students in high school will likely be high achieving in college (or they will at least achieve enough to graduate).

Each increase of $100 in cost for room and board is associated with an increase of .141% in the mean graduation rate. While this relationship is statistically significant, it does not show a powerful relationship between room and board costs (which generally measure quality of housing) and graduation rate. Room and board is generally in the thousands of dollars, and even if it was one thousand dollars higher (the dorms were quite a bit nicer), this would be associated with only a 1.41% increase in mean graduation rate.

Each increase of 10 percentage points in alumni who donate to the school is associated with an increase of 1.81% in the mean graduation rate. Having more alumni that donate to the school results not only in more funds for the school to improve its overall quality, but also it affirms the positive relationship that past students had with the college.

Each increase of 1 percentage point in expenditure per student as a percentage of tuition is associated with a decrease of 8.123% in the mean graduation rate. This variable is measuring how much of a student’s tuition is devoted back to them by way of instructional expenditures. It is slightly counterintuitive that an increase in instructional spending is related to a decrease in graduation rate, but perhaps this relationship alludes to the importance of college spending in categories other than instructional materials (like recreation facilities, food services, etc.). These outside of class factors could make a difference in whether a student chooses to continue pursuing a degree, especially if they are on the fence.

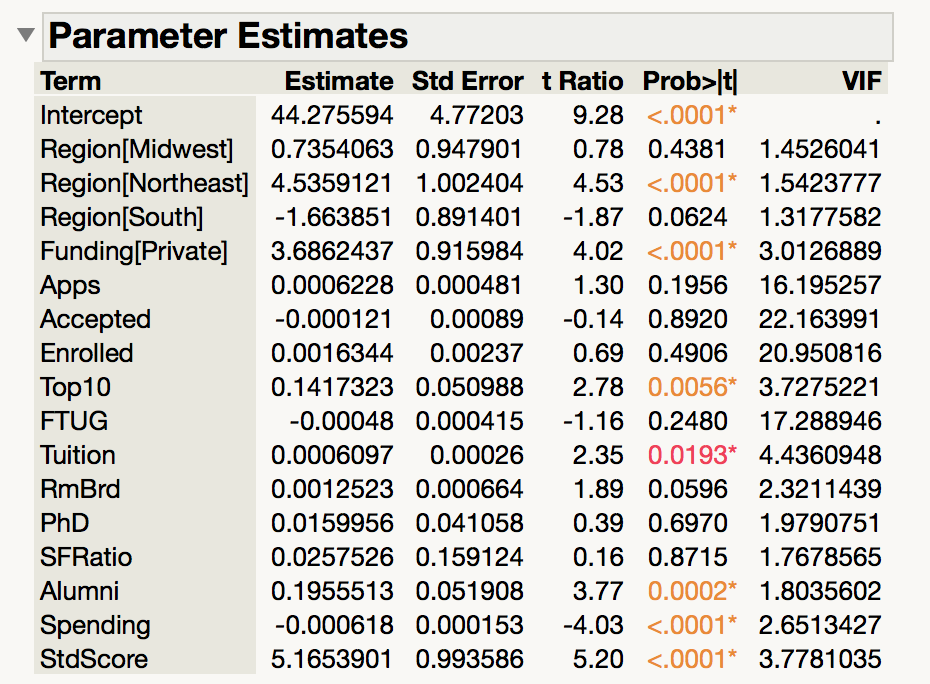
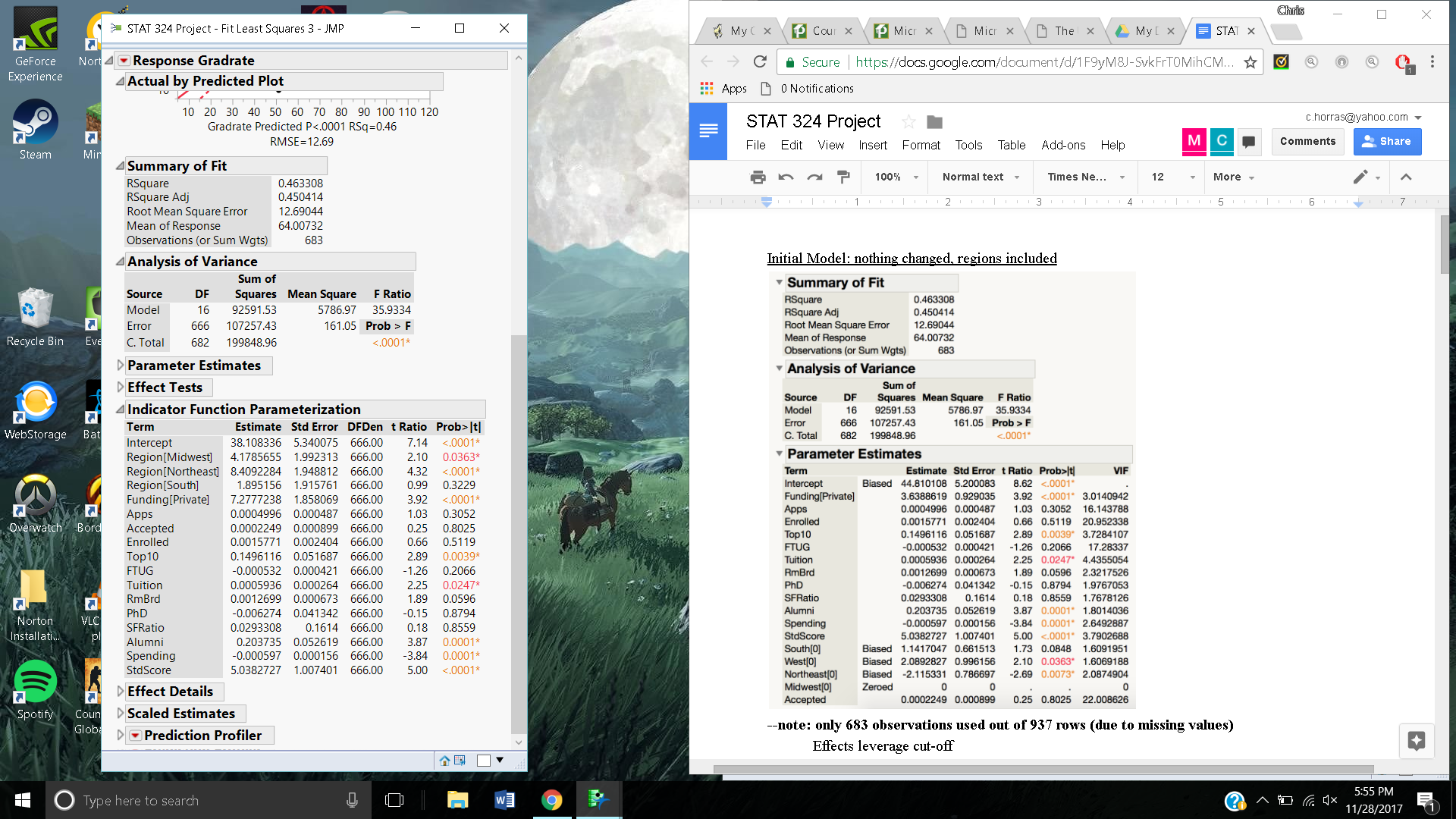
Each increase of 1 standard deviation in average SAT/ACT scores for the incoming class is associated with an increase of 4.988% in the mean graduation rate. When students perform well on standardized tests, this usually indicates their ability to perform well on tests in college, which would naturally be positively related to graduation rates.

Each increase of 1 percentage point in the percentage of applicants accepted by the school is associated with a decrease of 8.390% in the mean graduation rate. This variable has the most negative relationship to graduation rate. This relationship exists because when a school is less selective, students that may not have the academic habits for success will be accepted and will enroll. These students are then at a higher risk of not graduating college.

When looking at the relationship between college location and graduation rate, we compared all the regions to the West region. Colleges in the Midwest region are associated with a 5.193% higher mean graduation rate than colleges in the West, and Colleges in the Northeast region are associated with a 8.597% higher mean graduation rate. Colleges in the South did not have a significantly different graduation rate than colleges in the West.

**Appendix A: Initial Model**

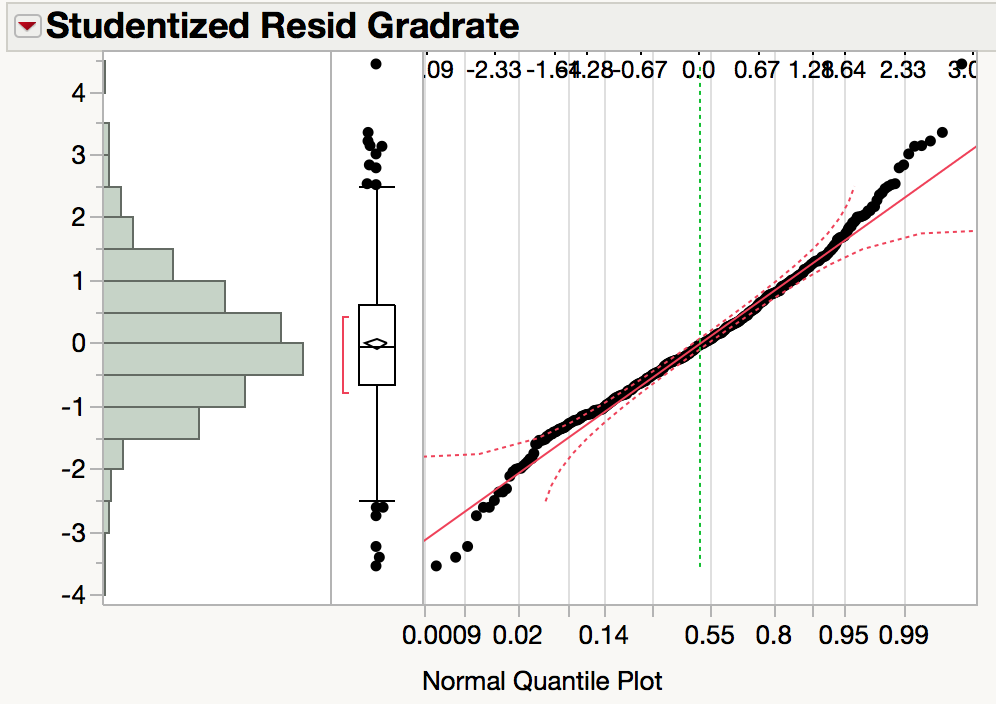
A.1: Fit Model for Initial Model



* Coded State variable into 4 Regions: West, Midwest, South, Northeast
* Ran a regression with all original variables plus new Region variable
* Checked VIFs, Assumptions, outliers, leverage, and influence

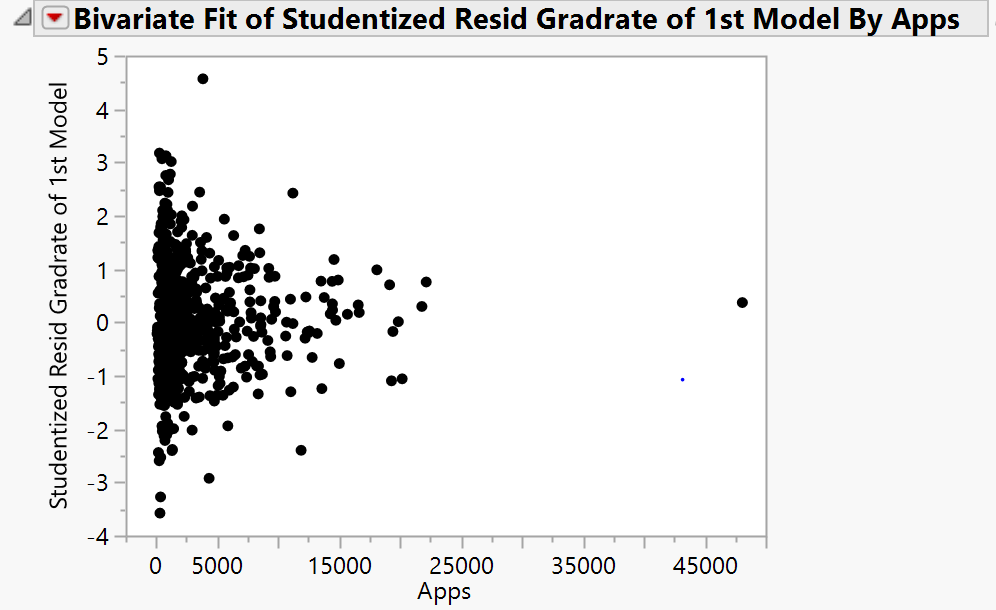
A.2: Assumption Checking

*A.2.1: Normal Probability Plot*

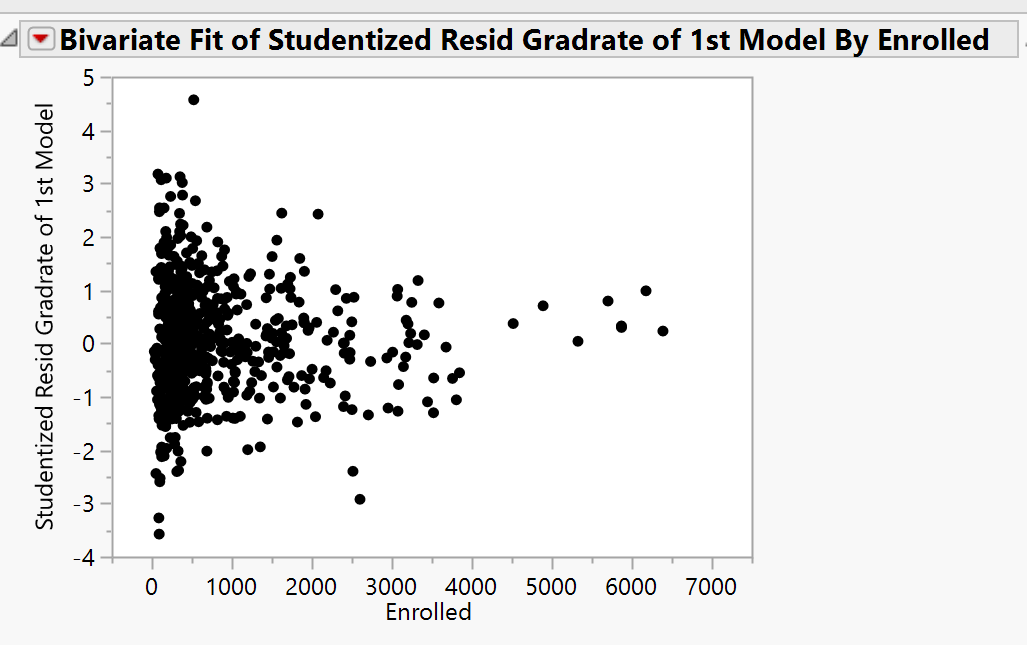


*A.2.2: Residuals vs. Predicted & Predictors Plots*

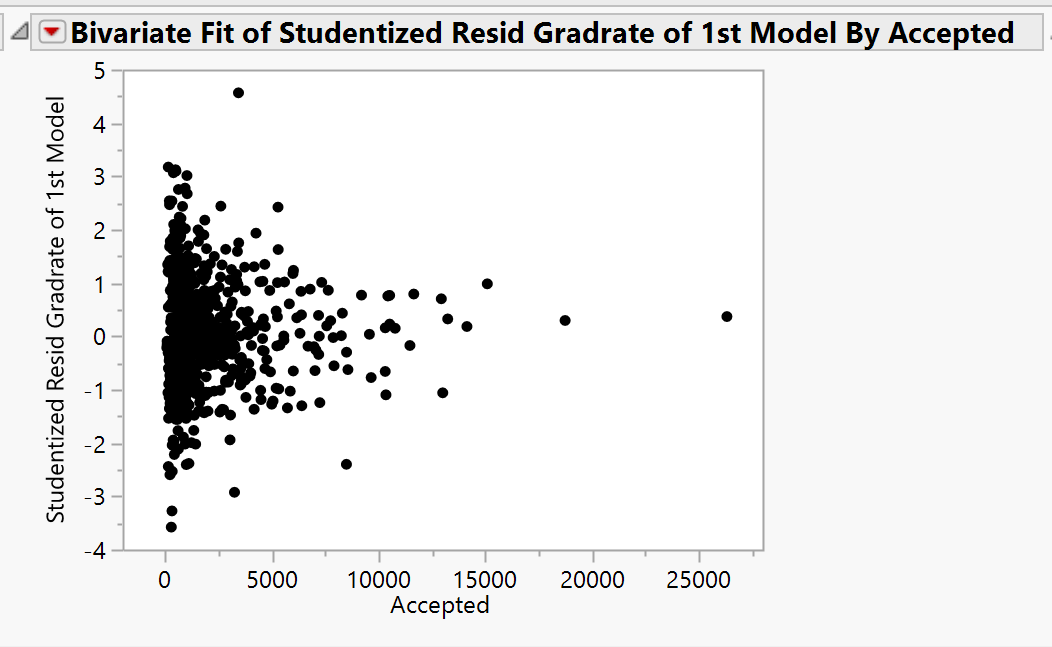
A.2.2.1



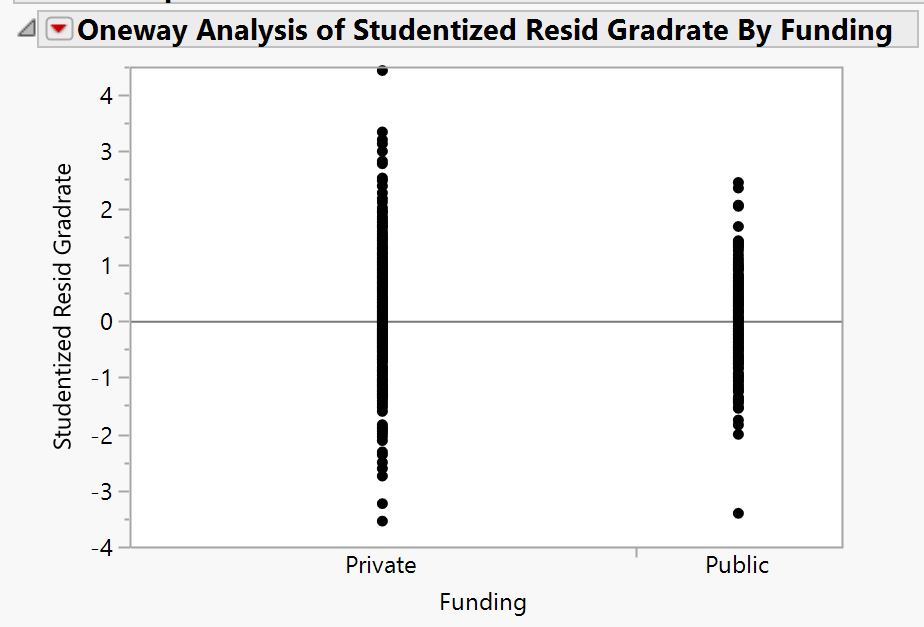
A.2.2.2



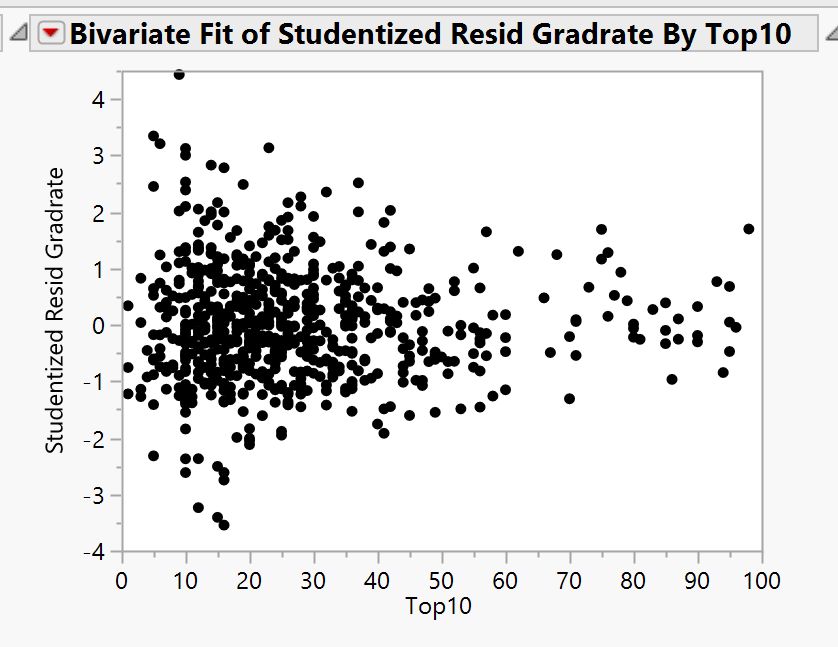
A.2.2.3



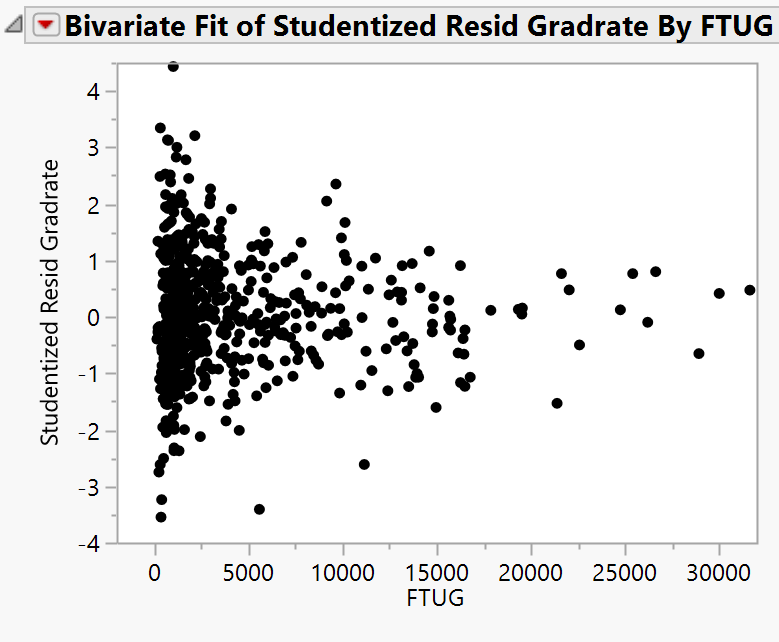
A.2.2.4



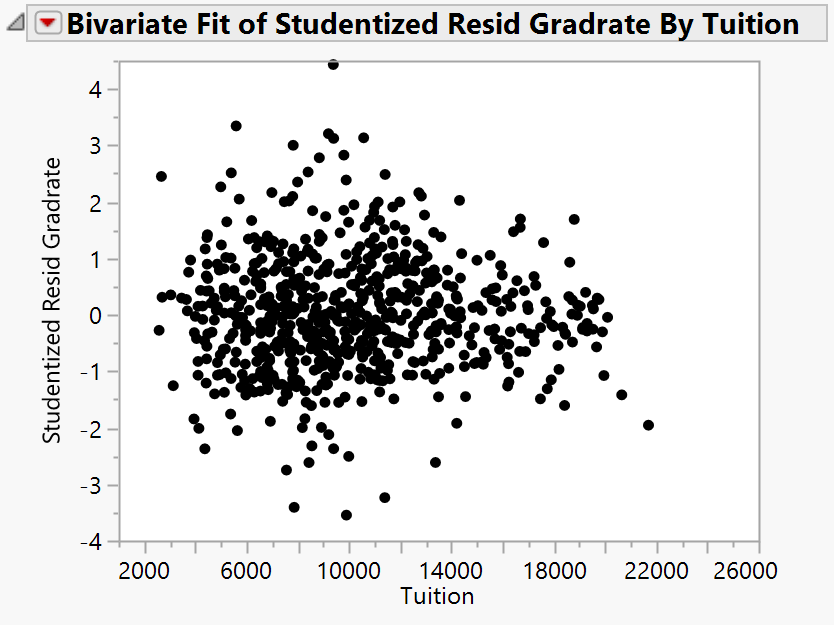
A.2.2.5



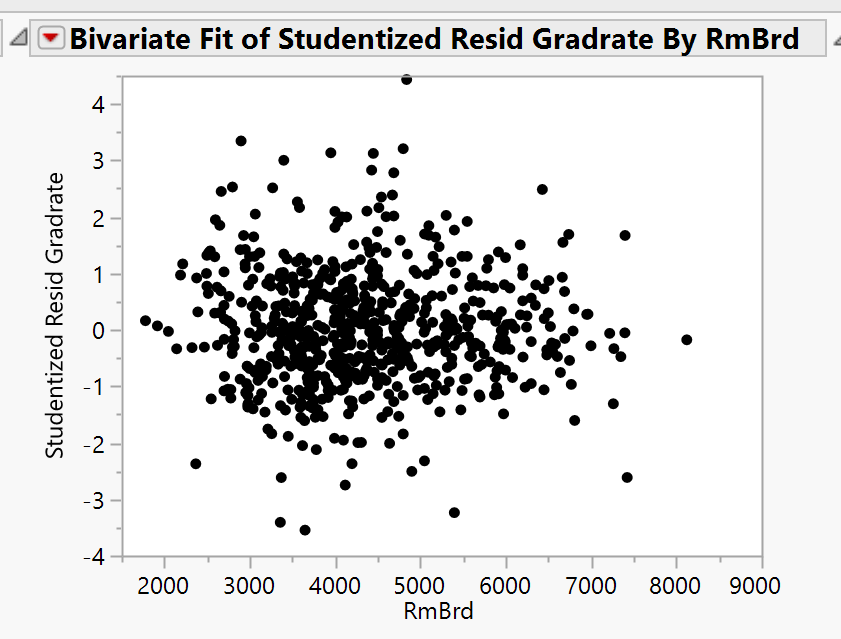
A.2.2.6



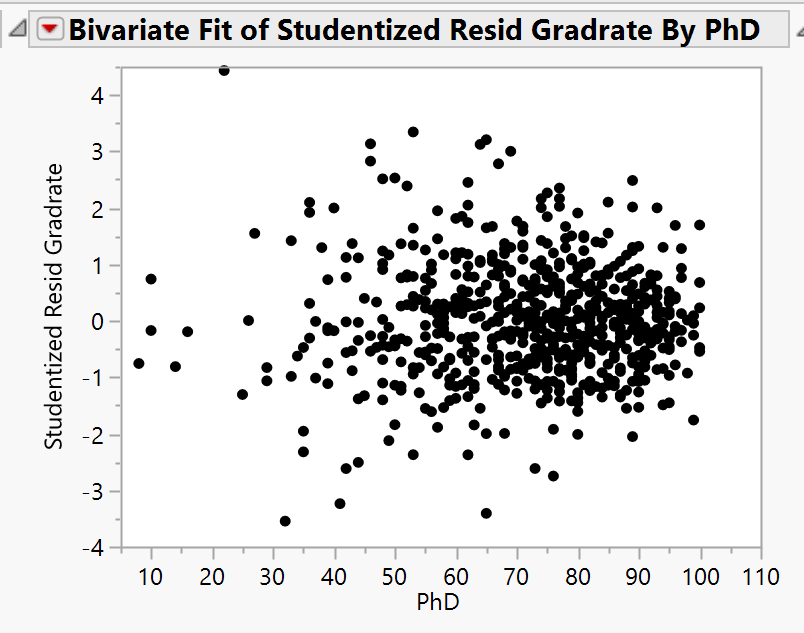
A.2.2.7



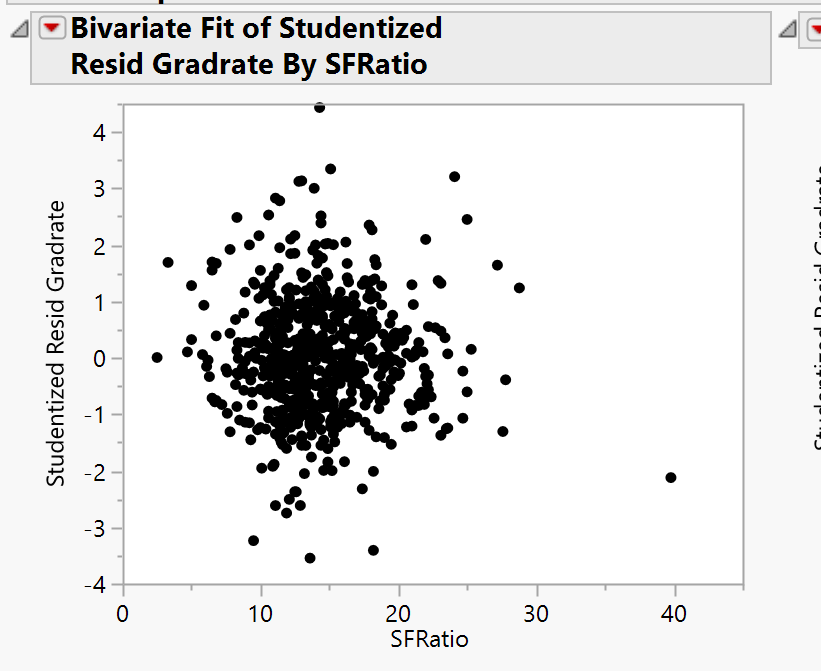
A.2.2.8



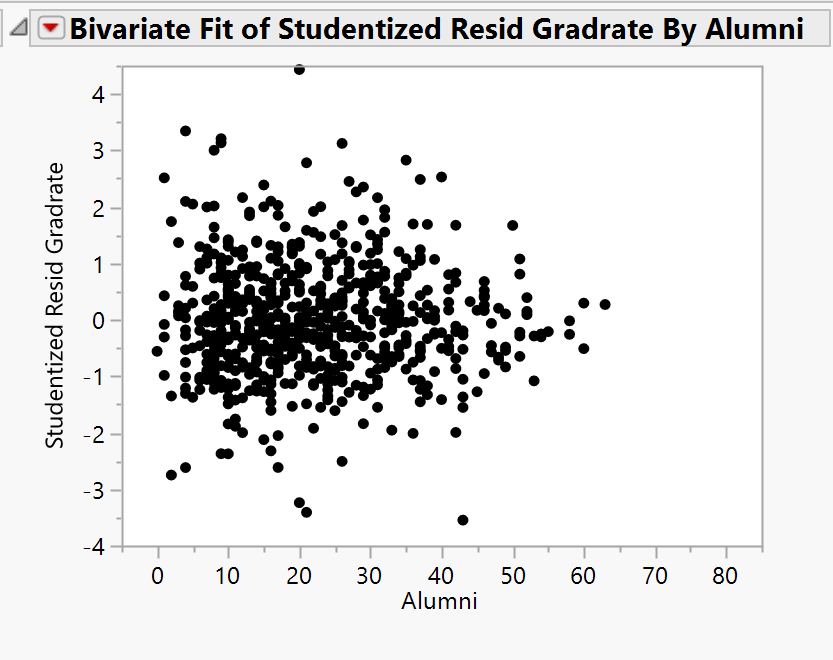
A.2.2.9



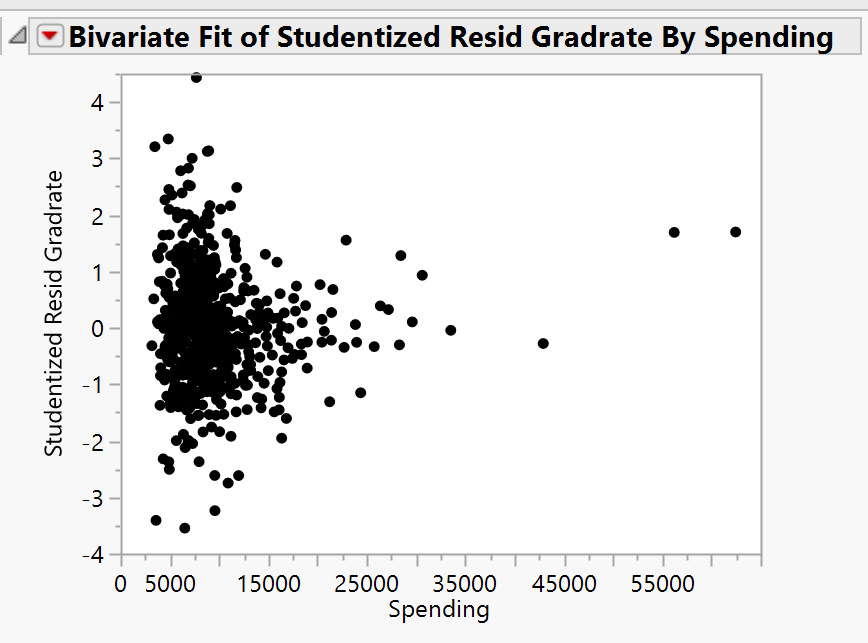
A.2.2.10



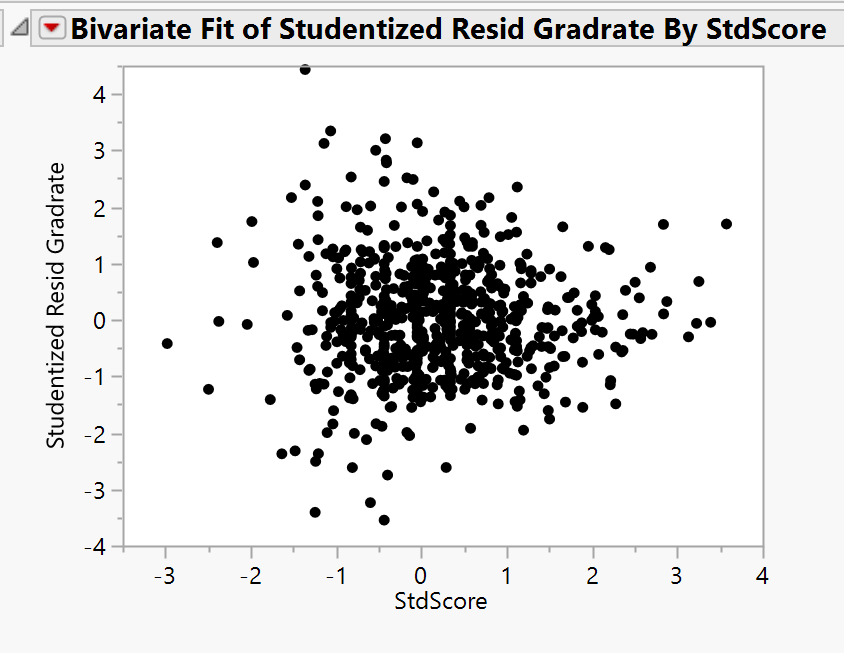
A.2.2.11



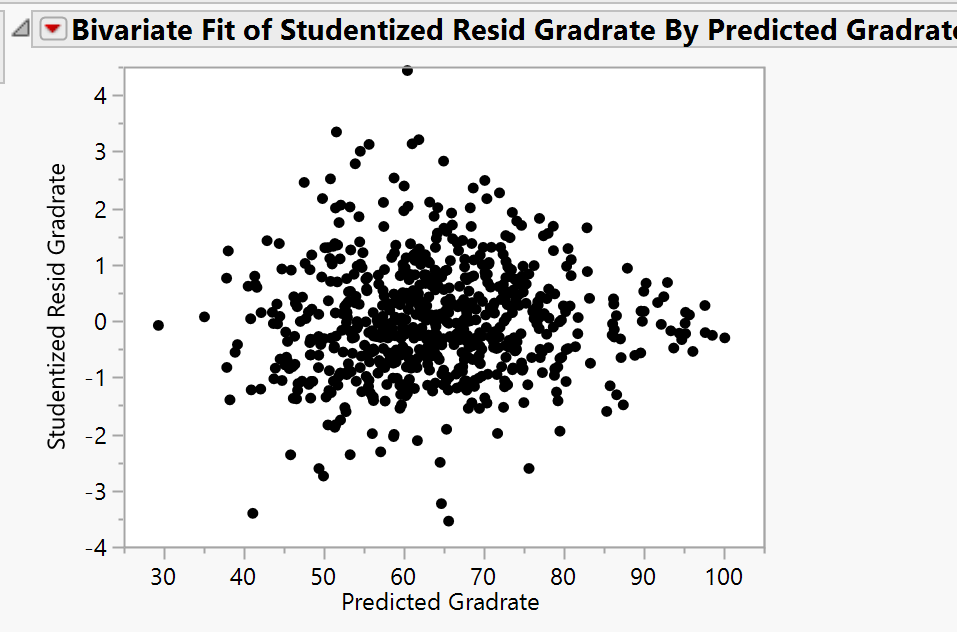
A.2.2.12



A.2.2.13

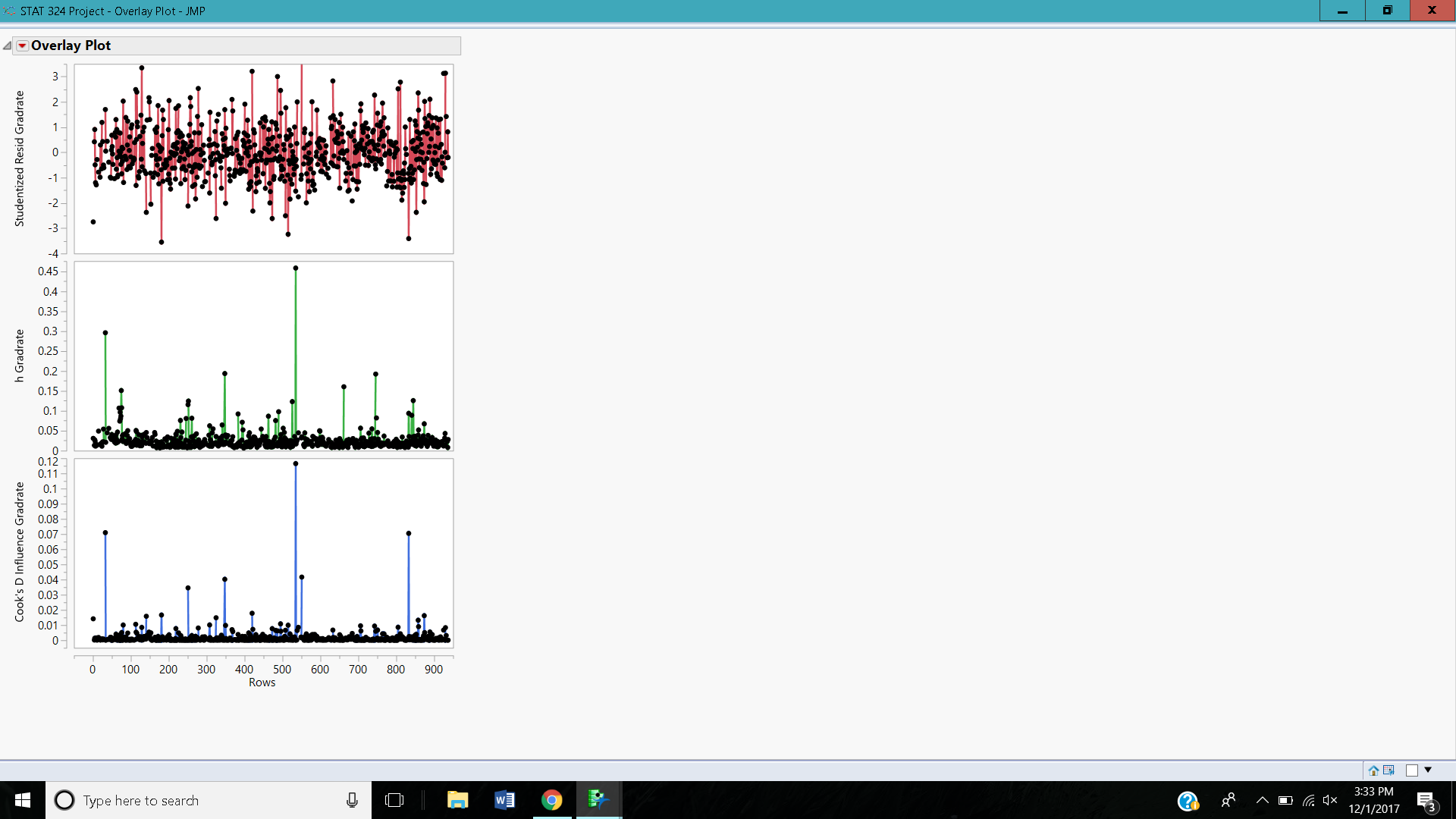


A.2.2.14



A.3: Checking for Outliers, Leverage, and Influence:

*A.3.1: Overlay Plots*



*A.3.2: Outliers*

|  |  |
| --- | --- |
| **Extreme High Outliers** | **Extreme Low Outliers** |
| Cazenovia College\* | Centenary College |
| Davis and Elkins College | Mount Saint Clare College |
| Lindenwood College | Texas Southern University |
| Salem-Teikyo University |  |
| Webber College |  |
| Wingate College |  |

\*= deleted from data set

*A.3.3: High Leverage*

cutoff =0.0703

Antioch University

Brown University

California Institute of Technology

Center for Creative Studies

Duke University

Indiana State University

Indiana Wesleyan University

Jamestown College

Johns Hopkins University

Michigan State University

Pennsylvania State University

Princeton University

Purdue University

Rutgers at New Brunswick

Texas A&M University

Texas Southern University

University of California at Irvine

University of California at Riverside

University of California at San Diego

University of California at Los Angeles

University of California at Berkeley

University of California at Santa Cruz

University of California at Santa Barbara

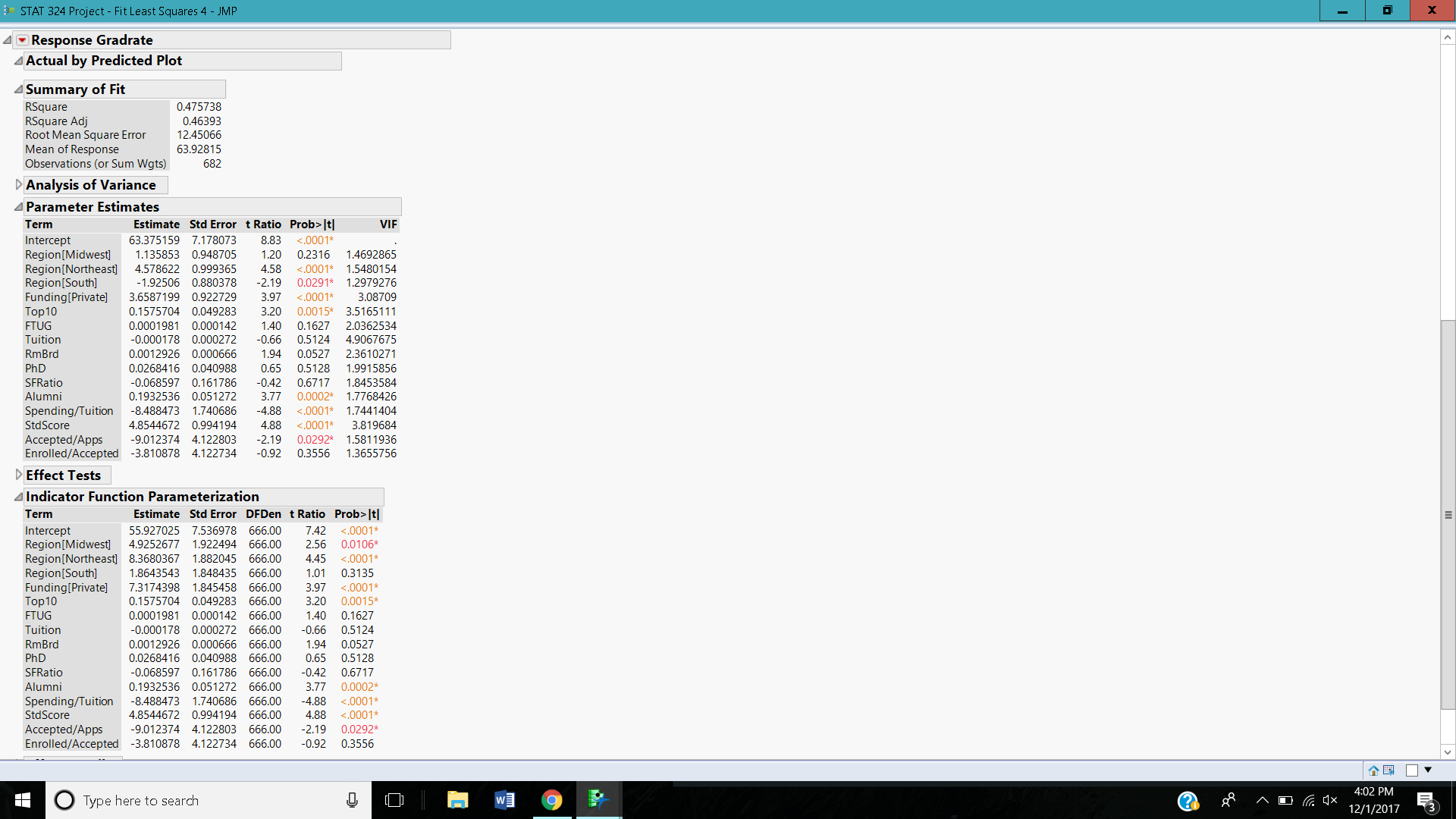
University of Illinois

University of North Carolina

University of Texas at Austin

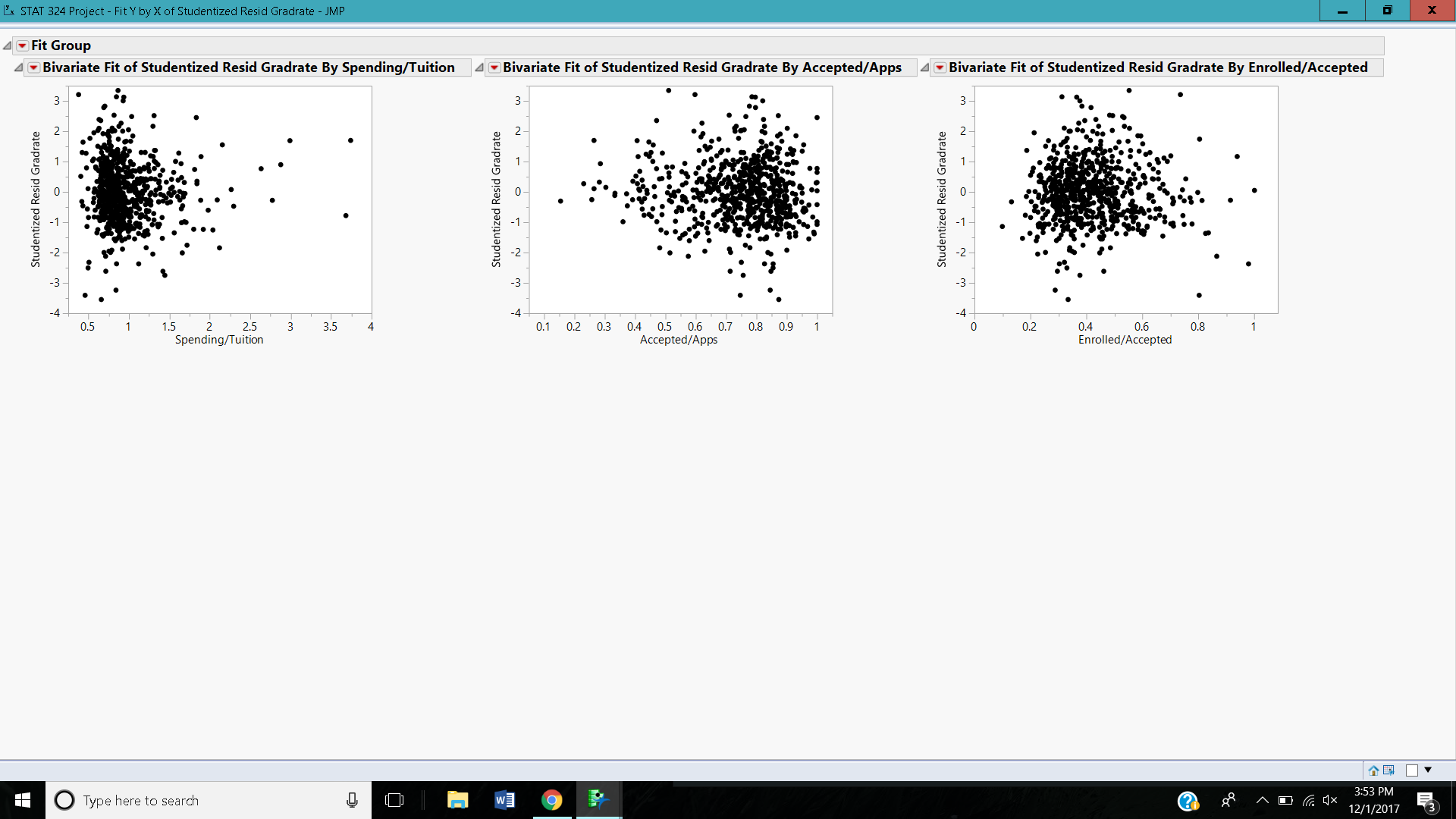
**Appendix B: Model 2 (after transformations)**

B.1: Fit Model for Model 2

****

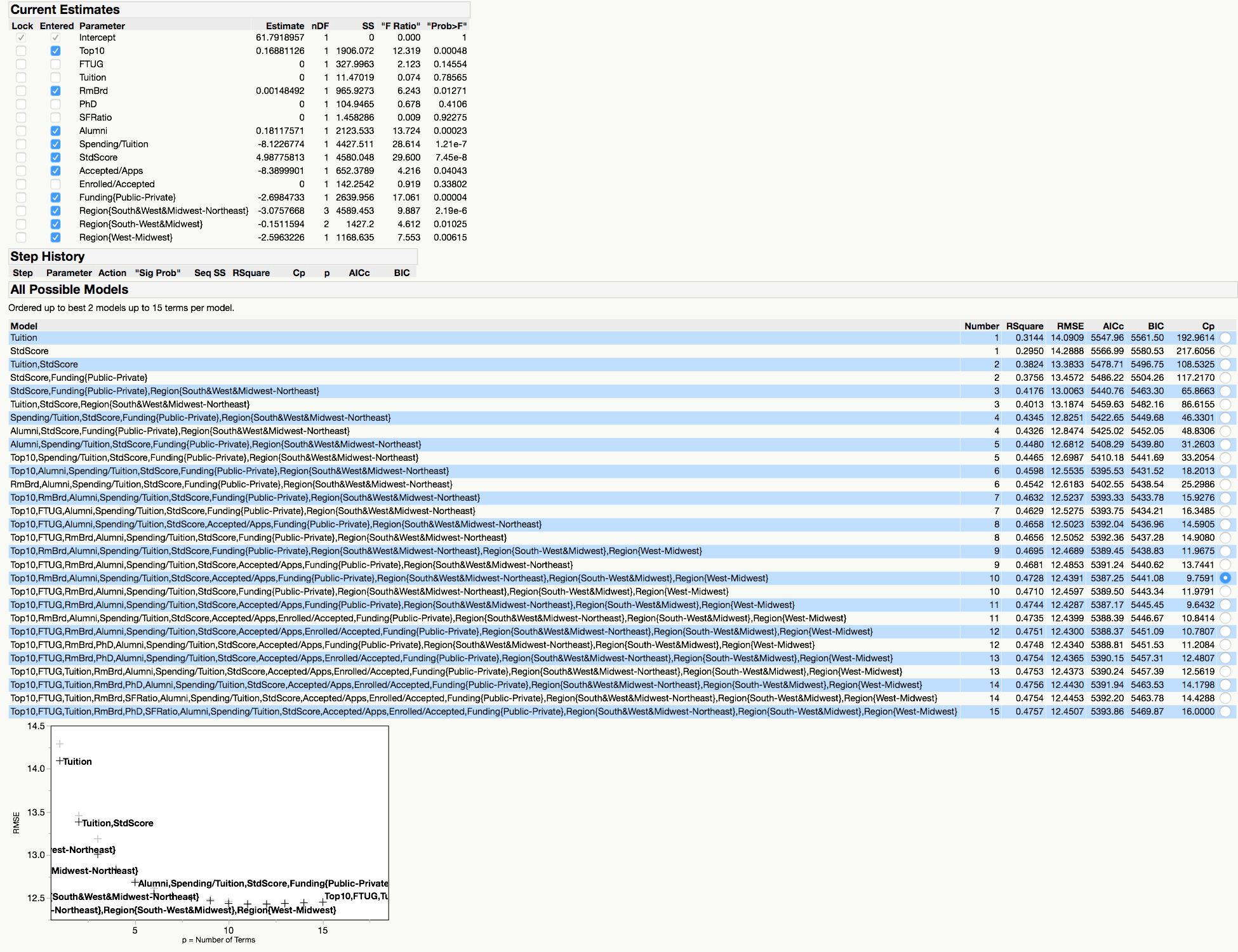
* Adjusted initial model by creating Accepted/Apps, Enrolled/Accepted, and Spending/Tuition
* Removed Accepted, Apps, Enrolled, and Spending from initial model

B.2: Residuals vs. Spending/Tuition, Accepted/Apps, Enrolled/Accepted



**Appendix C: Best Subsets and Final Model**

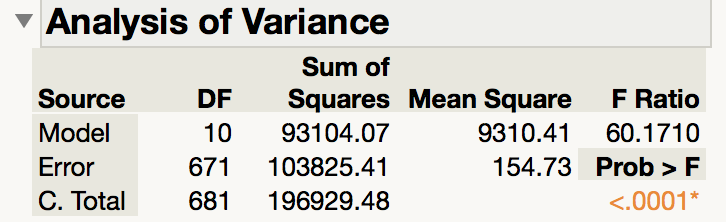
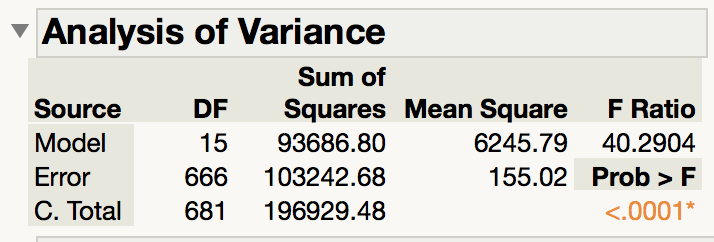
C.1: Best Subsets with transformed variables

****

* Ran best subsets:
  + With Accepted/Apps, Enrolled/Accepted, Spending/Tuition
  + Without Accepted, Enrolled, Apps, Spending
* Selected best model
* Centered all quantitative predictor variables
* Tried adding interactions between significant variables:
  + All quantitative variables\*each categorical variable
  + Other interpretable interactions: Top10\*Std Score, Accepted/Apps\*Std Score, Accepted/Apps\*Tuition
* No interactions were significant, so calculated partial f-test to confirm selection of final model
* Checked VIFs, assumptions, outliers, leverage, and influence

C.2: Partial F-Test

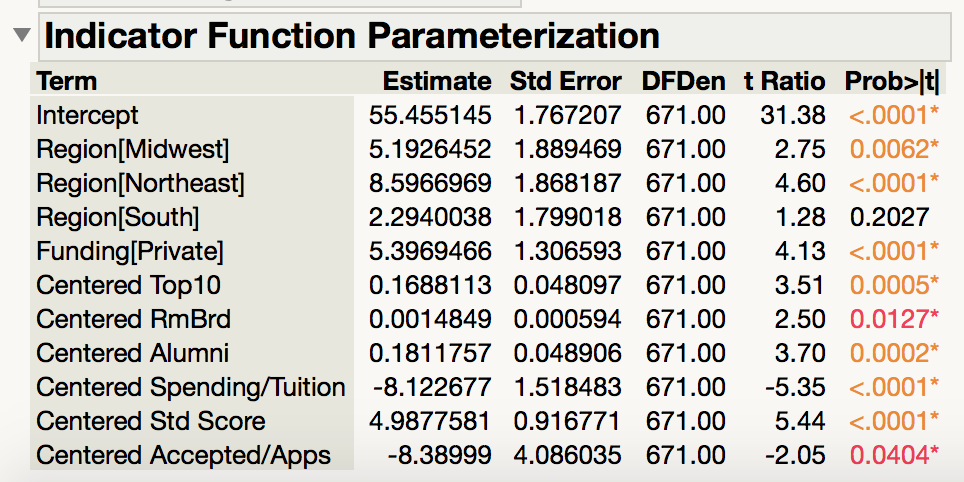
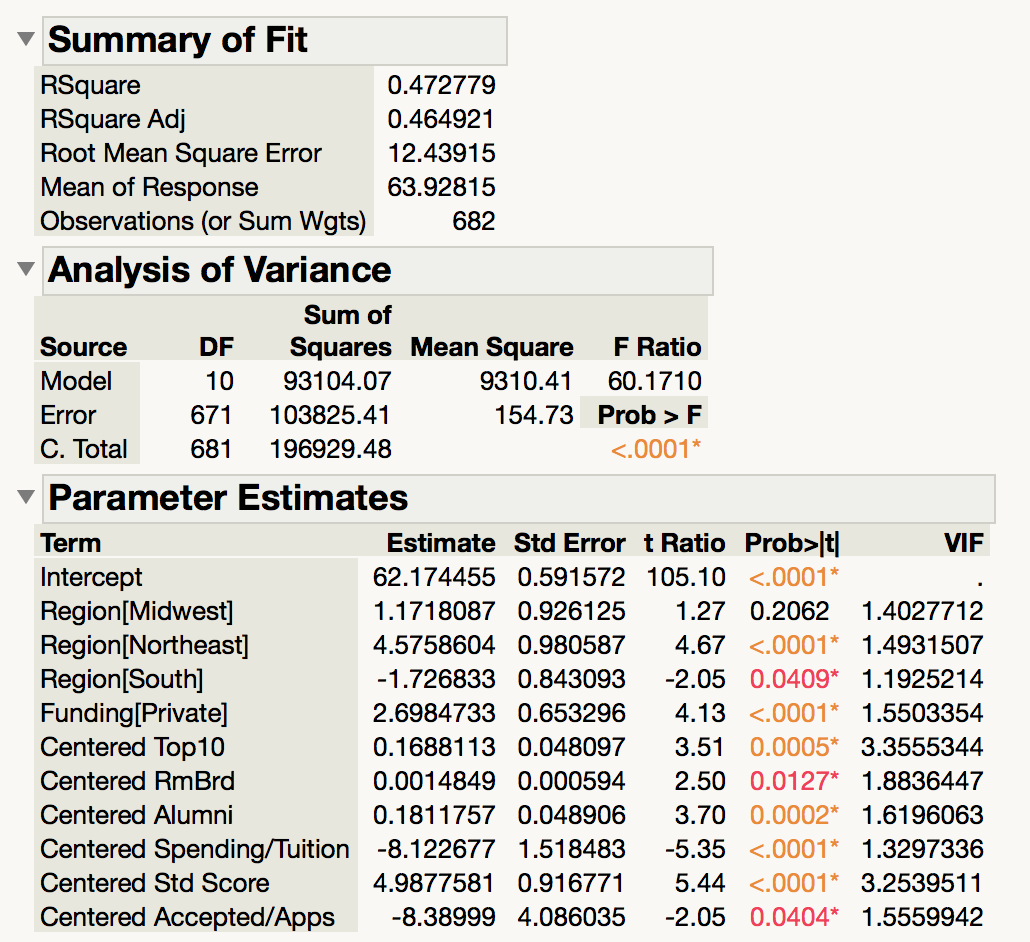
*Full Reduced*



F = [(103825.41-103242.68)/(671-666)]/155.02 = .626

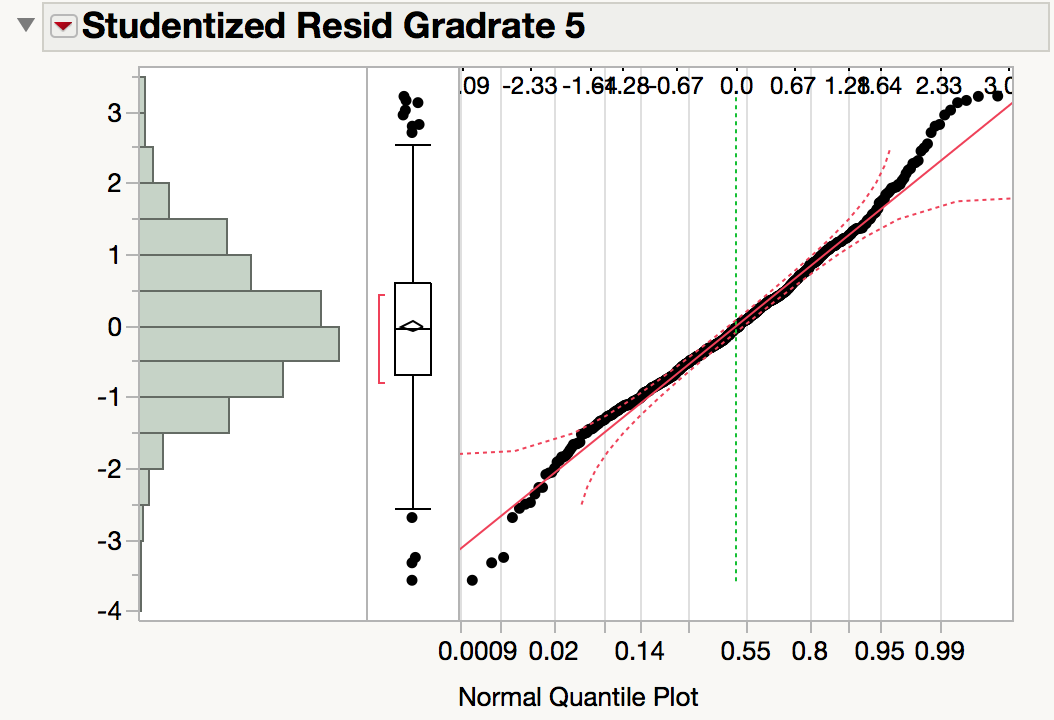
F (.05, 5, 666) = 2.2276

C.3: Final Model

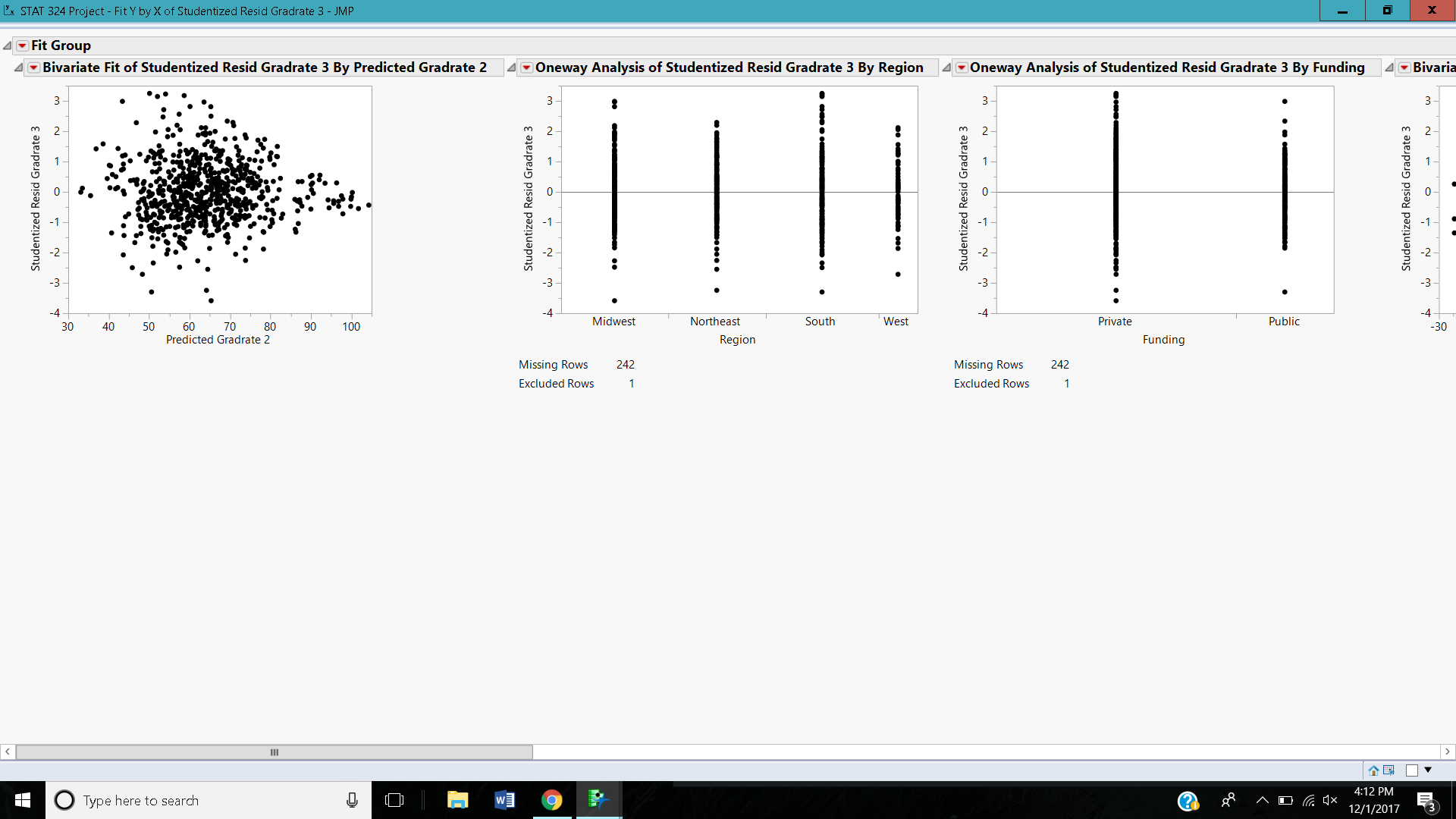


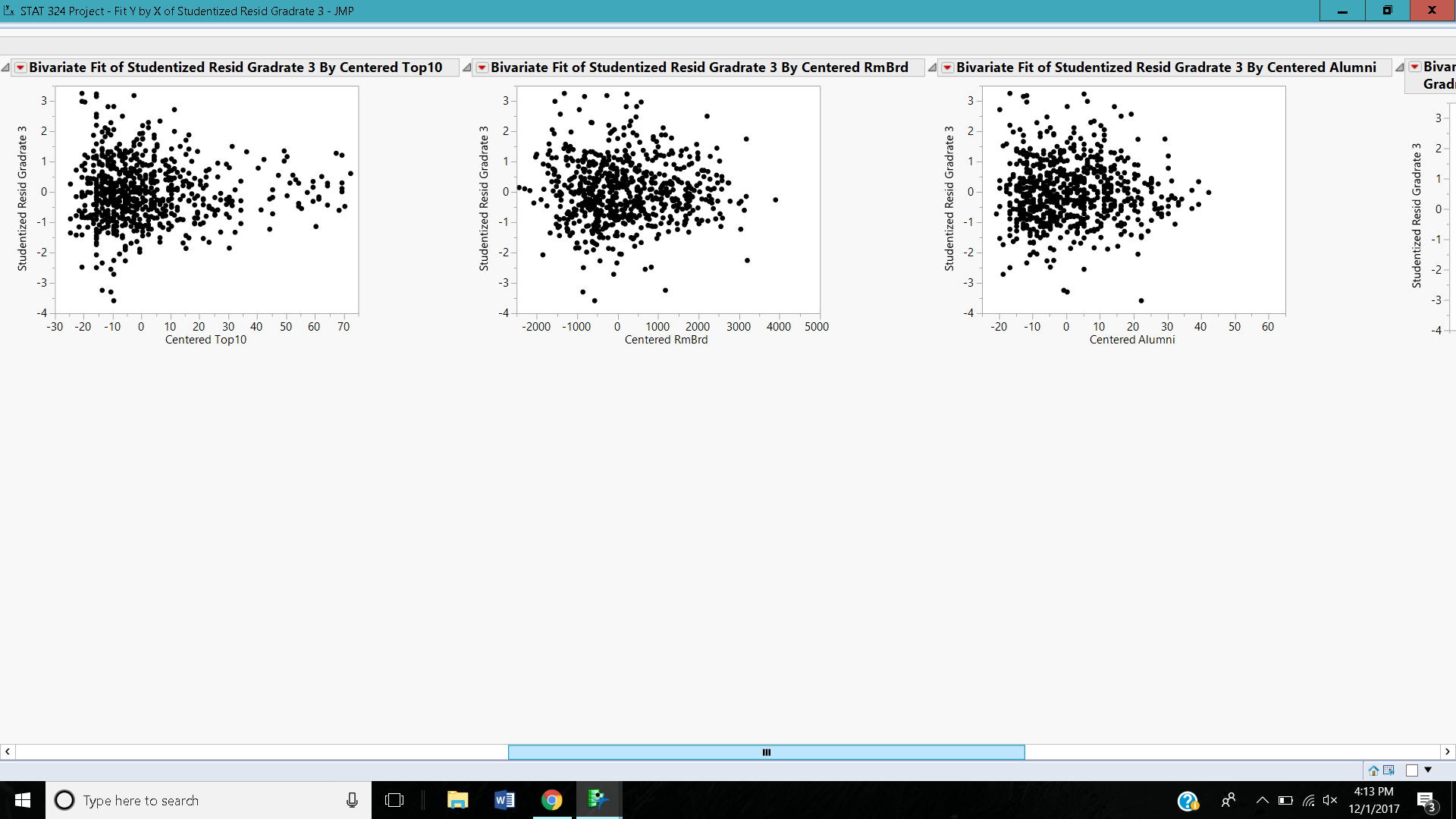
C.4: Checking Assumptions

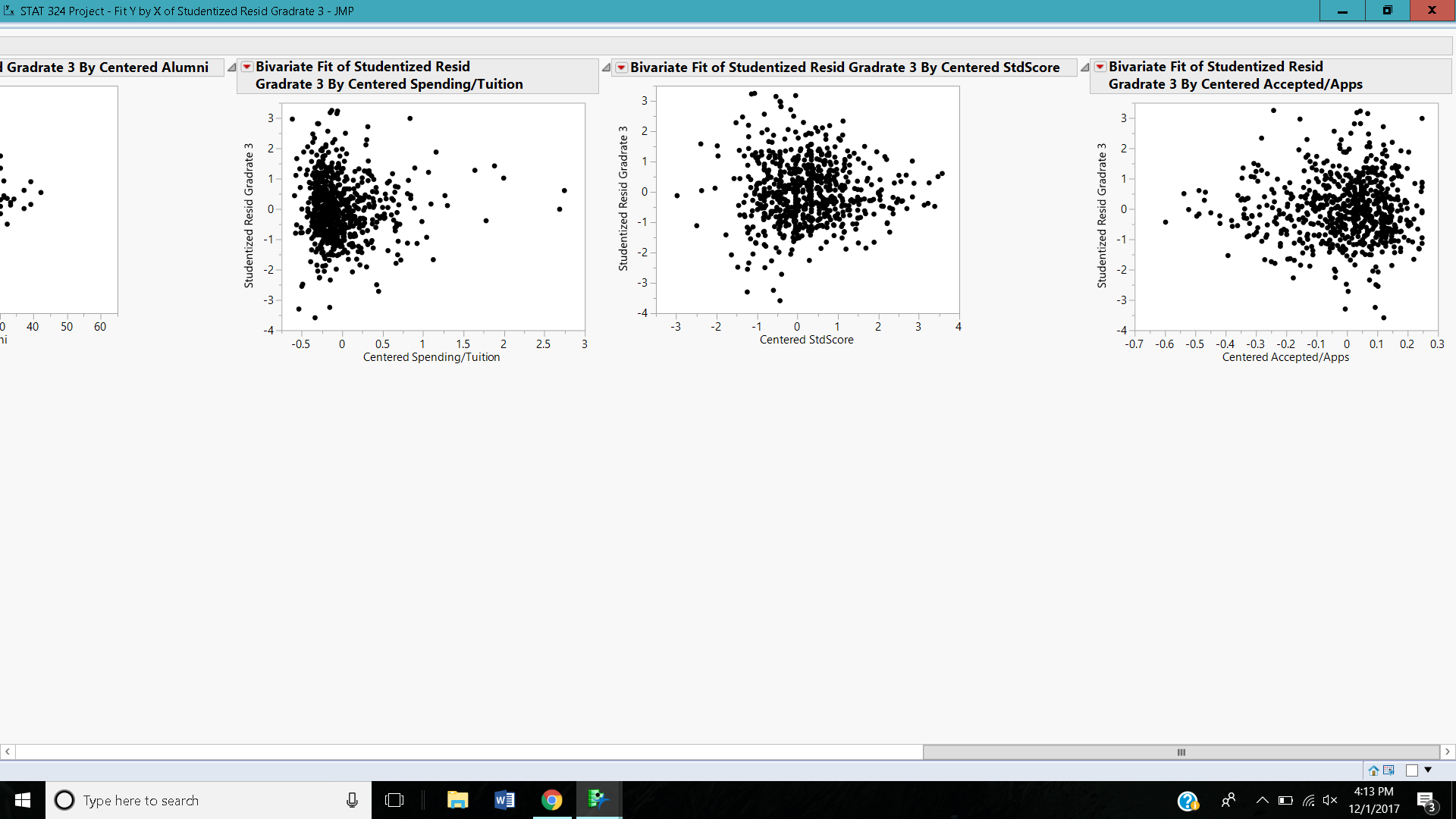
*C.4.1: Normal Probability Plot*



*C.4.2: Residuals vs. Predicted & Predictors Plots*

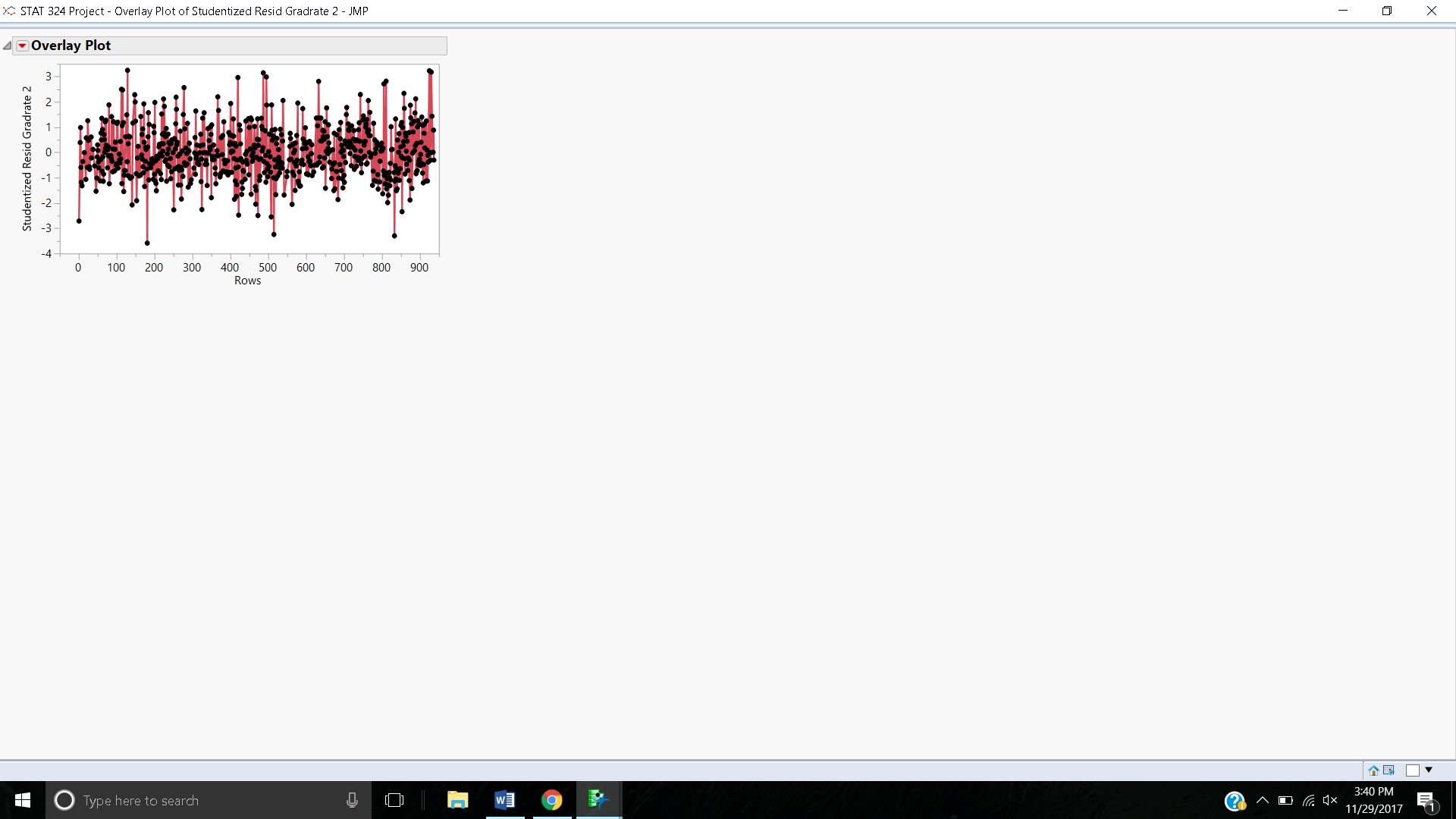


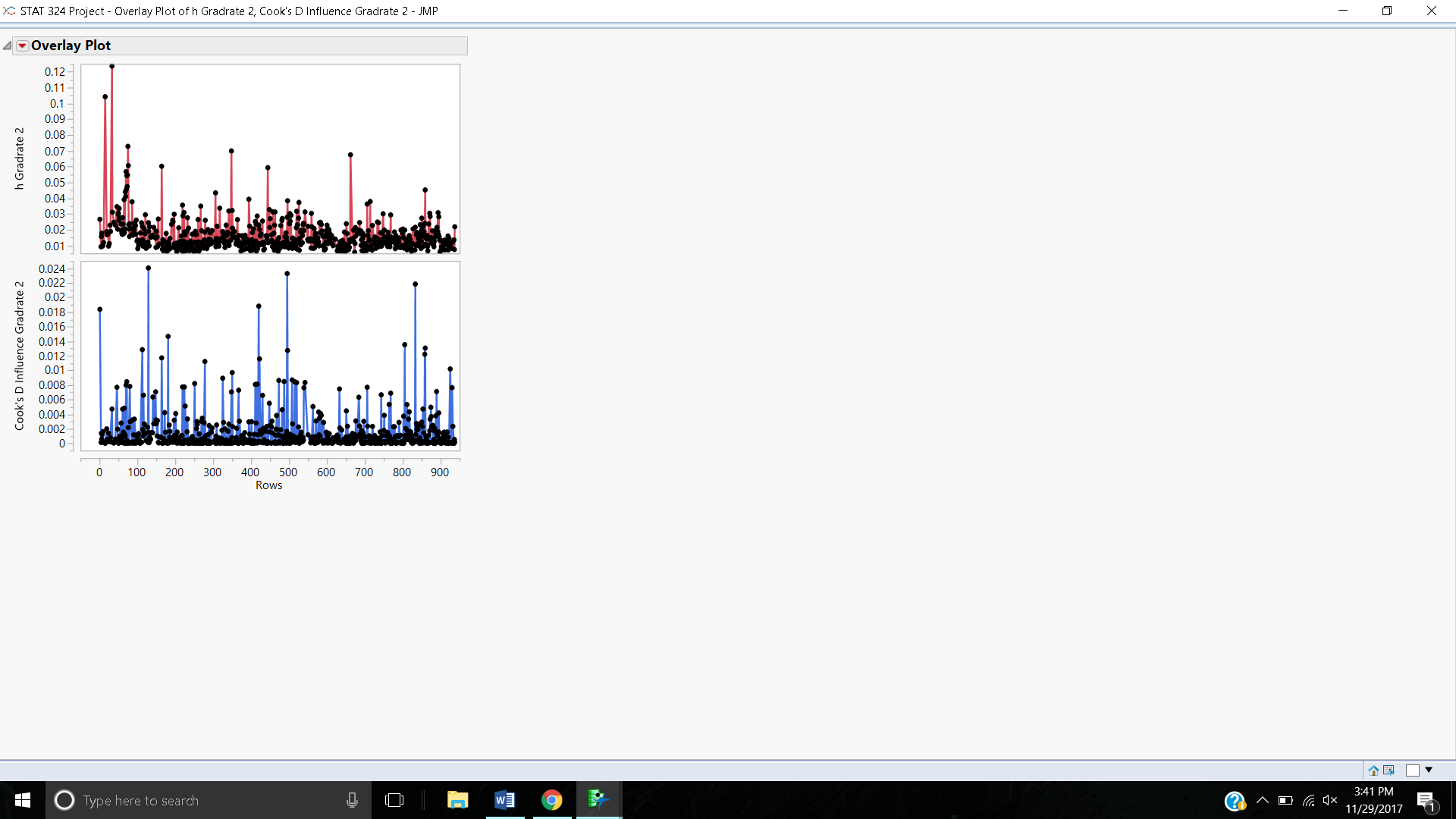




C.5: Checking Outliers, High Leverage, and Influential Observations

*C.5.1: Overlay Plots*





*C.5.2: Outliers*

|  |  |
| --- | --- |
| **Extreme High Outliers** | **Extreme Low Outliers** |
| Davis and Elkins College | Centenary College |
| Salem Teikyo College | Mount Saint Clare College |
| Webber College | Texas Southern University |
| Wingate College |  |

*C.5.3: High Leverage*

cutoff= .04755

Antioch University

California Institute of Technology

Johns Hopkins University