Coming to Online Classes Near You: Project 1

Sukanya Barman, Arjun Ganesh, Theo O'Neill, and Disha Patel

STAT 3220-002

October 18th, 2020



**TABLE OF CONTENTS**

|  |  |
| --- | --- |
| **SECTION NAME** | **PAGE NUMBER** |
| INTRODUCTION | 3 |
| DATA SUMMARY | 3-6 |
| ANALYSIS | 7-11 |
| CONCLUSION | 11-13 |
| SAS CODE | 13-14 |
| APPENDIX | 14 |

## 

## **Introduction**

Movies are popular around the world and transcend national borders. They allow viewers to escape their daily lives and explore both realistic and fantastical narratives that they would otherwise likely never experience and can satisfy the deeply-rooted human desire for good storytelling. Movies also represent a significant economic opportunity; the global entertainment industry earned over $100 billion in profits in 2019 and only continues to grow.[[1]](#footnote-0) Understanding what causes movies to be successful, then, is of extreme importance to studios and investors in the industry.

To this end, we will explore which factors have the largest impact on the gross revenue of movies. We will also investigate several related questions which have been found to be significant in predicting profitability in previous studies (e.g. Wallström & Wahlgren 2018[[2]](#footnote-1)): how a movie’s budget affects it’s revenue, if movies with lower content ratings on average result in higher profits, and what genres of movies tend to generate the most revenue. Understanding the impacts of the factors could aid creators and investors in minimizing their risk and maximizing the likelihood that their movies will be profitable.

The analysis in the rest of the report will include many different tests like global and nested F-tests (when applicable), t-tests, checking for multicollinearity, regression assumptions, influential points/outliers, and testing interactions between different parameters to see if they help predict the gross revenue to make the model stronger.

## **Data Summary**

The data our group decided to analyze is information regarding various films created around the world. This data was collected primarily for other movie watchers to see how well the movie is doing and learn information about it so they can determine whether or not it is something they would be interested in. The data comes from multiple sources which were put together by one of our team members to assure that there were enough variables for making a prediction. The data comes from Kaggle datasets; the primary data we used comes from a dataset that was scraped from the IMDb website, and one variable comes from a different dataset that contains Metascores, which are ratings given to movies by Metacritic, a website that aggregates reviews from leading critics in the entertainment industry. Because of the large number of movies contained in these datasets, we subset the data to include only movies that were labelled by IMDb to be from 2016, the most recent year included in our primary dataset, to better analyze variables that might play a role in predicting the movie’s revenue in the modern day. Whenever a movie did not have a corresponding Metascore or did not have a budget listed, we removed it from the dataset in order to be able to build a more accurate model. We also added the Metascore variable columns to ensure our group had enough data to work with.

One observation from this manipulated dataset contains multiple variables: budget ($), the company that created the movie, the country the movie is from, the director, genre, gross income, the title, the rating (PG-13, R, etc.), the release date, runtime in minutes, IMDb score (rated by users) out of 10, the star actor of the movie, the amount of votes on IMDb for the movie, and the writer, which are all found on one dataset created by someone on Kaggle who used IMDb data[[3]](#footnote-2). Finally, the dataset containing Metascores was also found on Kaggle, and that data originated from Metacritic and consists of a score out of 100, determined by movie critics.

### Data Dictionary:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Variable | Description | Quantitative or Qualitative? | If quantitative, units? | If qualitative, levels and dummy variables |
| Budget | The movie’s monetary budget | Quantitative | Dollars ($) | n/a |
| Runtime | The length of the movie | Quantitative | Minutes | n/a |
| IMDb score | User film ratings on a scale of 1 to 10 using IMDb | Quantitative | 1-10 scale, 1 being worst, 10 being best | n/a |
| Votes | Amount of votes for each movie on IMDb | Quantitative | # of people | n/a |
| Metascore | Critic ratings on a scale of 1 to 100 using Metacritic | Quantitative | 1-100 scale, 1 being worst, 100 being best | n/a |
| Rating | Film ratings, age suggestions of who should watch | Qualitative | n/a | Levels:   * PG * PG-13 * R * Unrated   Dummy variable:   * 1 if PG * 0 if non-PG   (since EDA shows PG has highest gross mean) |
| Production company | Company that produced the movie | Qualitative | n/a | Levels:   * Major studios (Paramount, Marvel, Disney) * Independent studios   Dummy variable:   * 1 if major studio * 0 if independent studio |
| Genre | Genre or type of movie | Qualitative | n/a | Levels:   * Comedy * Crime * Horror * Drama * Biography * Adventure * Action * Animation   Dummy variable:   * 1 if animation * 0 if non-animation   (since EDA shows animations have highest gross mean) |
| Country | Country the movie was created for | Qualitative | n/a | Levels:   * Canada * China * France * Ireland * Japan * Malta * UK * USA   Dummy variable:   * 1 if Japan * 0 if non-Japan   (since Japan has the highest gross mean) |

**EXPLORATORY DATA ANALYSIS**

We have chosen five predictor variables to test against our response variable, gross revenue. The five predictors are: adjusted budget,runtime, IMDB\_score, IMDB\_votes, Metascore. It is important to note both the gross revenue and budget were scaled more appropriately to ensure visually appeal of our scatter and plot of mean plots. In no way did this change our questions or EDA. (See scaling in code).

Quantitative Relationships versus Gross Revenue:

* Adjusted budget - The budget has the second strongest correlation realizing that a general theory would be the higher budget films may produce higher gross revenues because we assume there are more features in that movie to attract higher ticket sales. But this is just an informal analysis.
* Runtime - There seems to be not much of a correlation between runtime and gross revenue as most of the movies sampled seem to hover around 100 minutes. a
* IMDB\_score - There seems to be not much of a strong correlation with the score versus gross correlation. However though it is not linear, we do see that the scores above 8 had gross revenues of at least $150 million or more.
* IMDB\_votes - This predictor variable had the strongest correlation with gross revenue. The one outlier is Deadpool. It had over 600,000 votes and it is labeled as an outlier as all other votes were around the 30 to 200 thousands range.
* Metascore - Again, there seems to be not much of a strong correlation with the metascore versus gross correlation. Since these scores are compiled by professional critics, the metascores hover around 40-80. Even if it received a higher metascore than 80, the revenue was still around the same as some of the lower scored movies.

Qualitative Relationships versus Gross Revenue:

* Country- The plot of means for countries with gross revenues is an interesting analysis. Here are the mean values of y for each level: {Canada: 102.47, China: 95.447, France: 11.446, Ireland: 14.014, Japan: 138.14, Malta: 52.853, UK: 55.884, USA: 80.207)}. It is very interesting to see that Japan has the most movies cumulating the highest amount of gross revenue.
* Rating- The plot of means for ratings corresponds with my initial theories. Here are the mean values of y for each level: {PG: 157.64, PG-13: 82.694, R: 38.318, Unrated: 0.0351}. There were no G rated movies that made it into this table of movies. This shows that lower ratings allow all ages of the audience to come and watch the movie increasing ticket sales.
* Genre-The plot of means for genres with gross revenues is very scattered. Here are the mean values of y for each level: {Action: 95.067, Adventure: 85.79, Animation: 192.63, Biography: 39.803 ,Comedy: 36.485,Crime: 48.607, Drama: 31.572, Horror: 63.221}. The Animation genre has the highest amount of gross revenue out of all genres. A potential theory to keep in mind is that Animation caters to all audiences or age groups.

After analyzing the scatter plots & correlation quantitative variables and plot of means for the qualitative variables, the variables we conclude to be useful in predicting gross revenue is: Budget, IMDB\_votes, Metascore, Rating, and Genre.

## **Analysis**

**INTRODUCTION**

In the analysis section below, we started off by deciding which alpha level we will use throughout the section. The alpha will be necessary for any t-tests, F-tests, etc. that we perform in the analysis. The alpha value that we chose was 0.05 as it is the most standard alpha value to use for analysis. The general process that we will follow starts by working with just the quantitative variables to see which will be important to work with. Then, we thought it would be beneficial to test qualitative variables with an interaction between the qualitative variables (rating and genre) to see if any were significant enough to keep in the model. Finally, to build the model we planned to test a qualitative and quantitative interaction (rating and budget) to see if it is important to the model. After adding and removing variables and interactions, we decided to go through the process of checking assumptions, checking for multicollinearity, and find/evaluate influential points/observations. Our “variables of interest” from the EDA include the budget, votes, metascore, genre, and rating of movies.

**STAGE 1**

The first task is to see which variables and interactions are important predictors of gross revenue of a movie, so we decided to start with quantitative predictors. The three quantitative variables tested to predict the response variable (gross revenue in millions) were budget (in millions of dollars), votes, and metascore. These were the quantitative variables selected as they seemed to have the strongest linear relationship with the gross revenue from the EDA previously conducted. The “variable of interest” selected based on the EDA was the budget as our subquestion proposed was: does the budget provided to create a movie help determine how much money it makes? Budget made the most sense as a researcher to make the variable of interest since as a group we decided that it is very possible that the budget of a movie plays a role in how good a movie is which then determines how many people will actually pay to watch it. No interactions were conducted in this step because it is not necessarily a great idea to conduct quantitative/quantitative interactions. Our starting model was: .

We then conducted a global F-test to see whether or not the model was statistically significant. Our hypotheses were: the null (H0): all betas = 0 and the alternative (Ha): at least one beta is not equal to 0. After running the test in SAS, the output (appendix Table 1) showed that the F-statistic was 103.10 with a p-value of <0.001. Additionally, an F-critical value of about 2.67. Therefore, for , the p-value is less than alpha and the F-critical value is less than the F-statistic. Therefore, we reject the null hypothesis and conclude the model is statistically significant.

Next, we decided to perform individual t-tests on one of these quantitative variables which is the budget as it is the “variable of interest.” In order to see if the relationship is significant, the hypothesis of the t-test was: the null (H0) being = 0 and the alternative (Ha): . The output (appendix Table 1) shows that the ’s p-value is <0.001. This p-value is less than which is 0.05. Therefore, we reject the null hypotheses in the individual t-test and can conclude that the budget is significantly different from 0 and deemed as the “most important (quantitative) predictor.” Thus, the stage 1 end model ends up being:

**STAGE 2**

In this stage, we began to test and incorporate our qualitative variables. Before we begin, let’s note that the stage 2 start model is . The qualitative predictors that we deemed important from the EDA were genre and rating. As these are qualitative predictors, we turned them into dummy variables. We chose the dummy variables for each of the qualitative variables based on the EDA, choosing the level that generally produced the highest revenue mean to represent 1 in the dummy variable column. So, for genre, if the movie was animated it was represented as a 1 while the rest of the genres (ex. Comedy, Cime, Horror, Drama, etc.) were represented as 0. For the rating, if the movie was PG, it was represented as a 1 while the rest of the ratings (PG-13, R, Unrated) were represented by a 0. The “variable of interest” we selected was genre with the subquestion: does genre really impact the revenue made by a movie. Specifically, we picked it because we thought that the genre would impact the audience thus the revenue more than just the rating. Additionally, we thought that it would be beneficial to try and add a qualitative/qualitative interaction, so we used both the qualitative variables (rating and genre) and created an interaction with them. So, finally, the stage 2 model we ran in SAS was:

We then conducted a global F-test to see whether or not the model was statistically significant. Our hypotheses were: the null (H0): all betas = 0 and the alternative (Ha): at least one beta is not equal to 0. After running the test in SAS, the output (appendix Table 2 ) showed that the F-statistic was 68.15 with a p-value of <0.001. Therefore, for , the p-value is less than alpha. Therefore, we reject the null hypothesis and conclude the model is statistically significant with the added variables.

Next, we started off by performing a t-test on the interaction between the qualitative variables. In order to see if the relationship is significant, the hypothesis of the t-test was: the null (H0) being = 0 and the alternative (Ha): . The output (appendix Table 2) shows that ’s p-value is 0.7177. This p-value is greater than which is 0.05. Therefore, we fail to reject the null hypotheses in the individual t-test and can conclude that the interaction between genre and rating is not significantly different from 0. This means that the interaction and its main effects (DummyGenre and DummyRating) are removed from the model as well. Due to this, the qualitative “variable of interest” was not tested as it was removed after testing the interaction. Thus, the stage 2 end model ends up being the same as stage 1’s model with only the quantitative variables:

**STAGE 3**

In this stage, we begin working with more interactions that could be beneficial to the model. The model we ended stage 2 with was the same as the very first equation of:

However, we thought that it would be a good idea to test the interaction between the budget and the genre of the movie. We thought this would be an interesting interaction to assess as we thought that logically these two variables would be the best predictors of the gross revenue of a movie. Thus, we added this interaction and also added back the main effects removed from stage 2 (genre and rating). So, the model being fit into SAS and tested in this stage is:

We then conducted a global F-test to see whether or not the model was statistically significant. Our hypotheses were: the null (H0): all betas = 0 and the alternative (Ha): at least one beta is not equal to 0. After running the test in SAS, the output (appendix Table 3) showed that the F-statistic was 75.39 with a p-value of <0.001. Therefore, for , the p-value is less than alpha. Therefore, we reject the null hypothesis and conclude the model is statistically significant with the added variables.

Next, we started off by performing a t-test on the interaction between the quantitative/qualitative variables, specifically budget and genre like mentioned before. In order to see if the relationship is significant, the hypothesis of the t-test was: the null (H0) being = 0 and the alternative (Ha): . The output (appendix Table 3) shows that the ’s p-value is 0.0018. This p-value is less than which is 0.05. Therefore, we reject the null hypotheses in the t-test and can conclude that the interaction between budget and genre is significantly different from 0. Thus, the stage 3 end model ends up being:

**FINAL MODEL:**

***GrossRevenue* = β0 + β1*(budget)*+ β2*(votes)* + β3­*(metascore)* + β4­*(genre) +*  β5*(rating) +*  β6­*(budget\*genre) +* ε**

**SUMMARY OF MODEL BUILDING**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Stage | # of Predictors | Global F-test (p-value) | adj-R2 | MSE |
| 1 | 3 | <0.0001 | 0.7036 | 2957.22 |
| 2 | 3 | <0.0001 | 0.7036 | 2957.22 |
| 3 | 6 | <0.0001 | 0.7758 | 2237.46 |

The model at the end of Stage 3 appears to be the best model out of the available options. Though all 3 models pass the Global F-test with p-values of <0.0001, the Stage 3 model (which includes the 3 quantitative variables, 2 qualitative variables, and the interaction between budget and genre) has the highest adjusted R2 value at 0.7758 and the lowest MSE at 2237.46.

**EVALUATING THE MODEL**

*Regression Assumptions*

Now that we have created a basic model, we must evaluate this model. The first step in doing so is to ensure that the model meets regression assumptions by looking at the residuals, which are assumed to be an estimate of the true random error.

The first assumption is that errors have a mean value of zero; this assumption is tested by plotting the residuals against each of the independent variables or predicted y values () and look for clear trends or patterns. Looking at the scatter plots (Figure A), there is some clumping in some areas, but there aren’t any clear trends or patterns, so the assumption is met.

The next assumption is homoscedasticity, meaning that the errors have a constant variance () for all levels of the independent variables. To test this assumption, we must plot the residuals against the predicted y values () and look for patterns. Based on this plot (Figure B), there is clear evidence of a fanning out pattern. To correct this violation, we transformed the response variable (gross revenue of the movie) by taking the square root of all of the data points. After the transformation, the distribution of the residuals becomes much more randomly distributed, as shown by Figure C. Thus, the model becomes:

*y\*=√(GrossRevenue)* = β0 + β1*(budget)*+ β2*(votes)* + β3­*(metascore)* + β4­*(genre) +*  β5*(rating) +*  β6­*(budget\*genre) +* ε

This new model still meets the previous requirement, where there are no trends in the plots of the residuals versus each of the independent variables.

Using the new, transformed model, the next assumption is that the errors are randomly distributed, which is also met by the plot of the histogram (Figure D) because it roughly follows the normal distribution. In addition, the QQ plot looks fairly linear around the diagonal line (Figure E). The final assumption is that the errors are uncorrelated; this assumption can be met through the Durbin-Watson test for autocorrelation. The output (Figure F) shows that we fail to reject the Durbin-Watson test for correlation, meaning that there is no evidence of positive or negative first-order autocorrelation.

*Multicollinearity*

The model now passes the test for all the regression assumptions, so the next step is to check for multicollinearity using VIF after removing the interaction. The results (Figure G), show that none of the variables are anywhere close to 10, which is the cutoff for determining multicollinearity, and the average is close to 1 as well.

*Influential Points and Outliers*

Finally, we checked the model for outliers and influential points. There were many outliers, as shown by Figure H, including observations 13, 20, 87, 115, 84, and 33. In terms of the Cook's Distance statistic, which measures the overall influence of outlying observations on the beta values, observations 11, 14, 24, 33, and others may be considered influential (see Figure I). The Leverage statistics (Figure J) show that 33, 84, and 24 have the most leverage. Removing these influential points had little effect on our overall model, and so they were not removed because these points were representative of our population (movies released in 2016) and not errors in our data. We are trying to retain as much data as possible. Also, the movies that exerted the greatest leverage, observations 33 and 84, were *Finding Dory* and *Sausage Party*, which makes sense because both of these movies are animated and one is rated PG, both variables which play a large role in our model.

## **Conclusion**

*Interpretation*

Our final model to predict the gross revenue of a movie is the following:

Sqrt*(GrossRevenue)* = β0 + β1*(budget)*+ β2*(metascore)* + β3­*(votes)* + β4­*(genre) +*  β5*(rating) +*  β6­*(budget\*genre) +* ε,

where β0 =2.76, β1=0.022, β2=-0.0048, β3­=0.000028, β4­=0.58, β5=2.75, and β6­= 0.018.

The value of β1 implies that for every increase of $1,000,000 in a movie’s budget while other factors are held constant, the square root of its gross revenue will increase by approximately $22,000. This makes sense because a higher budget could allow for more advertising, enable more involved special effects, or hire big-name actors which would likely draw larger crowds. β2suggests that if a movie’s metascore increases by one point while other factors are held constant, the square root of its revenue will decrease by $4,800. This is perhaps surprising, as one might expect that higher ratings from critics might correlate with higher revenue, but could be explained by critics valuing different factors in films than the general public. β3suggests that for every additional vote for a film on IMDb while other factors are held constant, the square root of its revenue will increase by $28. This is reasonable as the effects of a single person’s vote and therefore attendance should not have a large effect. β4 implies that if the dummy genre variable increases by 1 (i.e. the movie is animated) while other factors are held constant, the square root of the film’s revenue will increase by $580,000. This seems like a large effect but is possible given the popularity and wide appeal of animated movies, which are often children’s films and therefore draw families. β5suggests that if the dummy rating variable increases by 1 (i.e. the movie is rated PG) while other factors are held constant, the square root of the movie’s gross revenue will increase by $2,750,000. This makes sense for similar reasons as animated movies. Finally, β6­ predicts that if the interaction between budget and genre increases by 1, the square root of the movie’s revenue will increase by $18,000.

Our final model has an adjusted R2 value of 0.786. This can be interpreted as meaning that 78.6% of the variation in our data can be explained by the independent variables that significantly affect the dependent variable. A disadvantage of our model is the number of independent variables, as that makes interpreting individual parameter estimates challenging. Interpretation is also complicated by our transformation of the response variable. We also found an RMSE of 2.23 million dollars, which suggests that we can expect our model to provide predictions of the square root of a movie’s gross revenue that are roughly within 4.46 million dollars of the true value. This is a relatively large margin of uncertainty, which casts some doubt on the usefulness of our model. We conclude that our model is useful but note that attention should be paid to its limitations if actually relied on to guide investments.

As a rudimentary test of our model we attempted to predict the gross revenue of the animated, PG movie Moana (2016), which had a budget of roughly $150 million, a metascore of 81, 264045 votes on IMDb. Our model predicted a gross revenue of $365 million, while its actual revenue was estimated to be $247 million. This is a non-trivial difference. We then applied our model to the R-rated drama Moonlight (2016), which had a budget of $4 million, and a metascore of 99, 926 votes on IMDb. Our model would predict a gross revenue of $5.75 million, while in reality it earned roughly $27.9 million. This is again a large difference and reinforces that caution is needed if using this model to make any serious decisions.

**Part 2: Future Research**

Our data are limited in that they are only drawn from movies released in 2016. We also note that IMDb users are not necessarily a good representation of the average moviegoer. We discourage extrapolating our model to films with gross revenues well outside of those our data contained and emphasize that, while judged to be useful by typical measures of regressions, the limited sample size and large amount of money at stake make practical implementations of this model difficult to interpret. This analysis could be improved by adding other variables to be considered to the model or choosing different base levels for the dummy variables, or by increasing the number of movies fed into the model.

**SAS Code**

\*/global f test stage 1;

proc reg data=mydata.movies2016 plots=none;

model tr1\_gross = tr1\_budget Metascore votes;

run;

data cutoff;

fcritical=quantile('F',.95,3,129); \*quantile('f', 1-alpha, numerator df, denominator df);

proc print data=cutoff; \*this will output the data table we created with the critical value;

run;

\*/ stage 2 and 3 analysis;

data mydata.movies20166 ; \*create a new data set saved to your library called HWImprove1;

set mydata.movies2016; \*set with the original;

dum\_genre=0; \*create a varaible called DummyCH where every observation will have a value of 0;

dum\_rat=0;

if genre= 'Animation' then dum\_genre=1; \*if the group is CH, then the dummy var will recode to 1;

if rating = 'PG' then dum\_rat =1;

gxr = dum\_genre\*dum\_rat;

gxb = dum\_genre\*tr1\_budget;

lny=log(tr1\_gross);

sqrty=sqrt(tr1\_gross);

run;

proc reg data=mydata.movies20166 plots=none;

model tr1\_gross= tr1\_budget Metascore votes dum\_genre dum\_rat gxb;

test dum\_genre,tr1\_budget,gxb;

run;

\*/ regression assumption;

proc reg data=mydata.movies20166 plots(only label)=(residualbypredicted residualplot qqplot residualhistogram);

model tr1\_gross = tr1\_budget Metascore votes dum\_genre dum\_rat gxb;

run;

proc reg data=mydata.movies20166 plots(only label)=(residualbypredicted residualplot qqplot residualhistogram);

model sqrty = tr1\_budget Metascore votes dum\_genre dum\_rat gxb/ dwprob;

run;

\* / checking for multicollinearity;

proc reg data=mydata.movies20166 plots(only label)=(residualbypredicted residualplot qqplot residualhistogram);

model sqrty = tr1\_budget Metascore votes dum\_genre dum\_rat/ vif;

run;

\* / outliers and influential points;

proc reg data=mydata.movies20166 plots(only label)=(cooksd rstudentbypredicted

rstudentbyleverage);

model sqrty= tr1\_budget Metascore votes dum\_genre dum\_rat gxb/ vif;

run;

proc reg data=mydata.movies20166 plots(only label)=(cooksd rstudentbypredicted

rstudentbyleverage);

model sqrty= tr1\_budget Metascore votes dum\_genre dum\_rat gxb/ vif;

reweight obs. = 33;

run;

proc reg data=mydata.movies20166 plots(only label)=(cooksd rstudentbypredicted

rstudentbyleverage);

model sqrty= tr1\_budget Metascore votes dum\_genre dum\_rat gxb/ vif;

reweight obs. = 85;

run;

proc reg data=mydata.movies20166 plots(only label)=(cooksd rstudentbypredicted

rstudentbyleverage);

model sqrty= tr1\_budget Metascore votes dum\_genre dum\_rat gxb/ vif;

reweight obs. = 87;

run;

\* / final model;

proc reg data=mydata.movies20166 plots(none);

model sqrty= tr1\_budget Metascore votes dum\_genre dum\_rat gxb;

run;

**Appendix**

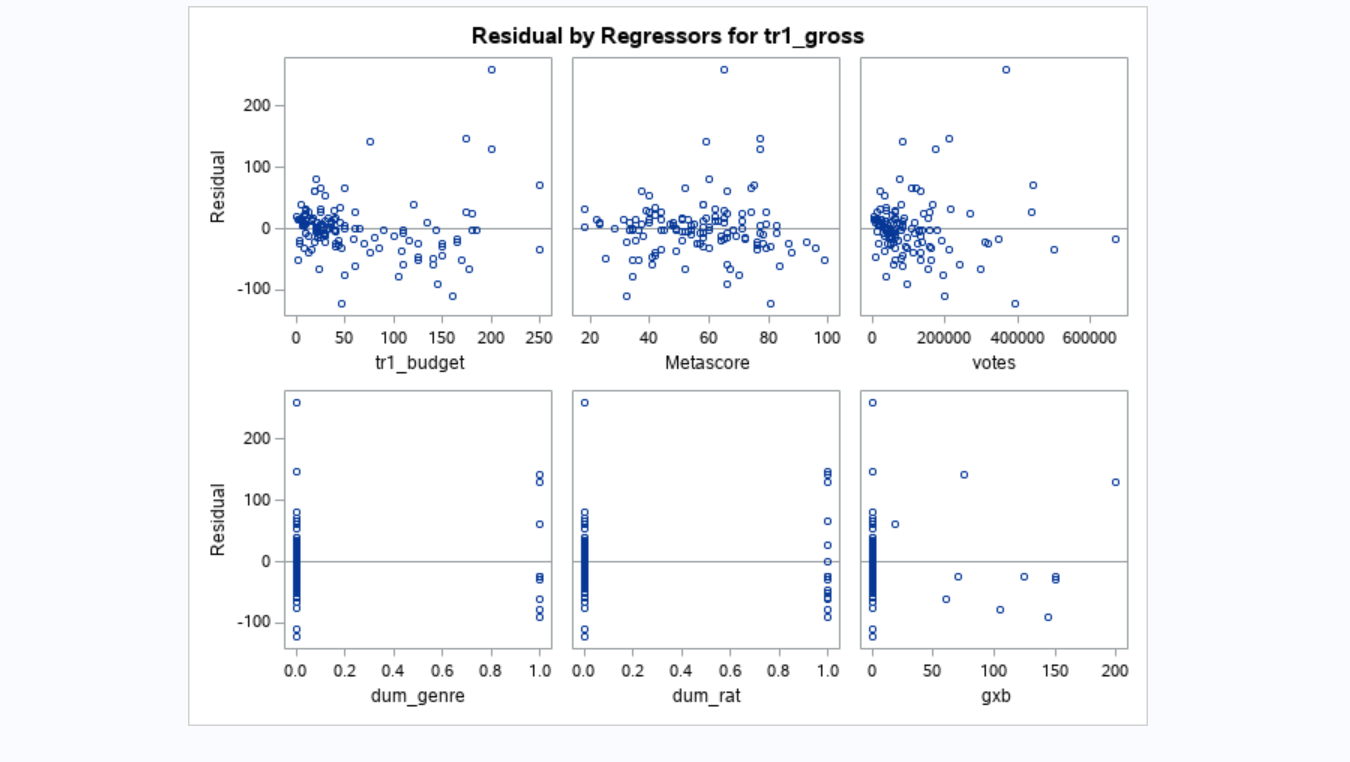
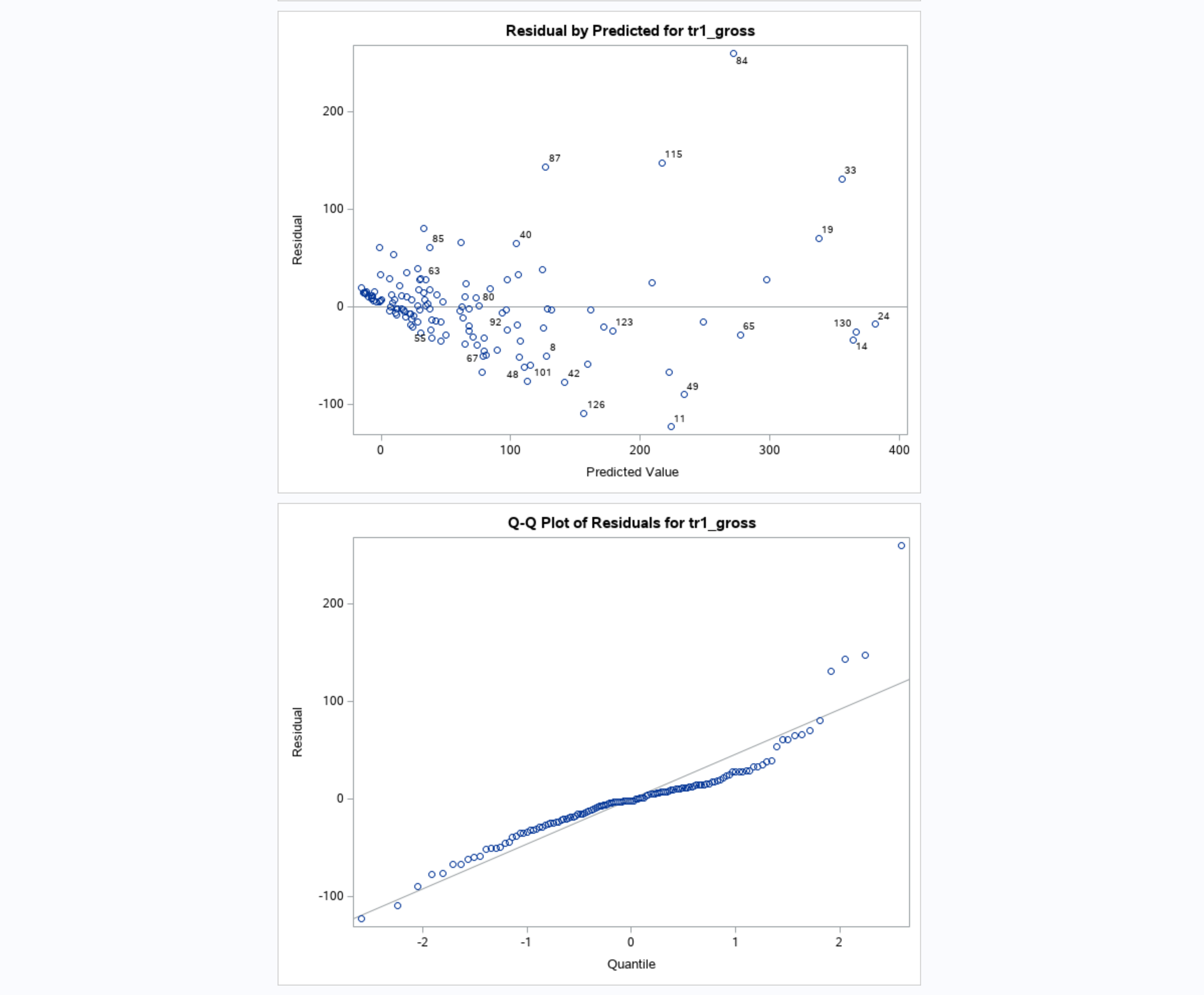


Figure A: Scatterplots of Residuals by Variable

Figure B: Scatterplot of Residuals 

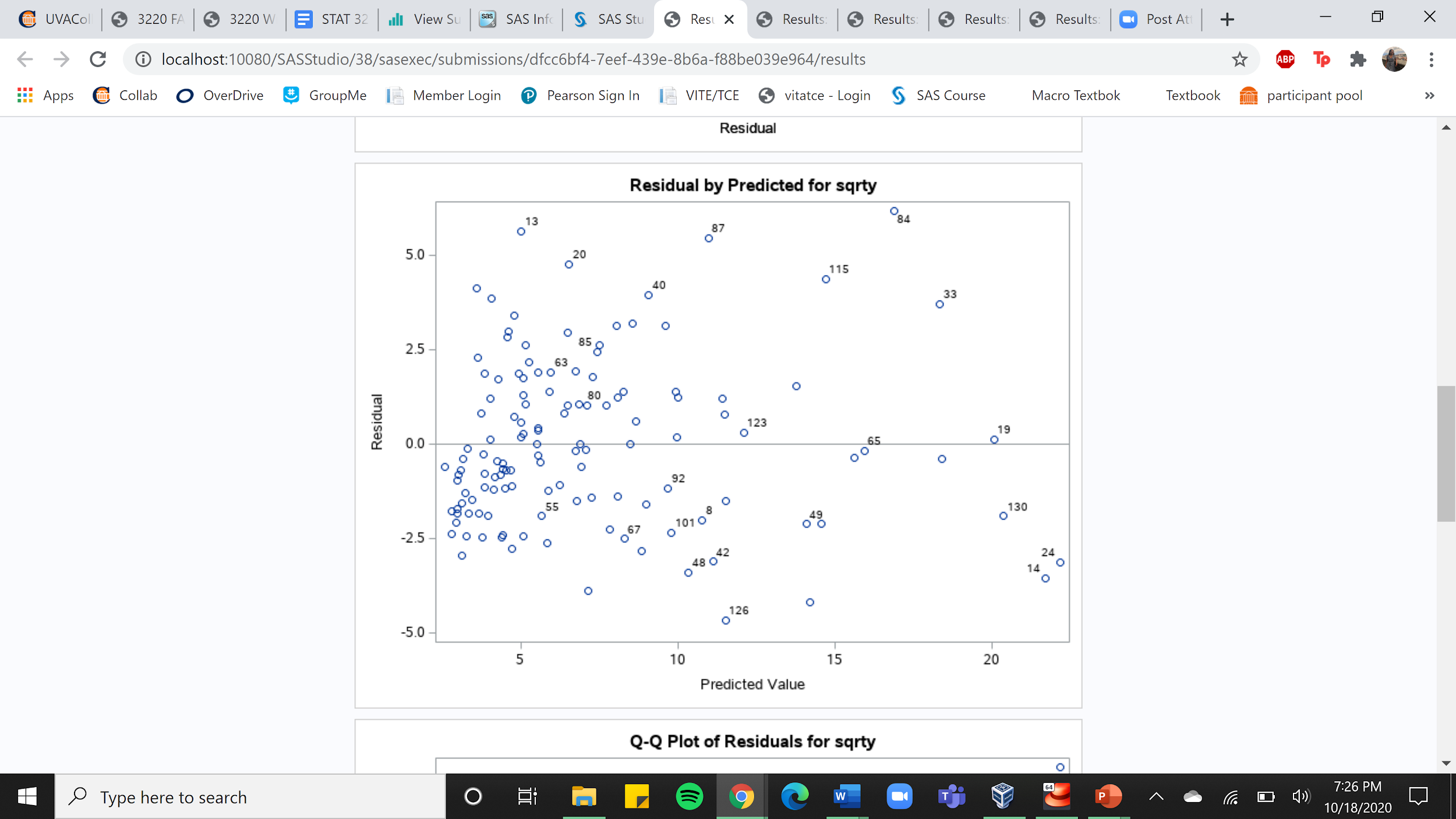


Figure C: Scatterplot of Residuals after Transformation

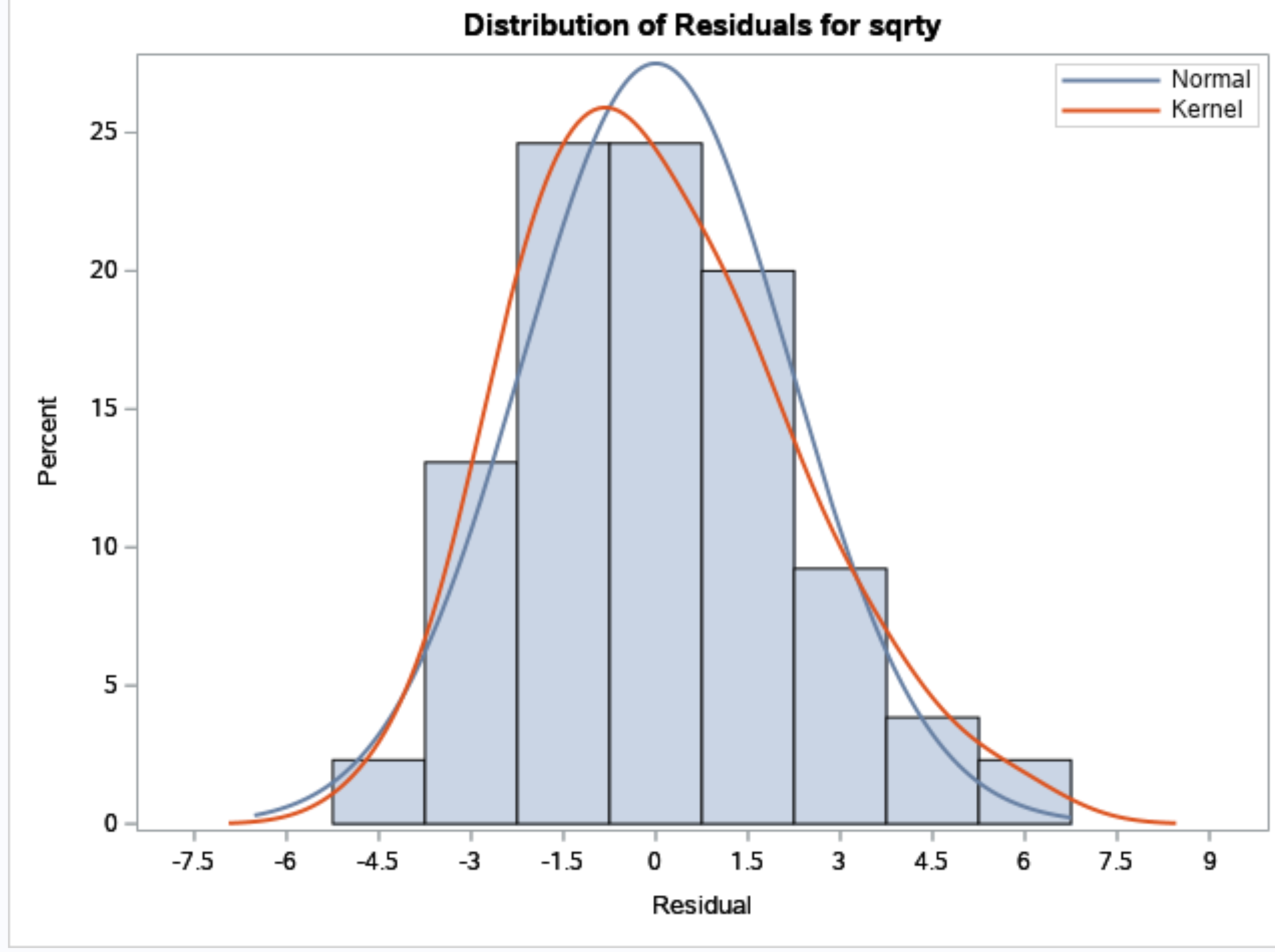


Figure D: Distribution of Residuals

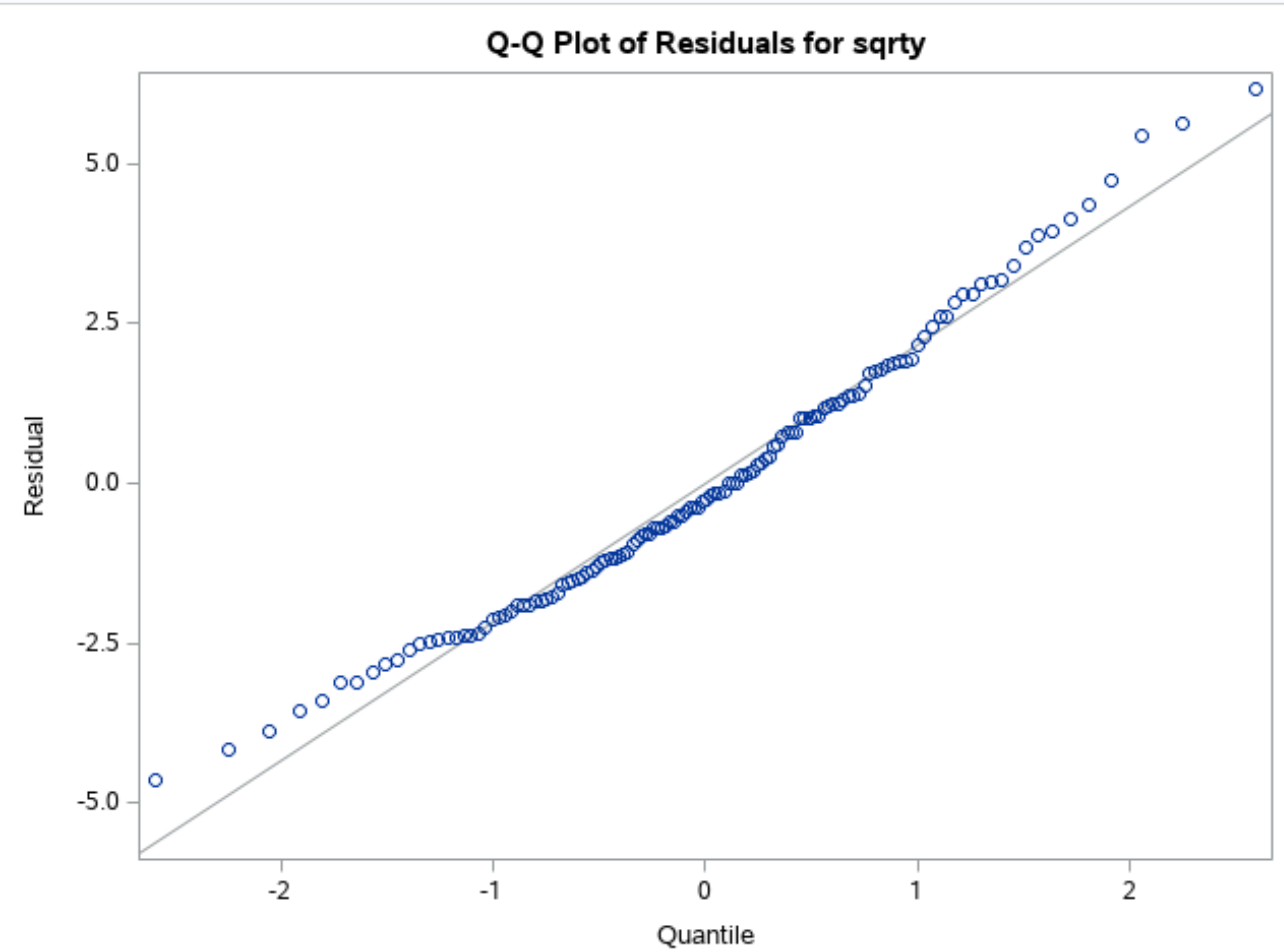


Figure E: QQ Plot

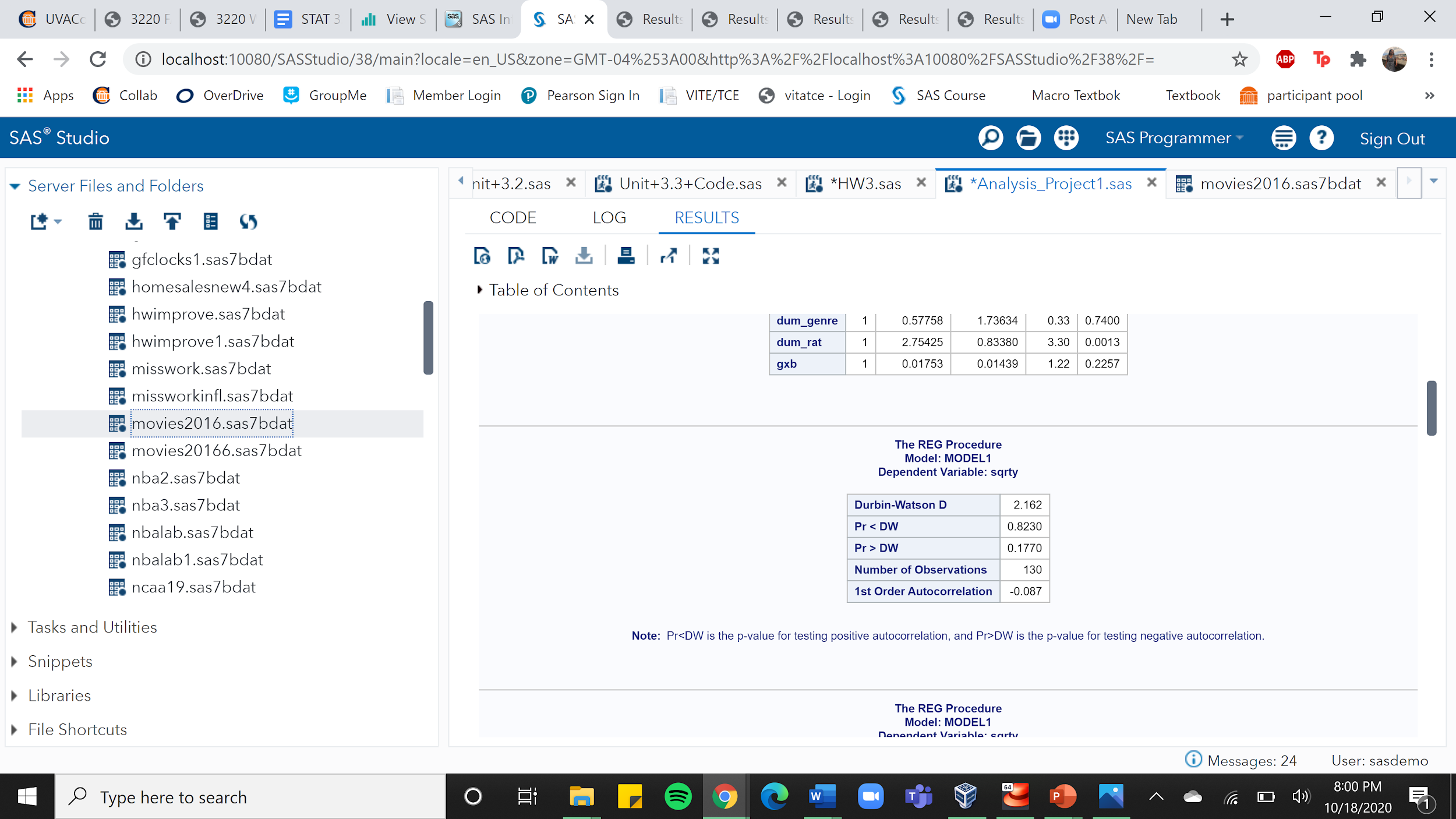


Figure F: Durbin-Watson Test

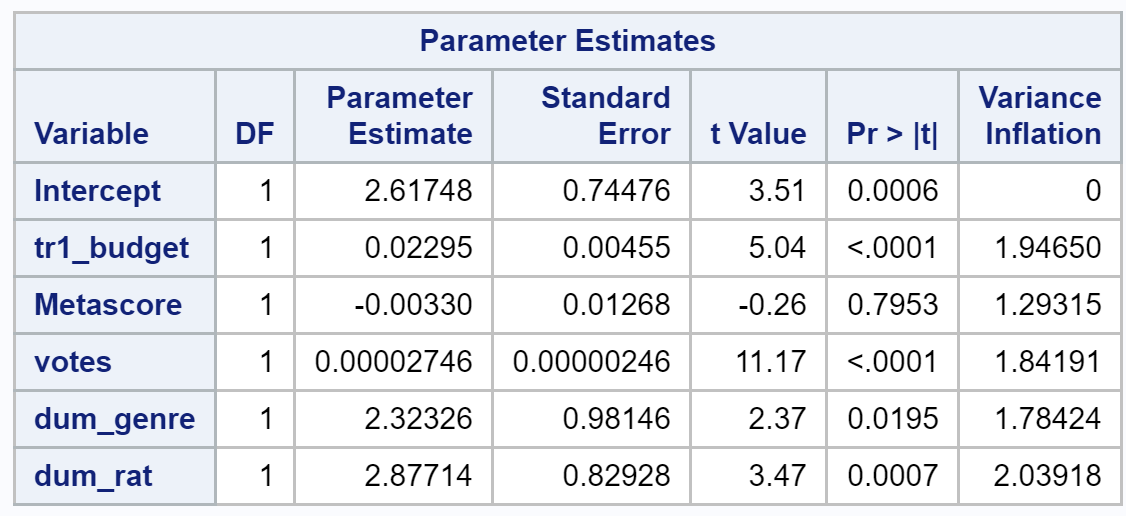


Figure G: VIF Values

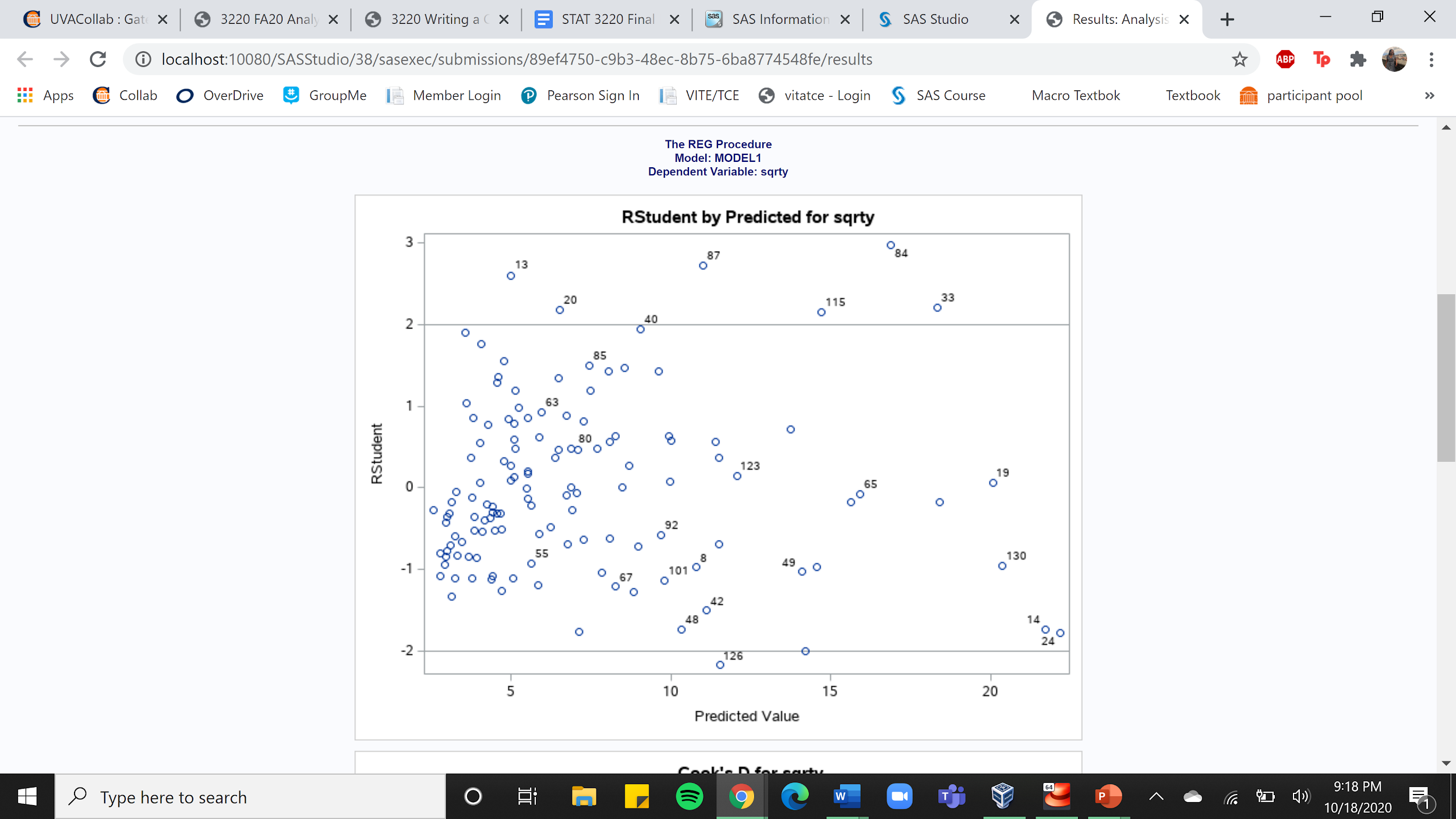


Figure H: Outliers (using RStudent)

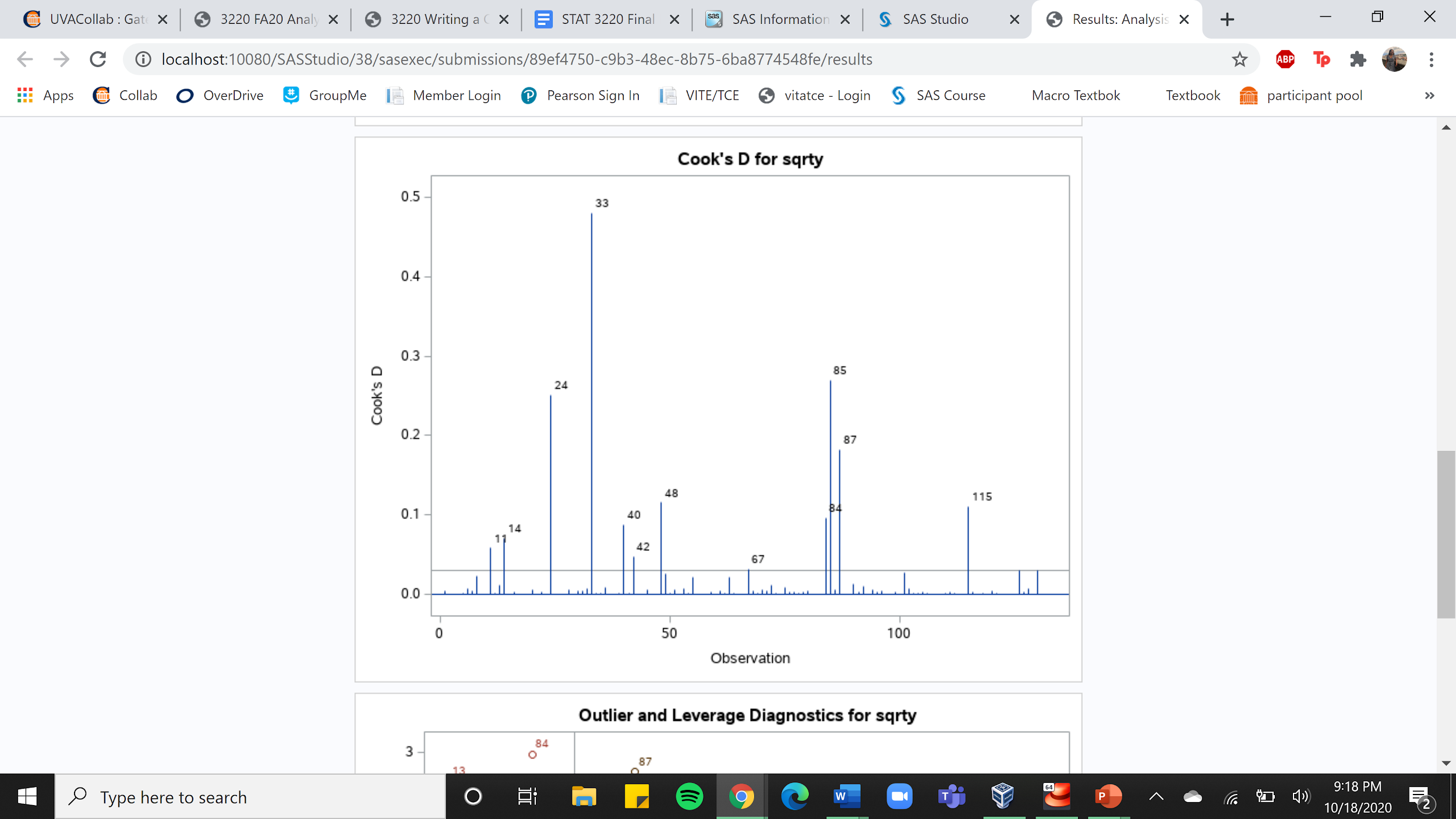


Figure I: Cook’s Distance Statistic Summary

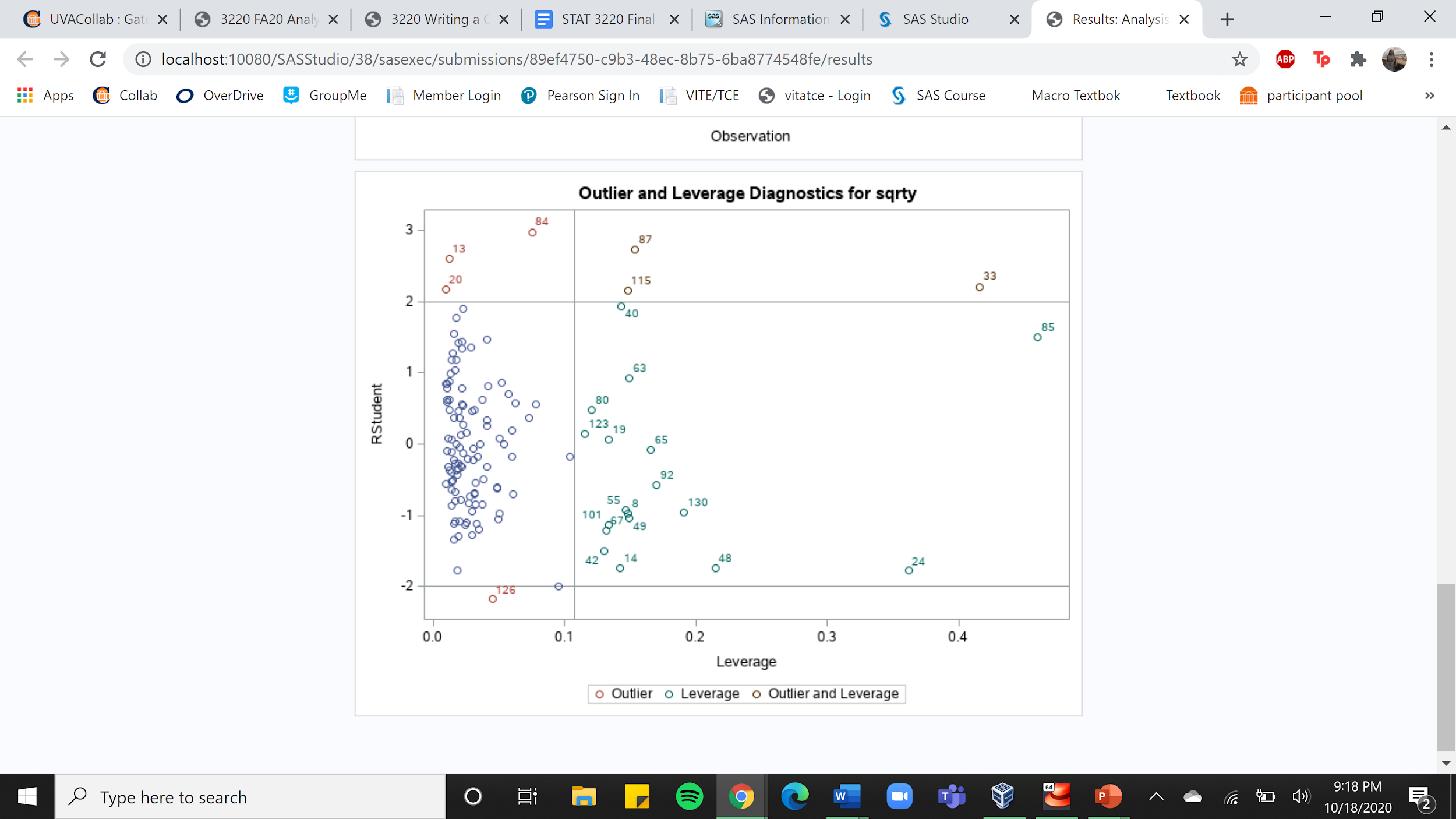


Figure J: Leverage and Outlier Diagnostics

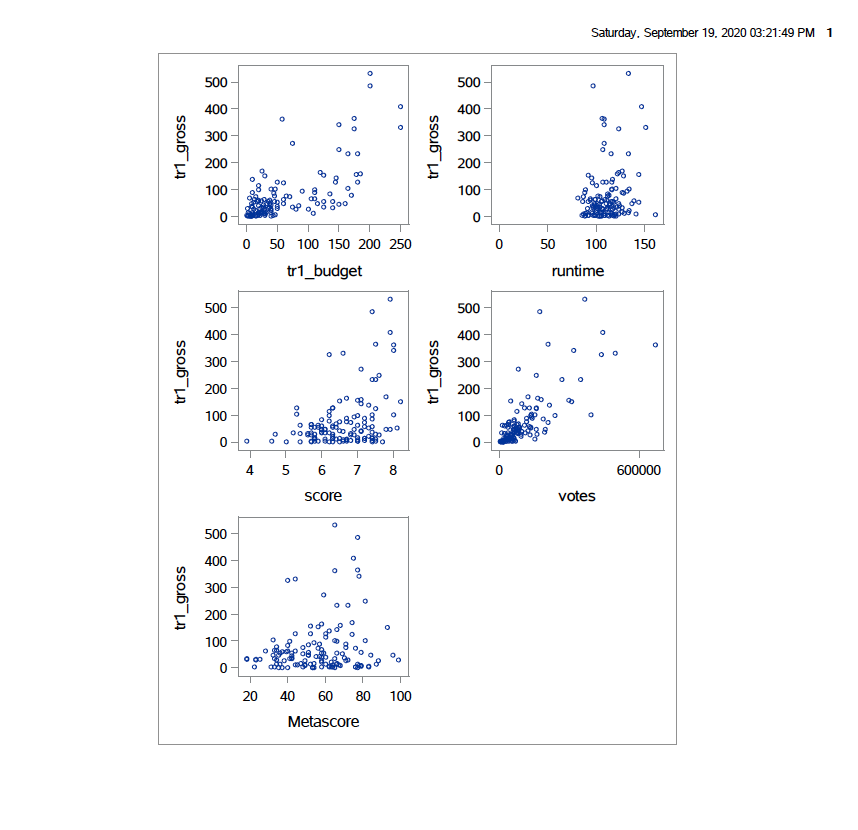
Figure K: Scatter plot of quantitative explanatory variables (budget, runtime, score, votes, metascore) vs response variable (gross revenue)

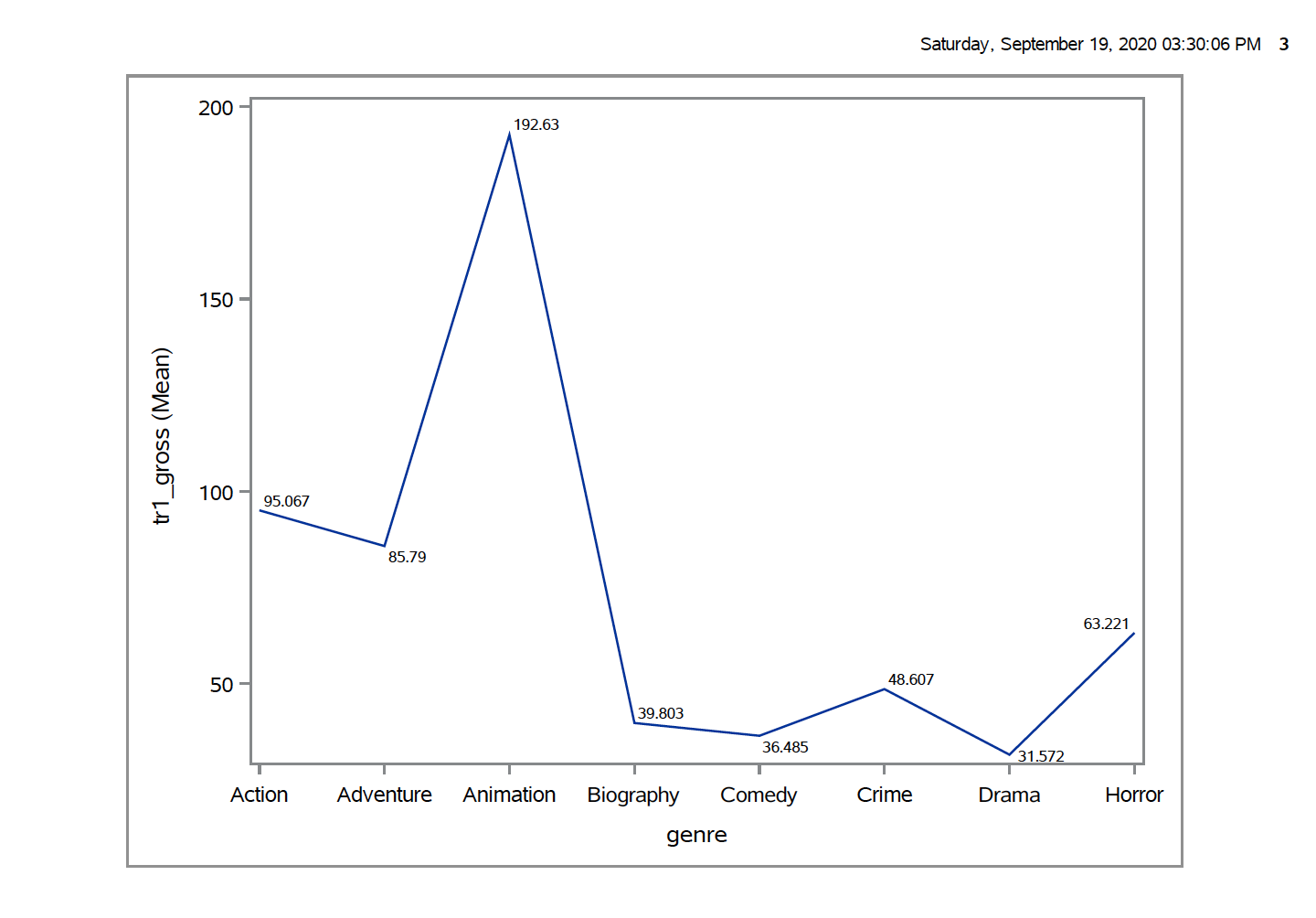
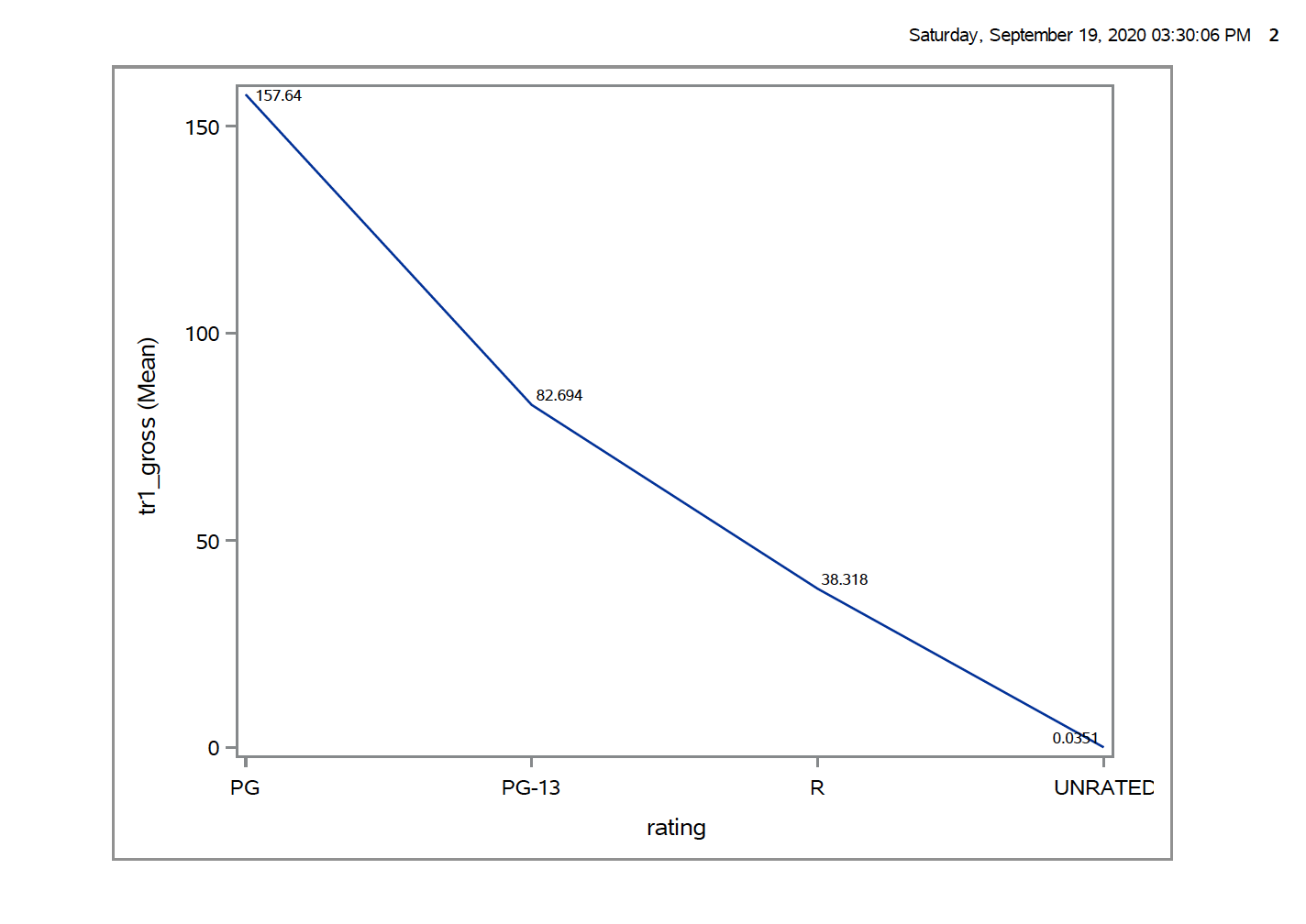
Figure L: qualitative explanatory variable (genre) vs gross revenue

Figure M: qualitative explanatory variable (rating) vs gross revenue

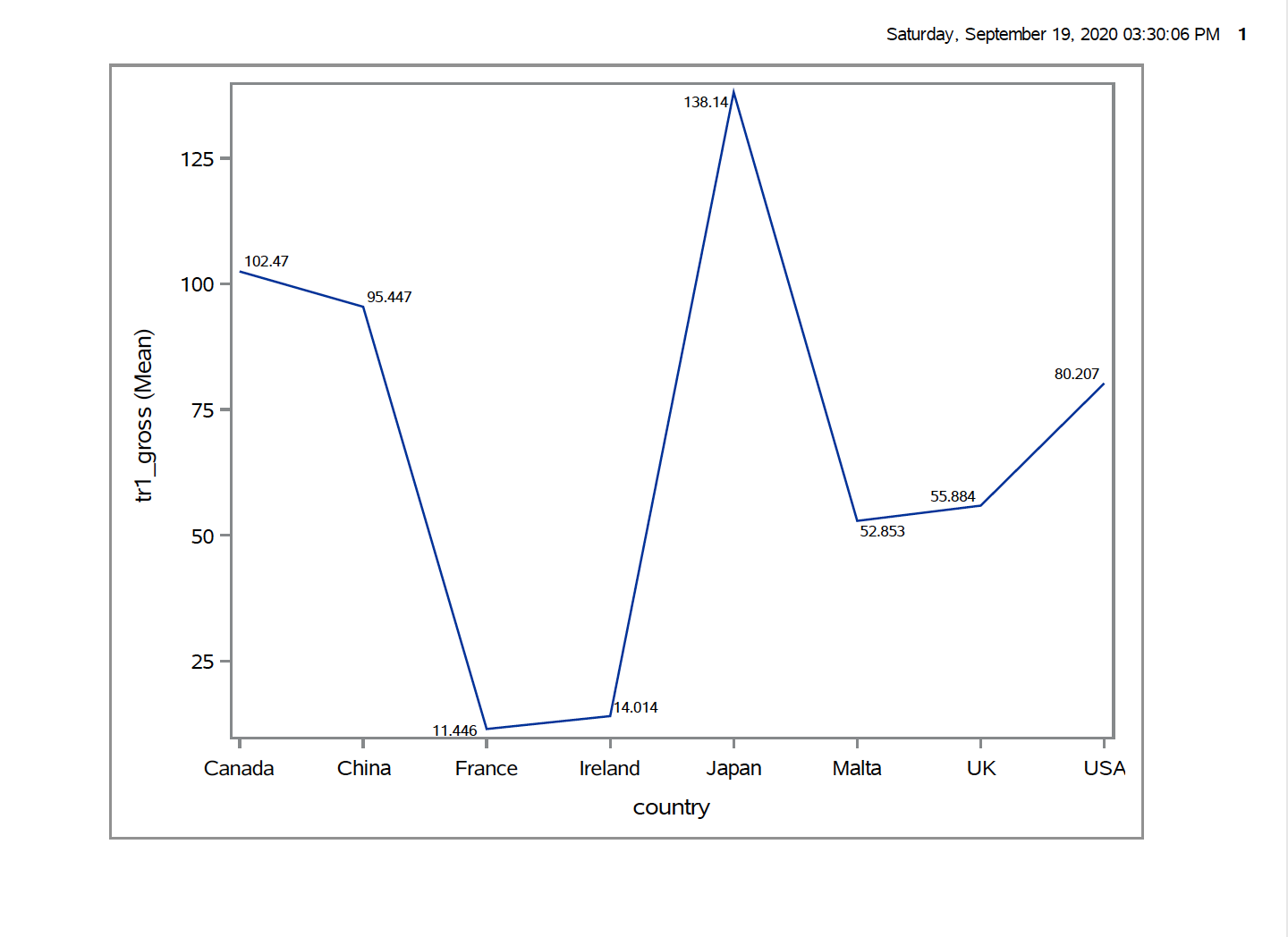


Figure N: qualitative explanatory variable (country) vs gross revenue

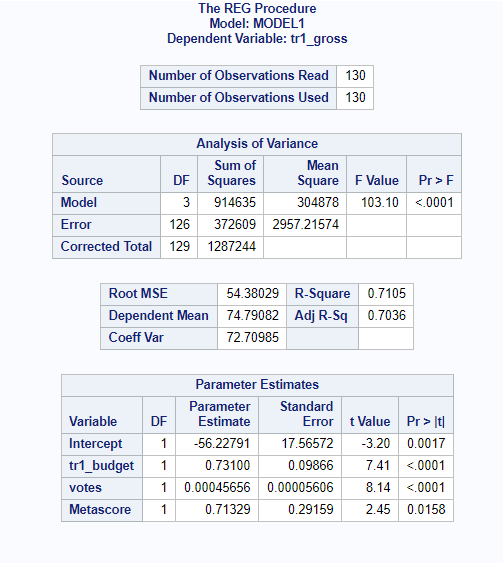


Table 1: Stage 1 and 2 End Model

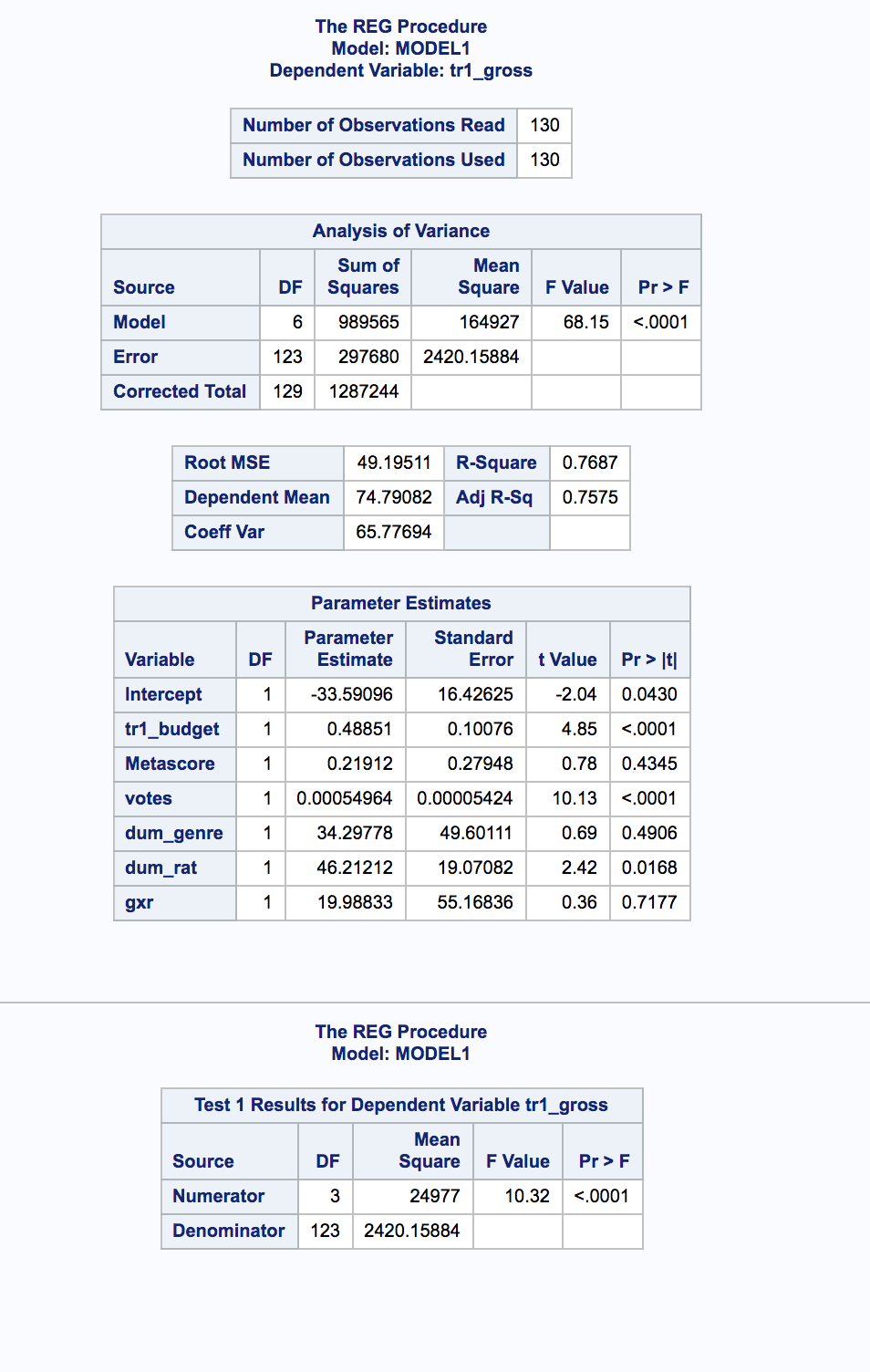


Table 2: Stage 2 Test



Table 3: Stage 3 End Model

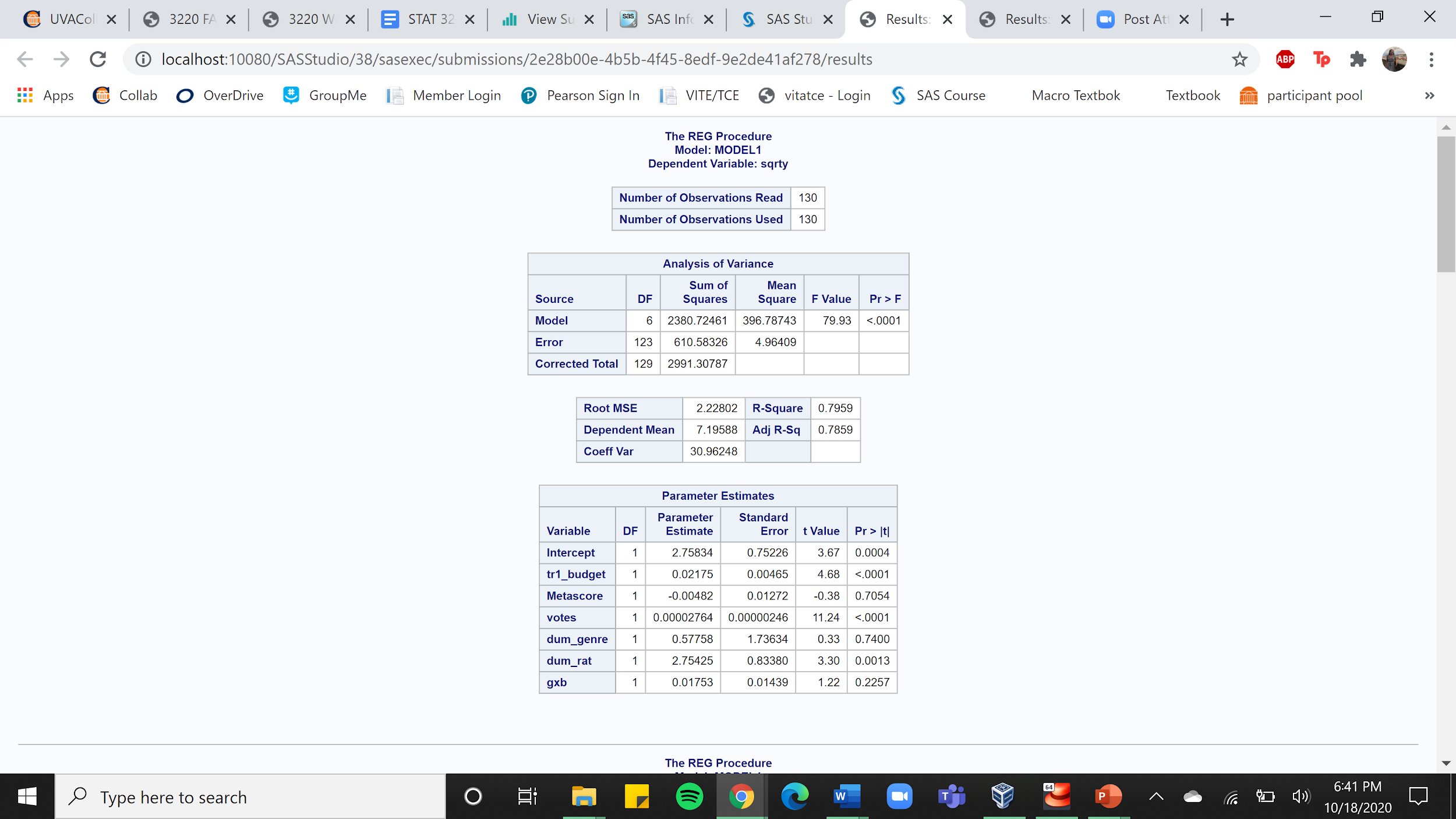


Table 4: Final Model after Transformation

1. https://www.forbes.com/sites/rosaescandon/2020/03/12/the-film-industry-made-a-record-breaking-100-billion-last-year/#4fe5a64034cd [↑](#footnote-ref-0)
2. https://www.diva-portal.org/smash/get/diva2:1211390/FULLTEXT01.pdf [↑](#footnote-ref-1)
3. <https://www.kaggle.com/danielgrijalvas/movies> [↑](#footnote-ref-2)