Analyzing the Influences that Impact the Pre-Course Survey and why the Composite Scores Between Both Surveys had no Statistically Significant Changes

Dhanraj Patel - 1003965168

2023-04-11

Contents

Abstract **1**

Introduction **2**

Data Cleaning **2**

Methodology **3**
Study design . . . 3
Methodologies for secondary research question . . . 3
Methodologies for primary research question . . . 4

Results **5**
Results for primary research question . . . 5
Results for secondary research question . . . 6

Conclusion and discussion **6**

Appendix **7**

Abstract

The main purpose of this study is to investigate if an R manual was effective in teaching a class the R programming language with no prior assumed programming experience before. In order to do this the students in the class were asked to fill out surveys in the beginning of the course and about the half way through the course. The questions were all ordinal and inquired about their past programming experiences and their current understanding of R at the point of taking the survey. Using the technical questions we can calculate a composite score that represents the students R proficiency and compare them between the two surveys to compare proficiency progress. Due to low sample size in the second survey we used a non parametric test to compare the surveys and found no statistically significant deviation in the central mean with a P value of 0.06608 (greater than 0.05). In addition, using linear regression and step-wise for model selection, we found out which of the other questions in the survey most contributed to the composite score and to what degree. From our results we found that the most important variables were R previous experience, R confidence level, university level previous statistics experience and high school level previous statistics with all these variables having a P value below 0.05.

Introduction

As the world becomes more technologically advanced and data-centric, learning how to program in R is becoming an increasingly vital skill. However, with the vast amount of associated learning resources and the endless amounts of new additions to the language, learning R may seem like a daunting endeavor. To alleviate this issue, the University of Toronto Department of Ecology & Evolutionary Biology has created an R manual to help students learn the R language for students with varying amounts of R and statistical prior experience. To identify if the manual is an effective teaching tool the faculty have been running a test pilot in a second year ecology course. In this course, students will be required to use the R programming language and will be given the R manual as an aid to help them learn R. In addition, it is important to note that the students were not required to have any R or any programming knowledge prior to taking the course, but some students are previously experienced. As the students are progressing through the course with the R manual, they will be asked at two points to fill out surveys about their past experiences, their current comfort in R and their experience using the R manual. The first time the students will take the survey will be at the beginning of the course, while the second time the students will take the survey will be about half way through the course. By taking these surveys the main goal of this study is to utilize various statistical methods to answer the primary research question of identifying if the R manual effectively improves students' R literacy skills. As a result we will be using non-parametric test that will allow us to compare the survey results without requiring the underlying population for assumptions to be met which would be the case with low sample size. In addition, it is important to note that all students were able to submit their beginning of the course survey answers which we will utilize to answer our secondary research question of what factors in the pre course survey influenced the students R language proficiency. As for the secondary reseach question it is important to clarify that it will measure the students proficiency in R at the begining of the course and not include proficiency gained throughout the course. We will be measuring the proficiency from the technical questions asked in the survey by calculating a weighted composite score based on the difficulty of the specific technical skill and then fitting the composite score as a response variable in a linear regression model with the other survey questions as predictors. From this we will be able to identify which and to what degree that student's pre class experiences (e.i. past programming or statistical classes) influenced their composite scores.

Data Cleaning

In this study we will be using two sets of data with the first data set being the survey done right at the beginning of the course, we will start by going over how I cleaned the first data set which included the survey done right at the start of the course. The first thing I did the to clean the data was to remove the questions ask for text responses. While these, questions are important to gain a deep understanding of the thoughts of the students in regards to R and the R manual, they will not be used in this study so they were removed. The main reason for removing these questions has to due with the fact that they had a low frequency answer rate and are difficult to analyze in a statistical setting. The justification for this step is that these answers are all written subjective accounts that would not be useful in an analysis. The last issue that needed to be cleaned is the fixing the formatting produced by excel so the data is more readable and usable in R. The missing data existed in questions that were follow up questions to another question. For example students were asked "Have you ever programmed in another language other than R?" and as a fallow up were asked "Which language?". The students who answered 'yes' to the initial question answered the secondary follow up question, but students who answered 'No' to the first question let the follow up question blank. This situation appeared a few times and for each, I determined the best way to deal with this is to fill in None to the missing answer. This is mainly because a non answer in this case is a viable answer and so to represent this answer I choose to write None to represent the missing but still viable answer. The main reason I did this is to have a complete set of data to apply statistical methods onto, missing data would hamper the analysis.

In addition, a second data set is used in this study which contains the matched data of the students from the survey done in the beginning of the survey with the survey done at the middle of the course. For this data set the only step that was done to clean the data was renaming the column headers exactly like how was

done for the first data set to improve the readability of the study and make using the data easier in the study. This concludes the data cleaning sections.

Methodology

In this section we will be discuss the methodologies used to investigate our research question.

Study design

Before diving into the methodologies its important to remind ourselves with what the main goal of the study is and how we can tailor our methodologies to best adhere to these goals. In this study students in an ecology course will be working in R throughout the course and have been given an R manual to aid them in learning R. Using the students survey answers from the beginning of the course and at the midpoint of the course using various methodologies our goal is to examine if the R manual has helped the students improve their R skills using data from the surveys which contain ordinal data. The questions in the survey included questions related to the students past experience with R and other programming languages, their own personal comfortability with R including their confidence in their proficiency with various R skills and questions related to their R learning process. Unfortunately, only 14 students were able to submit their mid course survey which influences the statistical tools we will be able to use. From this data, we will be able to answer our primary and secondary research questions. We will first begin by going over our methodology used to answer our secondary research question.

Methodologies for secondary research question

We will first start by going through the methodologies used to answer the secondary research question.

Choosing a statistical model

For this study we will be using linear regression as the main statistical model to examine the factors that most impacted the students proficiency in the first survey. The main justification for using the linear regression model is that we will be able to directly investigate the relationship between the response variable which will measure if the students self reported R proficiency skills with the explanatory variables which will measure the other factors that impacted their R learning. These factors include past programming and statistical experiences/confidence and questions about their learning process. In addition another advantage of linear regression is that we are able to isolate each explanatory variable and measure the outcome while controlling all other variables. With model selection and diagnostics we can identify the most important factors influencing their R proficiency score in the first surveys. In addition, using the coefficients in the model we can investigate the degree that the various factors influenced the students R proficiency score and so for these reasons this study will be utilizing linear regression modeling to answer to secondary research question.

Going into more details, we will first start by outlining what the response variable of the linear regression model will be. For this study the response variable will a numerical value indicating each student's overall R literacy score. It important to know that at the time of this survey the students were not yet taught any R or statistics so the answers are all based on their previous experiences. In each survey the students were asked to rate their confidence on a one to five scale in regards to their proficiency with multiple R literacy related questions. These questions included their abilities to import data, make scatterplots, calculate summary statistics of data using R and calculate the mean of a distribution using R. By analyzing their answers to these proficiency we are able to calculate a composite score that reflects their overall proficiency in regards to their R literacy. In the process of calculating a composite score for each student, it is important to keep in mind that the difficulty in each R skill varied so different weights need to be added to each skill to calculate

an accurate composite score for each student. The metrics we will use and their weights were decided as a group and are the following:

- Importing data: weight is 10 percent
- Making Scatterplots: weight is 20 percent
- Calculate summary statistics of data using R: weight is 30 percent
- Calculate the mean of a distribution using R: weight is 40 percent

In order to decide on the values of the weights we came together as a group and utilized our previous experience in R to decide on the proportions of the weights by analyzing how difficult each of the tasks were relative to each other. To give more details on how the questions were weighted, some of the considerations that were utilized included ease of understanding and statistical and statistical knowledge required to be confident in the questions. For example, learning how to important a package in R is quite easy and does not require any previous knowledge while calculating the mean of the distribution requires understanding what a distribution. The more difficult it was to be confident in each task the higher the question was weighted. This method introduces the limitation of subjectivity into our study due to the fact that other groups may have chosen to weight the tasks differently. Using these weights we are able to calculate a composite score the first survey which will be used as the response variable.

Moving onto the explanatory variables, these variables will include data from the other questions the students answered. These questions have to do with their past programming and statistics experience, their learning methodologies and overall confidence with R. We will begin by fitting a model with all these explanatory variables and use model selection and diagnostics to identify the most important factors influencing the students R proficiency (more on this later). Through these methods we will be able to finalize our linear regression model.

Linear regression assumptions

In order to ensure that the results of the linear regression models are correct we need to ensure that the assumptions of the linear regression models are followed. The assumptions of the linear regression models are the following:

- Normality of residuals
- Linearity of the data.
- Homogeneity of residuals variance
- Independence of residuals error terms

Ensuring these assumptions are valid is a crucial step in ensuring that the results given in the model are both accurate and generalizable. In order to check the validity of the normality of residuals, linearity and homogeneity of residuals variance we will be using diagnostic plots. These plots are available in the appendix. As for independence of residuals error terms we can verify this assumption using a VIF (variance inflation factor) score which will measure the multicollinearity between the terms.

Model selection

In order to identify which of the explanatory variables had the greatest influence on the response variables we will be using various methods for our model selection. We cannot include all the possible explanatory variables as some variables may not be strongly correlated to the response variable and thus should not be included in the model. To find the right balance between the number of variables we will be use the step-wise model selection method. This method identifies the best models by cycling through every possible combination of explanatory variables by calculating the AIC and BIC values of each model which can be thought of as a measure to measure the goodness of fit of the model parameter. The lower the AIC and BIC values the better the model is comparatively.

Methodologies for primary research question

Moving onto the primary research question we will be comparing the composite scores between the the two surveys. We will be calculating the composite score for both surveys in the same manner as discussed in the

last section. With the following questions are weights.

- Importing data: weight is 10 percent
- Making Scatterplots: weight is 20 percent
- Calculate summary statistics of data using R: weight is 30 percent
- Calculate the mean of a distribution using R: weight is 40 percent

One main issue to keep in mind is that for the second survey we had only 14 students responses matched to the first survey. Due to the issue of a small sample size we will be using a non-parametric test due to the fact that we do not need certain assumptions to be held true to analyze the tests results. This is crucially important as the low sample size would lead many assumptions in statistical methods that include assumptions to fail. The specific test we will be using is the Wilcoxon signed rank test as it is a non-parametric statistical test used to compare two related or paired samples to determine if there is a significant difference between their central tendencies. In this case we will be using the test to identify if the means differ between the composite scores between the two surveys. After using the test we will be able to use numerical and graphical summaries to gain further insight on how the composite scores changed from the beginning of course survey to the mid course survey.

Results

We can move onto the results of the study.

Results for primary research question

In order to find out if the change of the composite scores changed between the beginning of class and mid class survey we used a non-parametric paired Wilcox test due to the fact that our sample size is small. The test resulted in a P value of 0.06608 which is greater than 0.05 that from using the Wilcoxon signed rank test we have found that there is not a significant difference between the central mean tendency between the two surveys composite score. As such we can infer that that the for the students who were able to submit data for the second survey that the R manual did not play a significant role in helping them learn R as the composite score did not increase by a statistically significant amount. Although we do need to keep in mind that only 14 students were able to submit data or the second survey so their results might not be indicative of the class. The students only represent roughly 10 percent of the class so an overwhelming majority of the class is not represented. In addition, due to the low sample size the power of the results will also be very low. The power refers to the accuracy of finding a significant statistical difference (with the p value correctly being calculated below 0.05) and with low power the accuracy of the results can not be seen as reliable.

To gain further insights on the composite scores we can compare them directly using numerical and graphical summaries.

Table 1: Composite score summary

Meaure	Q1	Median	Q3	IQR	Mean	Trimmean	Var	SD
Survey one	2	4	5	3	3.48	3.56	3	1.64
Survey two	4	4	5	1	4.03	4.17	1	1.11

From table_1 we are able to compare composite scores between the first and second survey. From the table we can see that survey two had a mean and trimmed composite score higher than survey one. This makes sense as the students become more proficient in R as they continue to use it throughout the course and thus raise their confidence levels. However, despite what we see from the table due to the low power caused by a low sample size the results can not be seen as reliable.

Results for secondary research question

After determining that the composite score did not have a statistically significant deviation we can move on and answer our secondary research question of understanding the quantitative and qualitative factors that the composite score of the first survey. To do this we need to look which variables were selected by the step-wise model and the values of the coefficients. There may have been a small increase in the mean, but it is likely not significant as the test has showed us.

After fitting all the predictors in a linear regression model and using step-wise model selection the following predictors were selected as the final variables. These 4 aspects most contributed to the students overall composite score.

- Student has taken a statistics class in high school (or equivalent)
- Student has taken a statistics class in university (or eqivelent)
- Student Has previous experience in R
- Student is confident in R

Table 2: Coefficient values in final model

coefficients	Coeff_vals	Error_rate	P_VALS
intercept	0.29727	0.16055	0.0663000
Highschool exp statistics	0.15913	0.28287	0.5747000
Uni experience Statistics	0.39726	0.17811	0.0274000
R experience	0.45925	0.08675	0.0000005
R confidence	0.44677	0.10147	0.0000216

In table_2 we are able to see the factors that most influenced the composite score for the proficiency for the first surveys composite score. From the table we can see that the self reported R experience had the highest influence on the composite score with a coefficient of 0.45925 As for an more detailed explanation of the coefficient value, by controlling every other variable each increase in experience scores increases the students proficiency score by a factor of 0.45925 This makes sense as those more experienced in R well be better equipped to handle technical R questions and thus result in a high composite score. In a similar vain, the students confidence in R also was found to be a significant predictor with a p value of 0.0000216 and a coefficient of 0.44677. From this we can surmise that one score unit increase would increase the proficiency composite score by 0.44677 (with other variables controlled). This also makes reasonable sense as those who are confident in their R skills probably have previous R experience so for similar reasons as before it makes sense that it is a significant contributor to a higher composite score. Moving on after preforming the step-wise function the last 2 variables that were found to be significant were if the students had university level statistics level experience and high school level statistics experience in that order, with coefficients being 0.39726 and 0.15913 respectively. It makes sense that previous statistics experience results in a higher composite score as in statistics R is widely used and thus would increase their composite score for similar reasons as the R experience variable. Also, many of the R proficiency questions in the survey required statistical knowledge for example finding the mean of a distribution. In these cases, having previous statistical knowledge would help increase the students composite score as they would be more familiar with the questions.

Conclusion and discussion

After analyzing the data we are able to conclude for our primary research question that from the data available that the composite scores did not statistically significantly differ from first beginning of course

survey compared to the second mid course survey. We used a non parametric Wilcox paired test to determine due to the reason that our sample size was small with only 10 percent of the students with results from the second survey. Moving on, when examine the numerical summaries of the two surveys we noticed that the first survey had a lower mean and trimmed mean and a higher variance and standard deviation. This makes sense as when the students took the second survey they had spent a few months learning R and thus had time to hone their skills and become more proficient in R raising the second surveys composite score. However, it was not raised enough to be statistically significant. The larger variation in the first survey can be explained by the fact that the survey only considered pre course knowledge which is more varied. Moving onto the secondary research question, when focusing on the first survey we found that the most significant factors impacting the composite score is the student's R experience, R confidence, University level statistical experience and lastly highschool level statistical experience. This also makes sense as more experience in R will lead them to getting increasingly more proficient in R and thus result in a higher composite score.

After attaining all our results it is important to reflect on how the results compare with the overall reseach questions and any implications it might have in the grander scheme. As for the primary reseach questions we found that previous experince with R and statitics boosts the composite scores of the students. This is an important distiction as it allows us to identify an important confounder. If the main goal is to see if the R manual effectivly teaches R, then we need to take into account the previous experience of students to identify if imporovment is based on their experince of the R manual. To recitify this we can either controlling for experience by ensuring all students in the study have the same level of past experience. As for the secondary research question we are not able to generalize our data based on the low match number causing low power an loss of normaility.

Limitations

In our study a significant limitation arose in the fact that the second survey only had 14 students including in the data. This lead to a lot of statistical tools (Glms) becoming unusable. Also it leads to a loss of generalizability and only 10 percent of students were represented in regards to our primary question. We also were not able to find out what factors and characteristics contribute to the second surveys composite score like we were able to do with survey one data. In addition it led to a loss of normaility in our data and a low amount of power in our resuls leading them to be unreliable.

Another limitation we had was that the calculation of the composite score was subjective based on our personal experiences and judgements regarding R. If other reseachers were consoluted, they might have differing weights for each question which might impact the results. .

Next steps

For a repeat study it is imperative that full data be obtained for the second study. Also to test the R manual specifically making sure that each students starts off at the same level of experience would help control confounders. With a large data set the issues regarding normality, low power and a lack of generalizability would be resolved. We would also be able to use more robust an powerful statistical tools to get more accurate results.

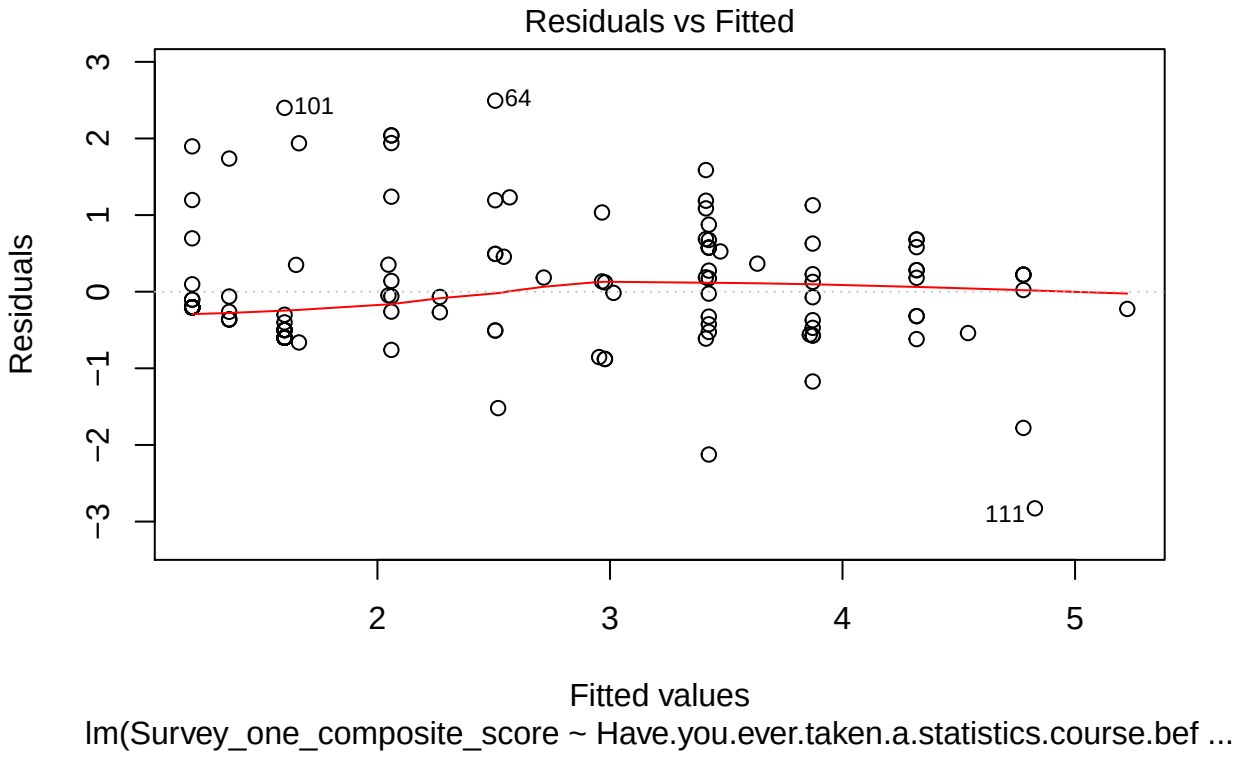
As for the subjectivity in the composite scores, in futur studies it can greatly help to use a wide variety of different weights and include new questions to see of the results are similar accorss the board. This will help limit subjectivity and create more reliable results.

As for the addition of new questions, it would be helpful to include questions regarding the specific amount of times the R manual was used and in addition it would help for the answer to be numerical (non ordinal) as that would help us control confounders and get a better idea of the efficacy of the R manual.

Appendix

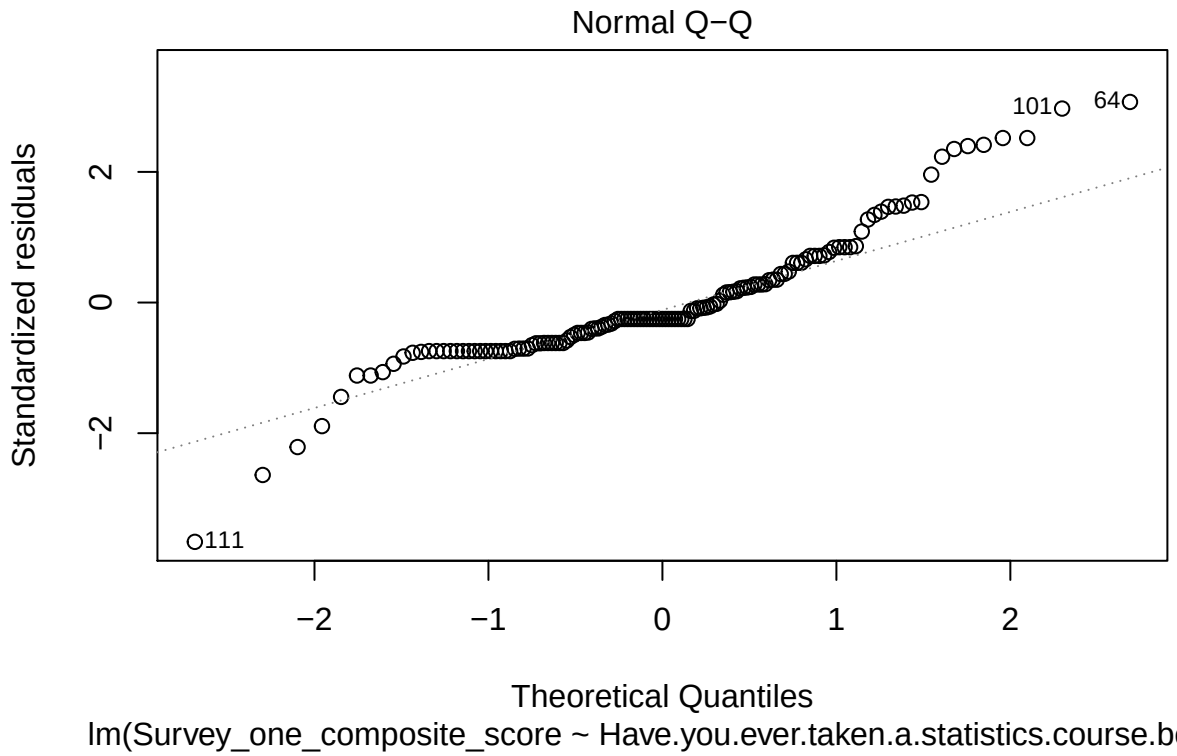
The following are the diagnostic plots for the final linear regression models assumptions.

```
plot(b,1)
```



As the residuals vs fitted plot line seems mainly horizontal we can conclude the data is mainly linear.

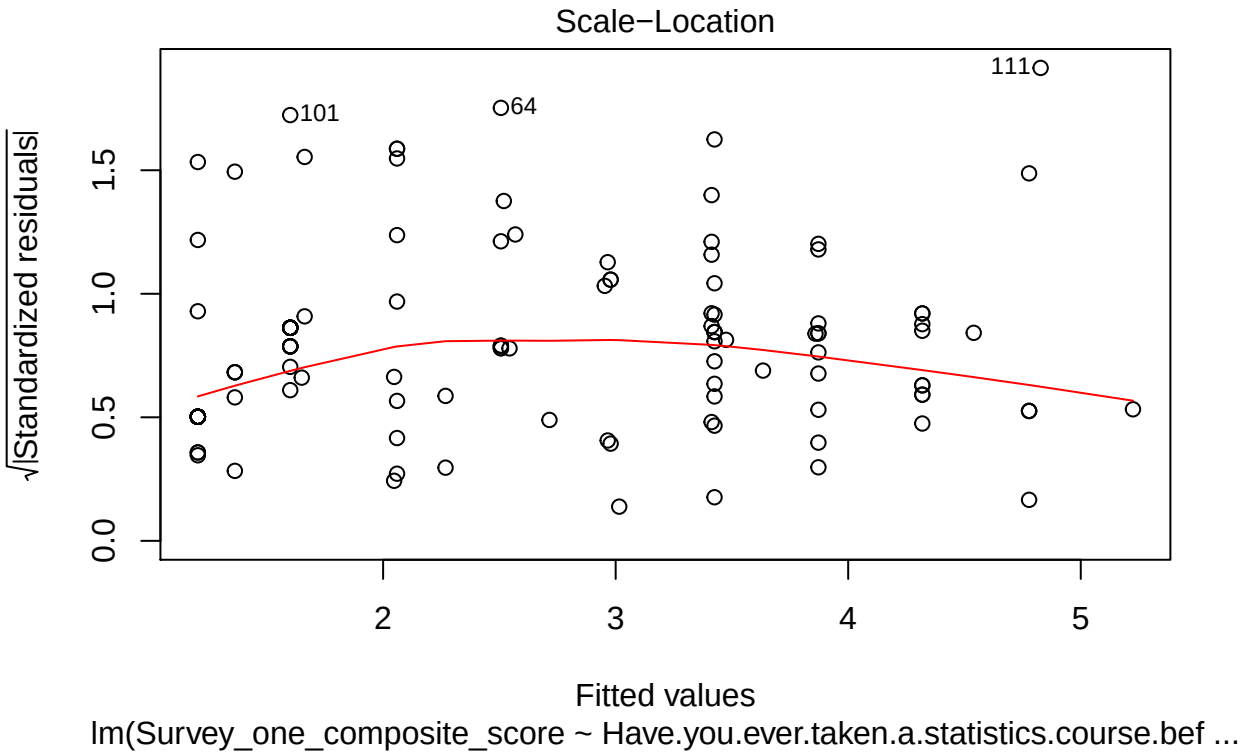
```
plot(b,2)
```



From the qq plot we see that the middle dots follow the line, but veer off dramatically on the ends. We can

see that the data is mostly normal.

```
plot(b,3)
```



From the plot we see that the line is noticeably curved downwards. The constant variance assumption is not completely broken but not strongly followed as well.