Exploring the Factors that Influence Self-Perceived Quality of Life

Group D

Jack Lewis, Jacob Sandish, and Rowan Olson

Abstract:

The goal of this project was to use General Social Survey data from adults in the United States in 2021 to create a multi-linear regression model and make inference about the self-ranking of individuals within society based on socioeconomic factors. Backwards elimination was used to test the initial predictor variables and a final model was constructed using marital status, education level, and income. The model can be used to make inference on a larger population because it meets the Independence condition for inference. However, it is limited to predicting these self-ranking values for those in the United States in 2021, meaning other geographical regions and time periods are not applicable.

Introduction:

There are a lot of different factors that could determine whether a person is happy with their life or not. These factors might be different for everyone, but finding common variables could produce potential benefits. An improved understand of the factors that contribute to well-being could help people make future decisions that might improve their own quality of life. The research could also provide insight to policymakers who could use the information to design policies that support happiness. Studying the cause of happiness could help medical professionals develop more effective plans on improving and combating mental health by reducing stress and anxiety. There are an infinite amount of potential benefits towards society if a correlation is found, but these are just a few that inspired us to conduct the study. This project had an end goal of determining if certain aspects of a persons life could predict how they would rate themselves on a 100 point scale.

Data:

The General Social Survey (GSS) is a sociological survey that has been conducted every 2 year since 1972. It is designed to measure attitudes, behaviors, and opinions of Americans on a wide range of topics. It is funded by the National Science Foundation, and is conducted by the University of Chicago. Our data set was collected from GSSdataexplorer.norc.org, which is a publicly available resource that compiles data from the GSS and allows researchers to pick certain variables to study.

The first step of data wrangling that we completed was to read in the data and replace some errors in the responses to "NA". We also removed variables that weren't suitable for further research, and renamed confusing variable names.

```
#Import Data
GSS = read.csv("GroupD-SelfRanking-Data.csv",
na = c(".i:  Inapplicable", ".n:  No answer", ".d:  Do not Know/Cannot Choose",
       ".s:  Skipped on Web", ".x:  Not available in this release",
       ".r:  Refused", "NA"))

#Remove unsuitable variables
GSS <- GSS %>%
  select(-size, -hrs2, -sphrs2, -evwork, -hapcohab, -god, -hompop, -income, -rincome,
         -wrkstat, -divorce, -family16, -relig, -pray)

#Rename Variables names
GSS = GSS %>%
  rename(age_kid_born = "agekdbrn", father_educ = "paeduc",
         mother_educ = "maeduc", earners_in_household = "earnrs",
         income = "realinc", father_own_business = "pawrkslffam")
```

The response variable for the model was a self ranking on a scale of 1 - 10 of how the respondent ranked their own life ('rank', quantitative). For this variable, we swapped all of the responses so that 10 was the highest, and 1 was the lowest to increase readability of our results. We then inflated the results by multiplying each one by 10 so that our future coefficients wouldn't have to be extremely small. Finally, we removed all cases from the data set that didn't give a rank during the GSS. This brought the data set from 4032 cases down to 1966.

```
#Reorder Rank
GSS <- GSS %>%
  mutate(rank = case_when(
    rank == "1 - top" ~ 100,
    rank == "2" ~ 90,
```

```
    rank == "3" ~ 80,
    rank == "4" ~ 70,
    rank == "5" ~ 60,
    rank == "6" ~ 50,
    rank == "7" ~ 40,
    rank == "8" ~ 30,
    rank == "9" ~ 20,
    rank == "10 - bottom" ~ 10
  ))

#Eliminate NA response variables from dataset
GSS <- GSS %>%
  filter(rank != "NA")
```

The next step of data wrangling that we did was creating a reasonable minimum value for the income variable, and converting the units of the variable from dollars to thousands of dollars. This will improve readability of our model in future steps of this study.

```
#Create reaonable minimum value for income (converting negatives into 0.001 for later log
GSS <- GSS %>%
  mutate(income = ifelse(income == -100.0, 0.001, income))

#Convert from dollars to thousands of dollars
GSS$income <- GSS$income / 1000
```

The rest of the data wrangling that we completed was to create reasonable minimum values, convert certain variables into quantitative variables, and create reasonable groups for categorical variables.

```
#Turn age into numeric
GSS <- GSS %>%
  mutate(age = ifelse(age == "89 or older", 90, age))

#Create reasonable minimum value for siblings
GSS <- GSS %>%
  mutate(sibs = ifelse(sibs == -97, 0, sibs))

#Turn childs into numeric
GSS <- GSS %>%
  mutate(childs = ifelse(childs == "8 or more", 9, childs))

#Turn educations into numeric
```

```r
GSS <- GSS %>%
  mutate(educ = ifelse(educ == "No formal schooling", 0, educ)) %>%
  mutate(father_educ = ifelse(father_educ == "No formal schooling", 0, father_educ)) %>%
  mutate(mother_educ = ifelse(mother_educ == "No formal schooling", 0, mother_educ))

#Turn earners in household to numeric
GSS <- GSS %>%
  mutate(earners_in_household = ifelse(earners_in_household == "Eight or more",
                                       9, earners_in_household))

#Turn weeks worked into numeric
GSS <- GSS %>%
  mutate(weekswrk = ifelse(weekswrk == "None or zero", 0, weekswrk))

#Reassign party id to 3 categories
GSS <- GSS %>%
  mutate(partyid = case_when(
    partyid == "Independent (neither, no response)" ~ "independent",
    partyid == "Independent, close to democrat" ~ "democrat",
    partyid == "Independent, close to republican" ~ "republican",
    partyid == "Other party" ~ "independent",
    partyid == "Not very strong democrat" ~ "democrat",
    partyid == "Strong democrat" ~ "democrat",
    partyid == "Not very strong republican" ~ "republican",
    partyid == "Strong republican" ~ "republican"
  ))

#Fix own gun variable
GSS <- GSS %>%
  mutate(owngun = ifelse(owngun == "REFUSED", NA, owngun))

#Reassign job satisfaction to binary
GSS <- GSS %>%
  mutate(satjob = case_when(
    satjob == "A little dissatisfied" ~ "dissatisfied",
    satjob == "Very dissatisfied" ~ "dissatisfied",
    satjob == "Moderately satisfied" ~ "satisfied",
    satjob == "Very satisfied" ~ "satisfied"
  ))

#Turn married into binary
```

```r
GSS <- GSS %>%
  mutate(marital = case_when(
    marital == "Married" ~ "Married",
    marital == "Divorced" ~ "Not Married",
    marital == "Never married" ~ "Not Married",
    marital == "Separated" ~ "Not Married",
    marital == "Widowed" ~ "Not Married"
  ))

#Convert variables to numeric
GSS$childs = as.integer(GSS$childs)
GSS$age = as.integer(GSS$age)
GSS$educ = as.integer(GSS$educ)
GSS$father_educ = as.integer(GSS$father_educ)
GSS$mother_educ = as.integer(GSS$mother_educ)
GSS$earners_in_household = as.integer(GSS$earners_in_household)
GSS$weekswrk = as.integer(GSS$weekswrk)

#Skim data
skim(GSS)
```

Table 1: Data summary

Name	GSS
Number of rows	1966
Number of columns	19
_____	
Column type frequency:	
character	6
numeric	13
_____	
Group variables	None

Variable type: character

skim_variable	n_missing	complete_rate	min	max	empty	n_unique	whitespace
marital	4	1.00	7	11	0	2	0
sex	22	0.99	4	6	0	2	0
partyid	10	0.99	8	11	0	3	0
satjob	625	0.68	9	12	0	2	0

skim_variable	n_missing	complete_rate	min	max	empty	n_unique	whitespace
owngun	53	0.97	2	3	0	2	0
father_own_business	265	0.87	2	3	0	2	0

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
year	0	1.00	2021.00	0.00	2021	2021.00	2021.00	2021.0	2021.00	
id_	0	1.00	2185.72	1283.88	6	1080.25	2145.50	3270.5	4470.00	
sibs	33	0.98	3.09	2.59	0	1.00	2.00	4.0	35.00	
childs	17	0.99	1.70	1.53	0	0.00	2.00	3.0	9.00	
age	134	0.93	52.46	17.04	19	38.00	53.00	66.0	90.00	
age_kid_born	588	0.70	25.58	6.17	9	21.00	25.00	30.0	57.00	
educ	15	0.99	14.81	2.83	0	12.00	15.00	16.0	20.00	
father_educ	456	0.77	12.49	3.81	0	12.00	12.00	16.0	20.00	
mother_educ	206	0.90	12.44	3.34	0	12.00	12.00	14.0	20.00	
earners_in_household	43	0.98	1.40	1.14	0	1.00	1.00	2.0	9.00	
weekswrk	55	0.97	31.27	23.17	0	0.00	46.00	52.0	52.00	
rank	0	1.00	64.20	17.18	10	60.00	60.00	80.0	100.00	
income	0	1.00	35.68	40.11	0	8.18	23.98	43.6	144.84	

The predictor variables that we chose to study initially were marital status ('marital', quantitative), siblings ('sibs', quantitative), children ('childs', quantitative), the age of the respondent when they had their first born child ('age_kid_born', quantitative), years of completed education ('educ', quantitative), years of father's education ('father_educ', quantitative), years of mother's education ('mother_educ', quantitative), amount of people earning an income in the respondent's household ('earners_in_household', quantitative), political party ('partyid', categorical), job satisfaction ('satjob', categorical), if the respondent owns a gun ('owngun', categorical), and income ('income', quantitative).

```
initial_model = lm(rank ~ marital + sibs + childs + age_kid_born + educ + father_educ +
                     mother_educ + earners_in_household + partyid + satjob + owngun +
                     income, data = GSS)
summary(initial_model)
```

```
Call:
lm(formula = rank ~ marital + sibs + childs + age_kid_born +
    educ + father_educ + mother_educ + earners_in_household +
```

```
    partyid + satjob + owngun + income, data = GSS)

Residuals:
    Min      1Q  Median      3Q     Max
-49.815  -7.066  -0.113   7.775  47.304

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)          43.73369    5.39742   8.103 3.00e-15 ***
maritalNot Married   -1.65376    1.38140  -1.197    0.232
sibs                  0.32506    0.26745   1.215    0.225
childs               -0.67041    0.53373  -1.256    0.210
age_kid_born         -0.13372    0.10736  -1.245    0.213
educ                  1.19623    0.25946   4.611 4.91e-06 ***
father_educ           0.07330    0.19987   0.367    0.714
mother_educ           0.18583    0.22509   0.826    0.409
earners_in_household -0.13715    0.69178  -0.198    0.843
partyidindependent   -0.98765    1.50544  -0.656    0.512
partyidrepublican    -0.43277    1.39243  -0.311    0.756
satjobsatisfied       1.36976    1.88529   0.727    0.468
owngunYES            -0.07708    1.22807  -0.063    0.950
income                0.08159    0.01474   5.536 4.61e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 14.22 on 603 degrees of freedom
  (1349 observations deleted due to missingness)
Multiple R-squared:  0.1639,    Adjusted R-squared:  0.1459
F-statistic: 9.094 on 13 and 603 DF,  p-value: < 2.2e-16
```

Model Selection:

After the data was ready to be studied and we had our initial model, we used backwards elimination to remove insignificant variables and reach our final model. During backwards elimination, we started with a full model and removed the variable with the highest p-value. Then the model is re-tested and this process is repeated until every variable had a p-value of below 0.05, and is therefore significant. After completing backwards elimination on our initial model, every single predictor variable was removed except marital status, income, and years of education.

This means that our final model includes 'marital', 'educ', and 'income'.

```
#Complete backwards elimination
best <- regsubsets(rank ~ marital + sibs + childs + age_kid_born + educ + father_educ +
                    mother_educ + earners_in_household + partyid + satjob + owngun +
                    income, data = GSS, nbest = 1, nvmax = 3,
                   method = "backward")

#Results from backwards elimination
with(summary(best), data.frame(rsq, outmat))
```

```
              rsq maritalNot.Married sibs childs age_kid_born educ father_educ
1  ( 1 ) 0.1150709
2  ( 1 ) 0.1538852                                          *
3  ( 1 ) 0.1560864                 *                        *
         mother_educ earners_in_household partyidindependent partyidrepublican
1  ( 1 )
2  ( 1 )
3  ( 1 )
         satjobsatisfied owngunYES income
1  ( 1 )                            *
2  ( 1 )                            *
3  ( 1 )                            *
```

$$\widehat{rank} = 44.17 - (2.149 * marital_N) + (1.228 * educ) + (0.08283 * income)$$

```
final_model = lm(rank ~ marital + educ + income, data = GSS)
```

Conditions:

The final model was then checked to be sure that multicollinearity does not occur by examining the VIF values for each variable in the model. Any VIF value that exceeds, or is close to the value of 5 would suggest that there is multicollinearity between two or more variables.

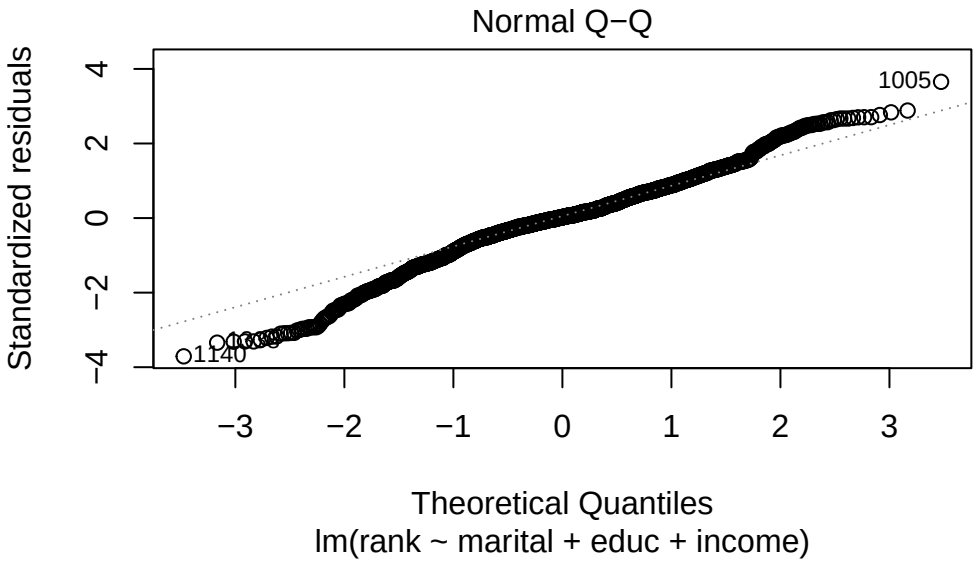
```
vif(final_model)
```

```
 marital     educ   income
1.109340 1.146197 1.250651
```

After running the VIF function, we find that the highest value is 1.25 which is less than 5. This indicates that there is not multicollinearity between the variables in the model.

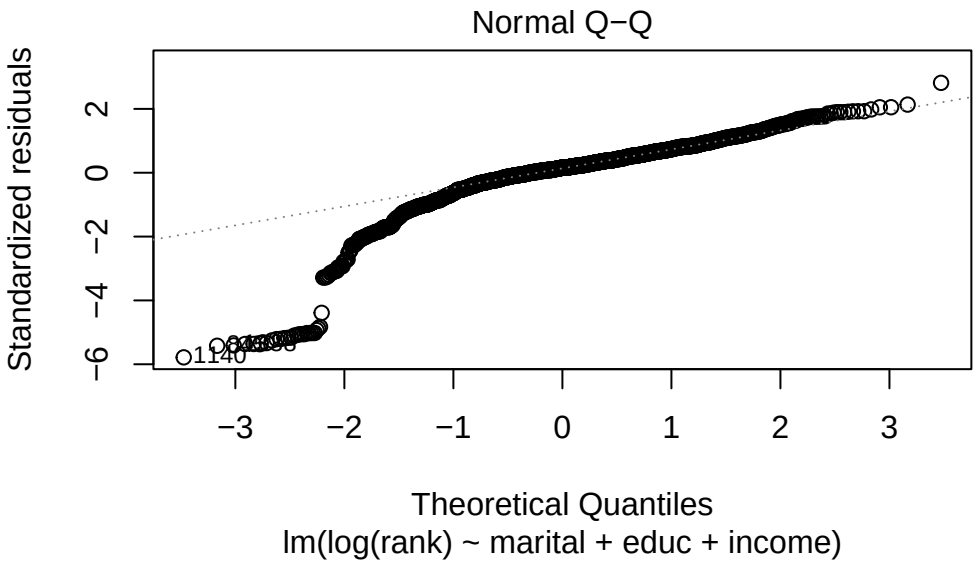
Next, normality was checked on the model.

```
plot(final_model, which = 2)
```

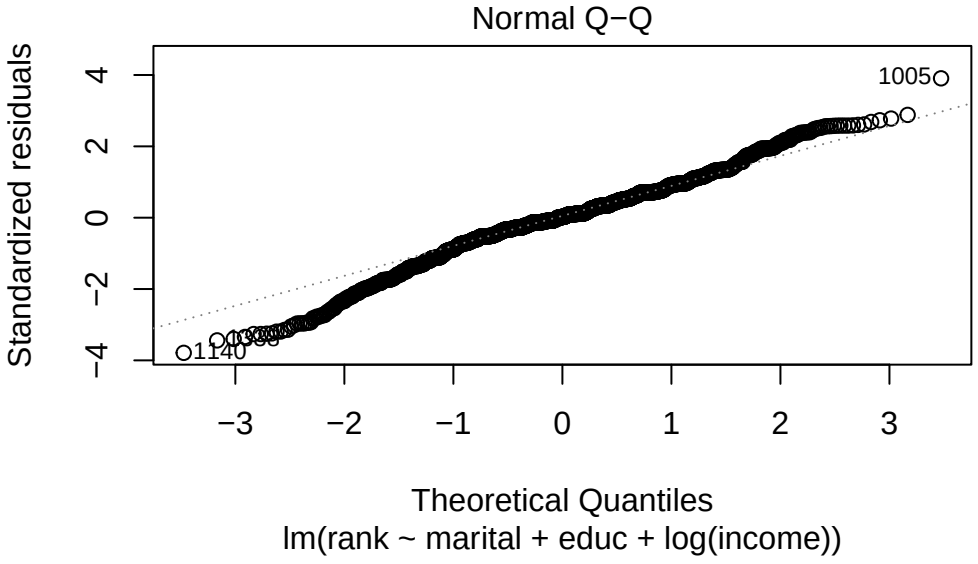


The Q-Q plot seems to indicate that it is not normal, given that the tails of plot stray away from the line. As a result, we tried to do a log transformation on the model on both the response and income.

```
transformed_response_final_model = lm(log(rank) ~ marital + educ + income, data = GSS)
transformed_income_final_model = lm(rank ~ marital + educ + log(income), data = GSS)
plot(transformed_response_final_model, which = 2)
```



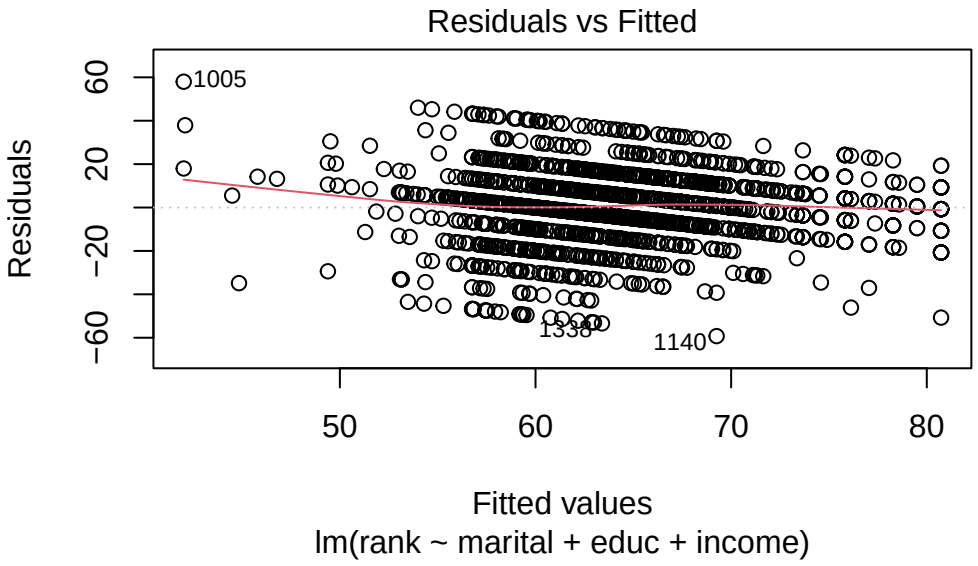

```r
plot(transformed_income_final_model, which = 2)
```



When performing a log on the response variable of rank, it seemed to eliminate most of the deviance from the line on the right side of the Q-Q plot, however exacerbated it on the left side. Furthermore, performing a log transformation on income had little to no effect at all. As a result, it seems best to use the untransformed model that we initially created. While it does not seem to perfectly conform to the standard of normality, we will continue to work with the model as it is close to normal, and allows for more clarity in coefficient interpretations.

Next, we checked to see if the linearity condition was met.

```
plot(final_model, which = 1)
```

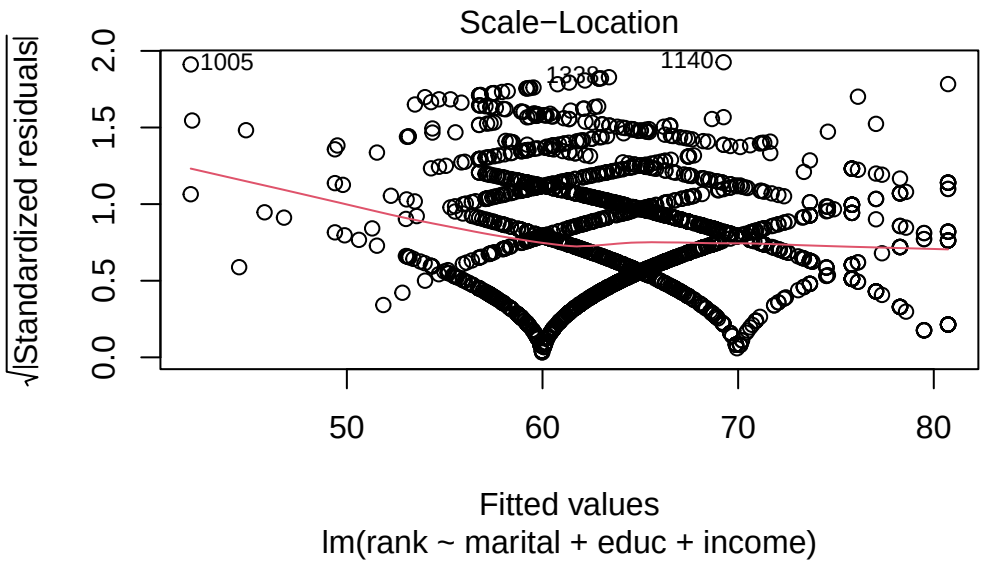


After looking at the residuals vs fitted plot we can see that the line, while not perfectly straight, definitely indicates that linearity is upheld for the model.

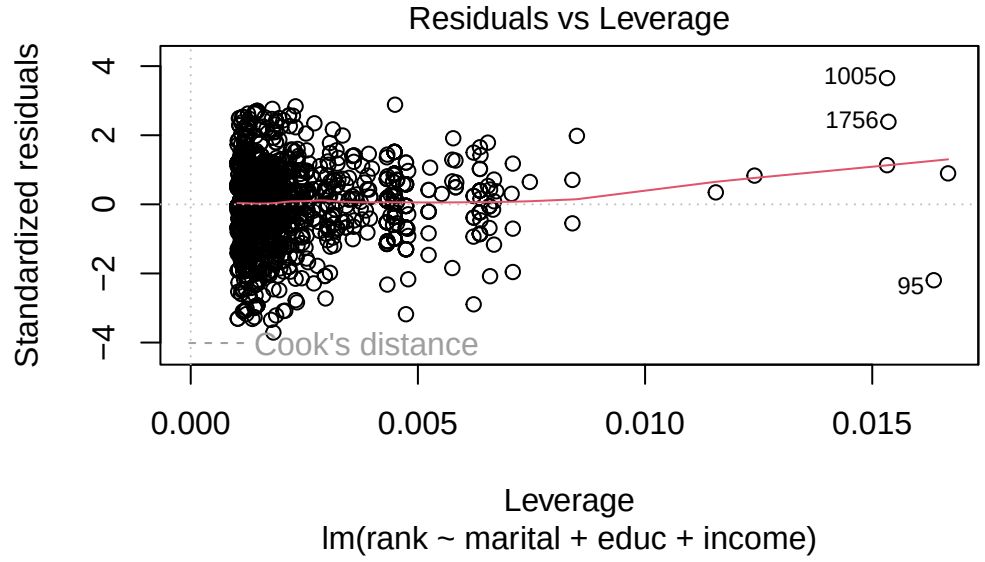
Then we examined the data to check if it held up to the independence condition. Given that the individual units of the data are random adults in the United States, we can assume that the data is independent. We know this because the source of the information, the General Social Survey, collects data from random adults across America to gain information for studying social dynamics.

Finally, we checked for Equality of Variance within the model.

```
plot(final_model, which = 3)
```




```r
plot(final_model, which = 5)
```



The first plot used to examine the Equality of Variance condition for inference was a Residual vs. Fitted plot. This was a little concerning, as there was a clear skew towards the middle of the plot; however, there does appear to be a band of ranges that the points are within on both sides of the line.

The second plot used to test for Equality of Variance was the Residual vs. Leverage plot. The plot shows no data points outside of the Cook's distance. Thus, there are no extreme outliers with high leverage. Given the results of these plots together, I would hesitantly say that they indicate that the Equality of Variance condition is met for the model.

Results:

The final model should be taken with a grain of salt due to the lack of normality and some issues with equity of variance. With that in mind, our model gives insight to a few variables that can cause a change in self ranking, however it is important to point out that we started with 12 variables, and by method of backwards elimination, found only 3 to have statistical significance with rank at the 5% level. Additionally, remember when interpreting the results that we changed the ranking scale to be out of 100 (1 being the lowing and 100 being the highest). Overall, the model explains 12.3% if the variability in rank. The coefficients and model are shown below…

```
final_model = lm(rank ~ marital + educ + income, data = GSS)
summary(final_model)
```

```
Call:
lm(formula = rank ~ marital + educ + income, data = GSS)

Residuals:
    Min      1Q  Median      3Q     Max
-59.257  -7.921   0.388   9.630  57.981

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)        44.16784    2.03495  21.705  < 2e-16 ***
maritalNot Married -2.14894    0.76351  -2.815  0.00493 **
educ                1.22833    0.13690   8.972  < 2e-16 ***
income              0.08283    0.01009   8.210 3.97e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 15.99 on 1943 degrees of freedom
```

```
  (19 observations deleted due to missingness)
Multiple R-squared:  0.123, Adjusted R-squared:  0.1217
F-statistic: 90.87 on 3 and 1943 DF,  p-value: < 2.2e-16
```

$$\widehat{rank} = 44.17 - (2.149 * marital_N) + (1.228 * educ) + (0.08283 * income)$$

The coefficient that stood out to us the most was the impact of marital status on rank. Compared to respondents who are married, we can expect the respondents who aren't to have a decrease of 2.149 points on their self-ranking, holding all else constant.

Another significant, yet not as surprising result was the effect of education and income on rank. For a 1 year increase in education, we can expect a 1.228 unit increase in the rank variable, holding all else constant. For a $1000 dollar increase in income, we can expect a 0.08283 point increase in how a respondent ranked their life, holding all else constant.

After reporting it should be noted that even though these three variables showed statistical significance with their effects on rank, the coefficients are extremely small since they are being used on a 100 point scale. For example, if someone who isn't married scored themselves at a 90, being married would only move up to a 92.15. Another example could be created for income. The difference between the expected score of someone making $50 thousand dollars a year a $100 thousand dollars a year is only 4.1415 points.

Conclusion:

The goal of this project was to predict the self ranking of individuals in the United States in 2021. It can be used to describe the possible relationship between an individual's social characteristics and their own perception of rank within society, for all United States citizens in the years around 2021.

This model is limited in the sense of geographic location and time frame. Since the data was exclusively taken by residents of the United States in the year 2021, it is important to recognize that this model can only be applied to other united states residents for the year 2021. For instance, one could not make a prediction about a resident of China in the year 2000 using this model. This makes sense because a different country would likely have a different culture and standards to which people rank themselves based off of. Moreover, the general time for which the prediction is being made is important to be near 2021 because income would vary with economic conditions and inflation over a period of time.

That being said, we should be able to make inferences about people within the region of the United States in the year 2021. This is because the General Social Survey takes a random sample of adults within America, making them independent from each other. Thus, for this select group of people we are able to both describe the relationship between socioeconomic

factors and self-ranking for survey participants as well as predict a self-ranking for other United States residents.

As stated earlier in this report, roughly half of the data entries were dropped due to missing values for rank. This model does not take any of those dropped entries into consideration, which is another limitation of this model. There is a possibility that all of the entries that were dropped share common characteristics and could have similar rank. As a result of ignoring this, the model could possibly be missing statistically significant information about a large percentage of the data that was used.

Ultimately, we see one of the biggest takeaways of this research not to be the idea that we can predict how people in the United States might rank themselves in society; rather, we think it's the possibility to statistically find what a culture deems to be socially significant. Our model, based upon American survey participants, found education, income, and marital status to be important. These are all central ideas to the United States tradition and the so-called "American Dream." Thus, moving forward, it would be fascinating to have people of other cultures take a similar survey to the GSS and see how they compare when similar models are created.