Analysis Of Income and College Major

Danish Tamboli

8/18/2020

Dependancies/Inital Setup

First we will load the dataset from the collegeIncome package.

```
library(collegeIncome)
data(college)
```

```
dim(college)
```

Dataset columns

```
## [1] 173  19
```

```
names(college)
```

```
##  [1] "rank"                                "major_code"
##  [3] "major"                               "major_category"
##  [5] "total"                               "sample_size"
##  [7] "perc_women"                          "p25th"
##  [9] "median"                              "p75th"
## [11] "perc_men"                            "perc_employed"
## [13] "perc_employed_fulltime"              "perc_employed_parttime"
## [15] "perc_employed_fulltime_yearround"    "perc_unemployed"
## [17] "perc_college_jobs"                   "perc_non_college_jobs"
## [19] "perc_low_wage_jobs"
```

From this we see that the Dataset consists of 19 fields, some of which will be useful to us for Inference.

```
head(college)
```

Dataset Summary

```
##   rank major_code                                    major major_category
## 1    1       2419                      Petroleum Engineering    Engineering
## 2    2       2416             Mining And Mineral Engineering    Engineering
## 3    3       2415                  Metallurgical Engineering    Engineering
## 4    4       2417 Naval Architecture And Marine Engineering    Engineering
## 5    5       2405                       Chemical Engineering    Engineering
## 6    6       2418                        Nuclear Engineering    Engineering
##   total sample_size perc_women p25th median  p75th   perc_men perc_employed
## 1  2339          36  0.9109326 25000  40000  50000 0.08906743     0.9115044
## 2   756           7  0.5154064 26000  37000  40000 0.48459355     0.7980501
## 3   856           3  0.5942076 26700  45000  60000 0.40579235     0.7871943
## 4  1258          16  0.6521298 26000  35000  45000 0.34787018     0.8465608
## 5 32260         289  0.4179248 31500  62000 109000 0.58207520     0.8515625
## 6  2573          17  0.4305368 23000  44700  50000 0.56946324     0.8474507
##   perc_employed_fulltime perc_employed_parttime
## 1              0.9206524              0.1774785
## 2              0.7110092              0.3623853
## 3              0.8833498              0.3387257
## 4              0.9366337              0.1673267
## 5              0.8086363              0.4020061
## 6              0.8756262              0.2040405
##   perc_employed_fulltime_yearround perc_unemployed perc_college_jobs
## 1                        0.7704431      0.08849558         0.6702970
## 2                        0.7093101      0.20194986         0.3867764
## 3                        0.7738366      0.21280567         0.7289116
## 4                        0.6527853      0.15343915         0.2460902
## 5                        0.6852821      0.14843750         0.5867515
## 6                        0.6567727      0.15254929         0.4624782
##   perc_non_college_jobs perc_low_wage_jobs
## 1             0.1821782         0.05544554
## 2             0.5158761         0.21560172
## 3             0.1759983         0.03014828
## 4             0.4107636         0.04323827
## 5             0.3860437         0.11801062
## 6             0.4057592         0.23472949
```

```
summary(college)
```

```
##       rank       major_code      major           major_category
##  Min.   :  1   Min.   :1100   Length:173         Length:173
##  1st Qu.: 44   1st Qu.:2403   Class :character   Class :character
##  Median : 87   Median :3608   Mode  :character   Mode  :character
##  Mean   : 87   Mean   :3880
##  3rd Qu.:130   3rd Qu.:5503
##  Max.   :173   Max.   :6403
##
##      total         sample_size       perc_women         p25th
##  Min.   :   124   Min.   :   2.0   Min.   :0.0000   Min.   :18500
##  1st Qu.:  4361   1st Qu.:  39.0   1st Qu.:0.3397   1st Qu.:24000
##  Median : 15058   Median : 130.0   Median :0.5357   Median :27000
##  Mean   : 39168   Mean   : 356.1   Mean   :0.5226   Mean   :29501
##  3rd Qu.: 38844   3rd Qu.: 338.0   3rd Qu.:0.7020   3rd Qu.:33000
##  Max.   :393735   Max.   :4212.0   Max.   :0.9690   Max.   :95000
##
```

```
##      median           p75th           perc_men       perc_employed
##  Min.   : 22000   Min.   : 22000   Min.   :0.03105   Min.   :0.0000
##  1st Qu.: 33000   1st Qu.: 42000   1st Qu.:0.29798   1st Qu.:0.7477
##  Median : 36000   Median : 47000   Median :0.46429   Median :0.8028
##  Mean   : 40151   Mean   : 51494   Mean   :0.47745   Mean   :0.7886
##  3rd Qu.: 45000   3rd Qu.: 60000   3rd Qu.:0.66033   3rd Qu.:0.8410
##  Max.   :110000   Max.   :125000   Max.   :1.00000   Max.   :0.9562
##
##  perc_employed_fulltime perc_employed_parttime perc_employed_fulltime_yearround
##  Min.   :0.5743         Min.   :0.0000         Min.   :0.5857
##  1st Qu.:0.7741         1st Qu.:0.2090         1st Qu.:0.7009
##  Median :0.8319         Median :0.2862         Median :0.7484
##  Mean   :   Inf         Mean   :0.2874         Mean   :0.7476
##  3rd Qu.:0.8974         3rd Qu.:0.3623         3rd Qu.:0.7896
##  Max.   :   Inf         Max.   :0.5518         Max.   :1.0000
##                         NA's   :1
##  perc_unemployed   perc_college_jobs perc_non_college_jobs perc_low_wage_jobs
##  Min.   :0.04383   Min.   :0.0633    Min.   :0.08278       Min.   :0.00000
##  1st Qu.:0.15899   1st Qu.:0.2974    1st Qu.:0.27995       1st Qu.:0.06957
##  Median :0.19723   Median :0.4160    Median :0.42020       Median :0.10857
##  Mean   :0.21140   Mean   :0.4478    Mean   :0.41498       Mean   :0.11481
##  3rd Qu.:0.25229   3rd Qu.:0.6170    3rd Qu.:0.52756       3rd Qu.:0.15353
##  Max.   :1.00000   Max.   :0.8383    Max.   :0.85364       Max.   :0.36566
##                    NA's   :1         NA's   :1             NA's   :1
```

The First thing that interests us from this are the different Majors and Major Categories

```
with(college,head(table(major)))
```

```
## major
##                              Accounting                    Actuarial Science
##                                       1                                    1
##      Advertising And Public Relations                 Aerospace Engineering
##                                       1                                    1
##                Agricultural Economics Agriculture Production And Management
##                                       1                                    1
```

```
with(college,table(major_category))
```

```
## major_category
##     Agriculture & Natural Resources                                Arts
##                                  10                                   8
##              Biology & Life Science                            Business
##                                  14                                  13
##         Communications & Journalism            Computers & Mathematics
##                                   4                                  11
##                           Education                         Engineering
##                                  16                                  29
##                              Health          Humanities & Liberal Arts
##                                  12                                  15
## Industrial Arts & Consumer Services                     Interdisciplinary
##                                   7                                   1
```

```
##                 Law & Public Policy                   Physical Sciences
##                                   5                                  10
##           Psychology & Social Work                      Social Science
##                                   9                                   9
```

We will not have a look at Majors as there exists many majors with just 1 entry which aren't sufficient to factor in their effect, Rather we will look into Major Categories

We should remove the Interdisciplinary category from the Major Categories as it has a sample size of just 1, which certainly not enough to draw conclusions around it.

```
library(dplyr)
```

```
## Warning: package 'dplyr' was built under R version 4.0.2
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
new_college <- college %>%
        filter(major_category != "Interdisciplinary")
table(new_college$major_category)
```

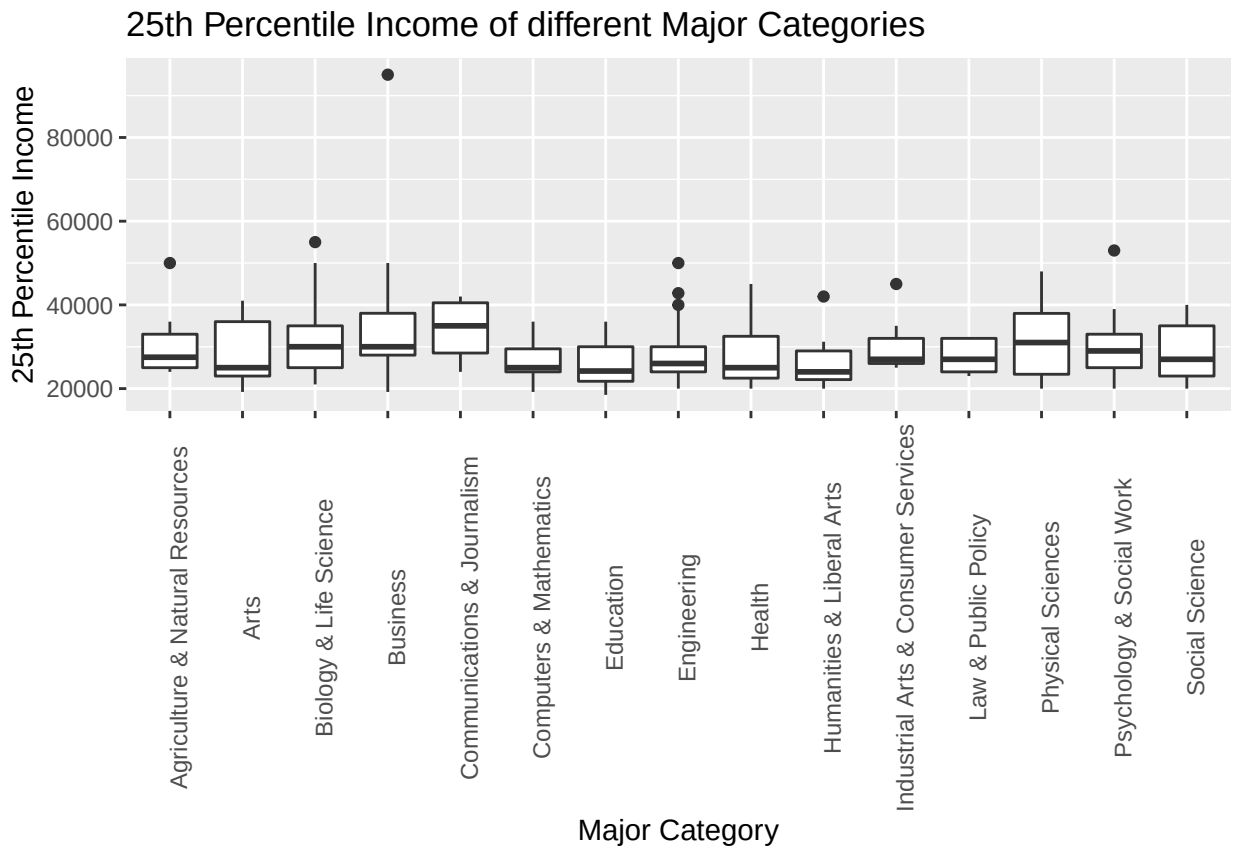
```
##
##     Agriculture & Natural Resources                                 Arts
##                                  10                                    8
##              Biology & Life Science                             Business
##                                  14                                   13
##         Communications & Journalism               Computers & Mathematics
##                                   4                                   11
##                           Education                          Engineering
##                                  16                                   29
##                              Health            Humanities & Liberal Arts
##                                  12                                   15
## Industrial Arts & Consumer Services                  Law & Public Policy
##                                   7                                    5
##                   Physical Sciences             Psychology & Social Work
##                                  10                                    9
##                      Social Science
##                                   9
```

Now that we have removed Interdisciplinary Major category we can proceed.

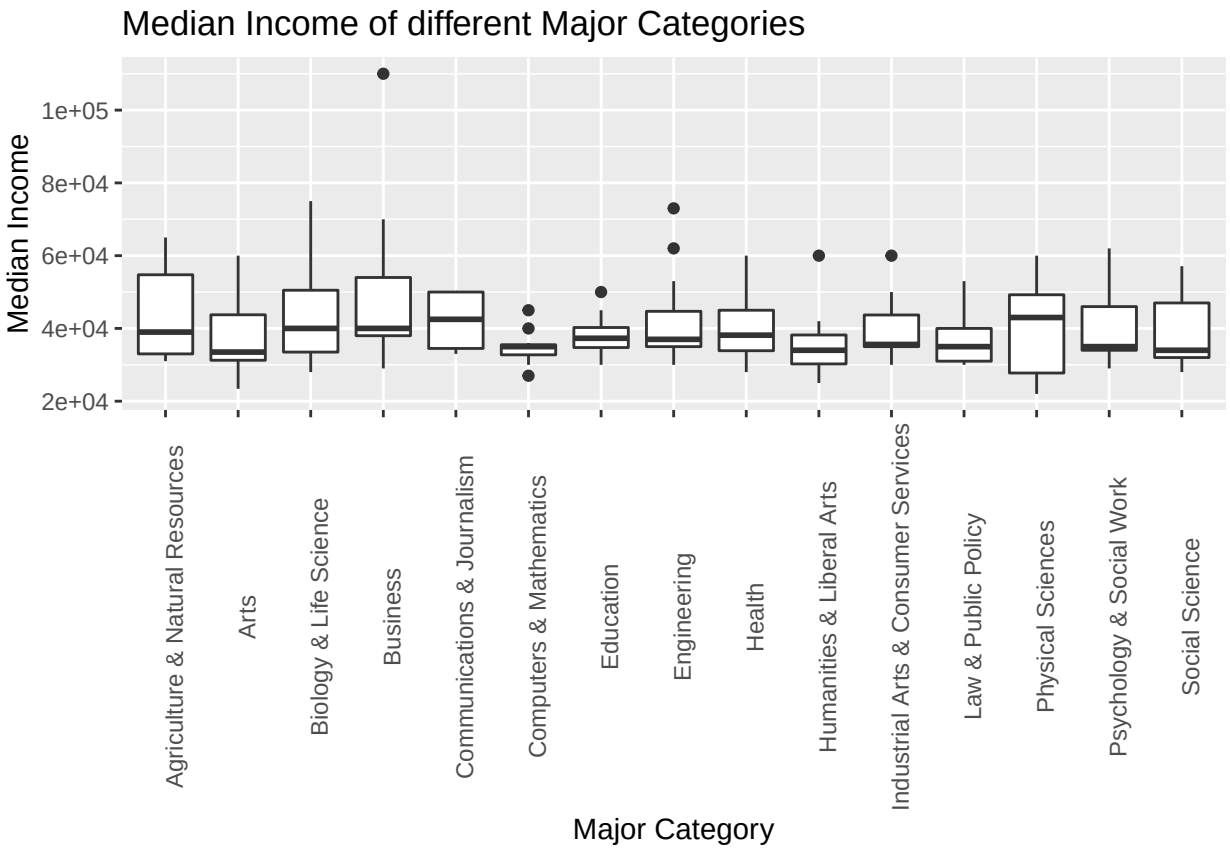
```
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 4.0.2
```

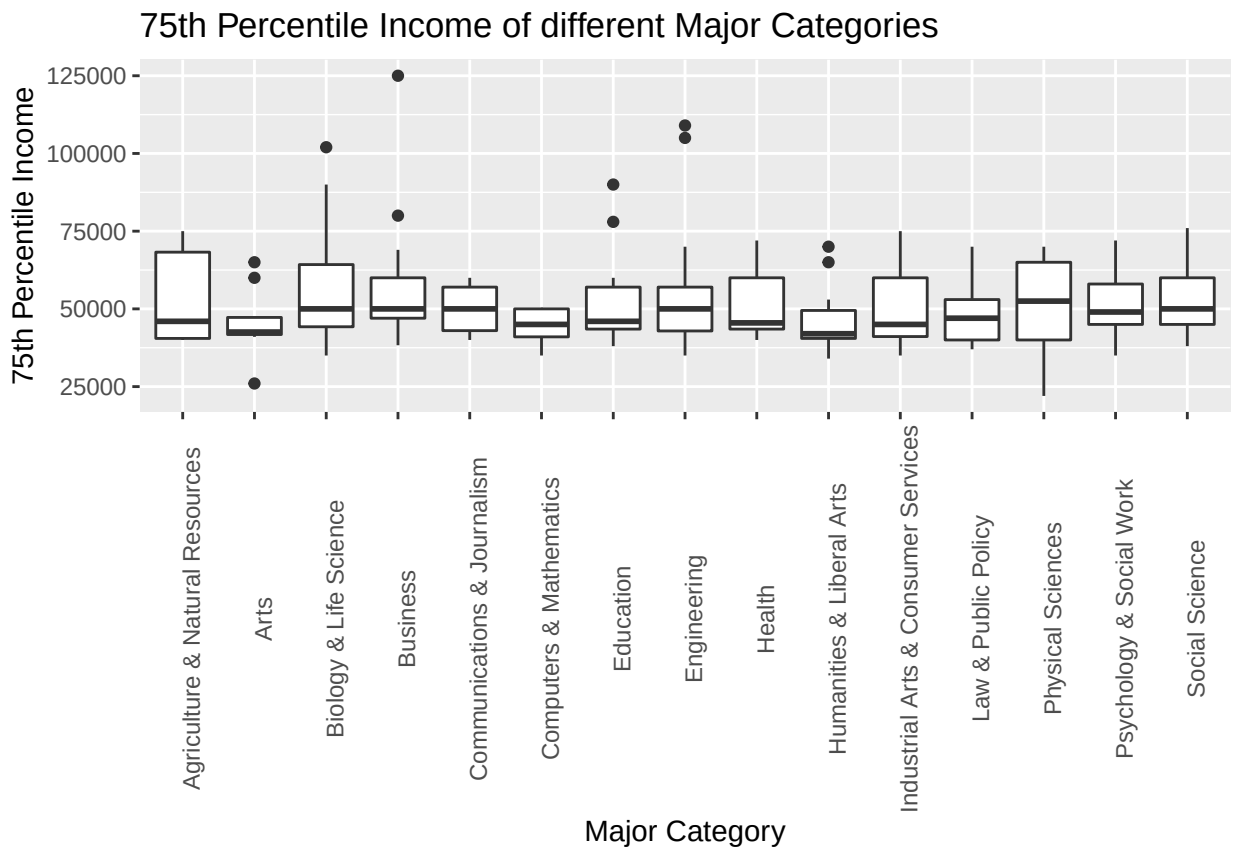
```
ggplot(data = new_college, aes(x = major_category, y = p25th))  + geom_boxplot(aes(x = major_category, y
```




```
ggplot(data = new_college, aes(x = major_category, y = median))  + geom_boxplot(aes(x = major_category,
```




```
ggplot(data = new_college, aes(x = major_category, y = p75th))  + geom_boxplot(aes(x = major_category, y
```



Upon looking at the 25th, 50th and 75th percentile of Incomes across various Major Categories we observe no major variation, We will take the Median as an outcome.

Linear Regression/ Linear Curve fitting

Upon looking further into the fields of the Dataset we see:

- Employment is divided into Full time and Part time and Unemployed.
- Jobs are divided into COllege, Non College and Low wage jobs.
- Income is also separated by Gender.

We will take some of these factors into account as predictors when fitting the outcome Median Income.

```
fit <- lm(median ~ major_category + perc_men + perc_college_jobs + perc_low_wage_jobs,new_college)
```

Fields we have haven't taken into account and why:

- perc_women, Taken into account when perc_men was included as predictor (Singularity)
- rank, major_code, major, Not taken into account as most fields have 1 sample, not enough to make conclusions, major_category taken instead.
- total, sample-size, Irrelevant with respect to our need.
- p25th, p75th, Not taken into account as no drastic variation noticed, hence median is a good estimate to select as outcome.

- perc_employed, perc_employed_fulltime, perc_employed_parttime, perc_employed_fulltime_yearround, perc_unemployed, Not taken into account as we are looking at Income regardless of Duration they are employed within a year or whether they work full time or part time.
- perc_non_college_jobs, Taken into account when perc_college_jobs and perc_low_wage_jobs were included as predictiors (Singularity)

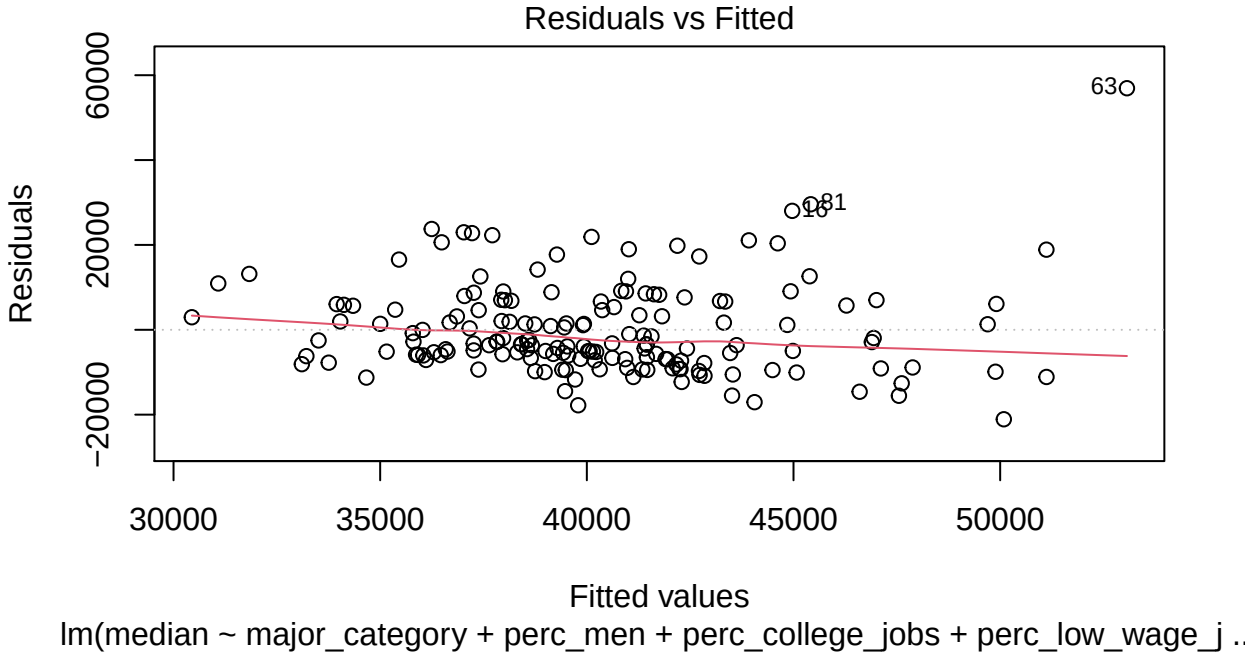
```
summary(fit)
```

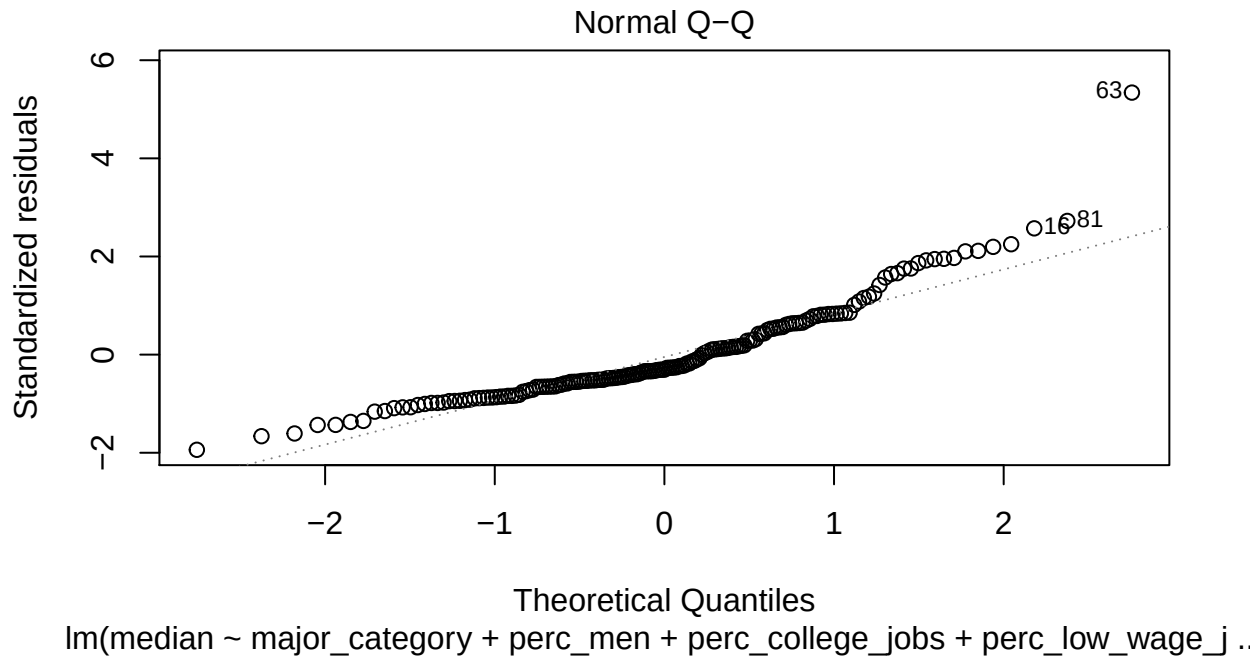
```
##
## Call:
## lm(formula = median ~ major_category + perc_men + perc_college_jobs +
##     perc_low_wage_jobs, data = new_college)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -21088  -6917  -3211   5965  56932
##
## Coefficients:
##                                                 Estimate Std. Error t value
## (Intercept)                                      44479.7     5864.2   7.585
## major_categoryArts                               -5429.2     5435.1  -0.999
## major_categoryBiology & Life Science               707.2     4772.0   0.148
## major_categoryBusiness                            5540.4     4846.5   1.143
## major_categoryCommunications & Journalism        -2776.6     6761.1  -0.411
## major_categoryComputers & Mathematics            -9607.2     5129.6  -1.873
## major_categoryEducation                          -5140.6     4591.3  -1.120
## major_categoryEngineering                        -2931.4     4185.3  -0.700
## major_categoryHealth                             -3700.9     4880.2  -0.758
## major_categoryHumanities & Liberal Arts          -9022.7     4711.1  -1.915
## major_categoryIndustrial Arts & Consumer Services -2731.0     5604.6  -0.487
## major_categoryLaw & Public Policy                -5612.6     6374.9  -0.880
## major_categoryPhysical Sciences                  -3268.4     5109.4  -0.640
## major_categoryPsychology & Social Work           -5102.7     5340.0  -0.956
## major_categorySocial Science                     -3059.1     5287.2  -0.579
## perc_men                                          5011.5     3958.1   1.266
## perc_college_jobs                                -7841.4     5283.5  -1.484
## perc_low_wage_jobs                                2108.6    16018.8   0.132
##                                                 Pr(>|t|)
## (Intercept)                                     3.01e-12 ***
## major_categoryArts                                0.3194
## major_categoryBiology & Life Science              0.8824
## major_categoryBusiness                            0.2548
## major_categoryCommunications & Journalism         0.6819
## major_categoryComputers & Mathematics             0.0630 .
## major_categoryEducation                           0.2646
## major_categoryEngineering                         0.4847
## major_categoryHealth                              0.4494
## major_categoryHumanities & Liberal Arts           0.0573 .
## major_categoryIndustrial Arts & Consumer Services 0.6268
## major_categoryLaw & Public Policy                 0.3800
## major_categoryPhysical Sciences                   0.5233
## major_categoryPsychology & Social Work            0.3408
## major_categorySocial Science                      0.5637
## perc_men                                          0.2074
```

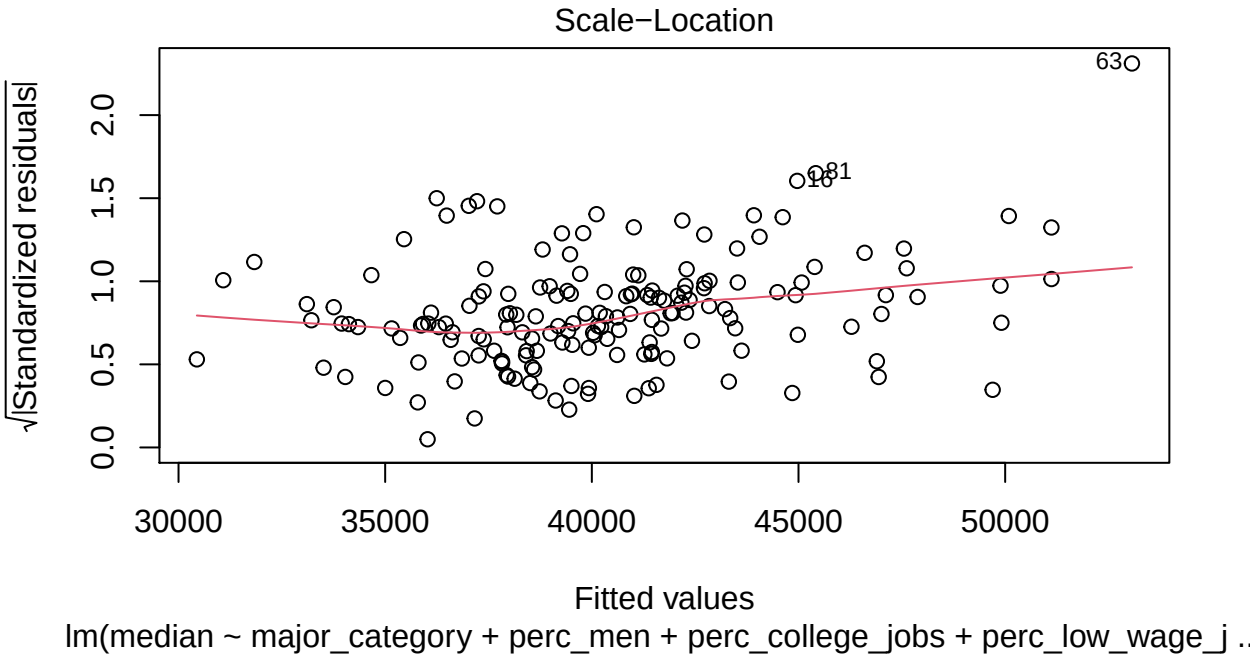
```
## perc_college_jobs                                    0.1398
## perc_low_wage_jobs                                   0.8954
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11330 on 153 degrees of freedom
##   (1 observation deleted due to missingness)
## Multiple R-squared:  0.1236, Adjusted R-squared:  0.02627
## F-statistic:  1.27 on 17 and 153 DF,  p-value: 0.2191
```

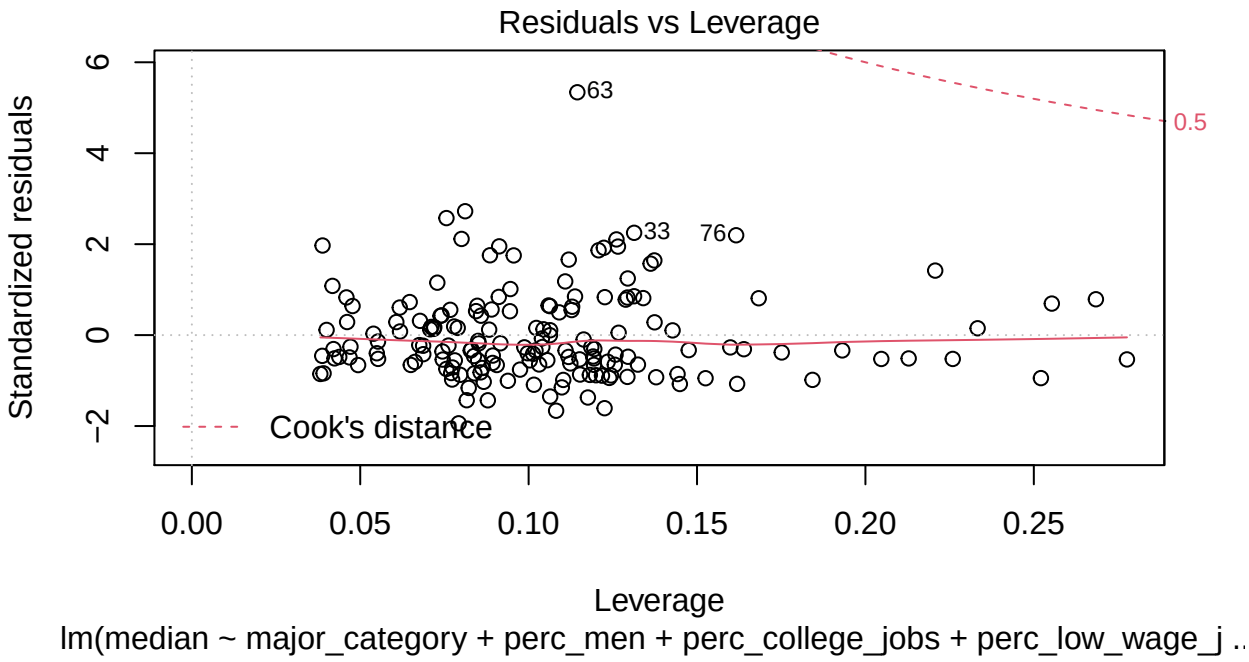
Looking at the summary, we see that holding gender and job category constant we don't see a major difference in Income across the Major Categories, indicating that Major Categories don't have a strong impact on Income.

```
plot(fit)
```




Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(median ~ major_category + perc_men + perc_college_jobs + perc_low_wage_j ..
63
16
81
−2
−1
0
1
2
6
4
2
0


Scale−Location
√|Standardized residuals|
Fitted values
lm(median ~ major_category + perc_men + perc_college_jobs + perc_low_wage_j ..
63
16
81
30000
35000
40000
45000
50000
0.0
0.5
1.0
1.5
2.0




```
#fit_residuals <- residuals(fit)
#fitted <- fitted.values(fit)
#plot(density(fit_rediduals))
#plot(fitted,fit_residuals)
```

Looking at Plot 1 (Residuals vs Fitted) we see that Normality assumptions are not far off, Overall there doesn't seem to be an effect of College Major Category on Median Income of an Indivisual in this particular study.