Analysis of income by college major

Dependencies

First we load the dataset from the `collegeIncome` package. Next we load the `broom` package for tidier display of regression output and the `dplyr` package for working with data frames.

```
library(collegeIncome)
data(college)

library(broom)
library(dplyr)
```

```
## Warning: package 'dplyr' was built under R version 3.4.2
```

```
##
## 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
```

Exploratory data analysis

Let's take a look at the data to get acquainted with the contents.

```
head(college)
```

```
##   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
## 1  2339          36  0.9109326 25000  40000  50000 0.08906743
## 2   756           7  0.5154064 26000  37000  40000 0.48459355
## 3   856           3  0.5942076 26700  45000  60000 0.40579235
## 4  1258          16  0.6521298 26000  35000  45000 0.34787018
## 5 32260         289  0.4179248 31500  62000 109000 0.58207520
## 6  2573          17  0.4305368 23000  44700  50000 0.56946324
##   perc_employed perc_employed_fulltime perc_employed_parttime
## 1     0.9115044              0.9206524              0.1774785
## 2     0.7980501              0.7110092              0.3623853
## 3     0.7871943              0.8833498              0.3387257
## 4     0.8465608              0.9366337              0.1673267
## 5     0.8515625              0.8086363              0.4020061
## 6     0.8474507              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
```

```
tail(college)
```

```
##     rank major_code                    major            major_category
## 168  168       3302 Composition And Rhetoric Humanities & Liberal Arts
## 169  169       3609                  Zoology    Biology & Life Science
## 170  170       5201   Educational Psychology  Psychology & Social Work
## 171  171       5202      Clinical Psychology  Psychology & Social Work
## 172  172       5203    Counseling Psychology  Psychology & Social Work
## 173  173       3501          Library Science                 Education
##     total sample_size perc_women p25th median  p75th  perc_men
## 168 18953         151  0.8459344 30000  42000  65000 0.1540656
## 169  8409          47  0.7643203 50000  65000 102000 0.2356797
## 170  2854           7  0.8644561 33000  46000  58000 0.1355439
## 171  2838          13  0.8128766 22000  29000  38000 0.1871234
## 172  4626          21  0.5847764 39000  48000  58000 0.4152236
## 173  1098           2  0.3212961 22500  38400  45000 0.6787039
##     perc_employed perc_employed_fulltime perc_employed_parttime
## 168     0.7636511              1.0041209              0.1016484
## 169     0.6757741              0.8792842              0.1889597
## 170     0.7932137              0.9613045              0.1179815
## 171     0.8017061              0.8414807              0.2807614
## 172     0.7403101              0.8203650              0.2846461
## 173     0.8194622              0.7470044              0.3622428
##     perc_employed_fulltime_yearround perc_unemployed perc_college_jobs
## 168                        0.7687849       0.2363489         0.6798525
## 169                        0.6058012       0.3242259         0.3260464
## 170                        0.7406321       0.2067863         0.3928227
## 171                        0.7271024       0.1982939         0.2131006
## 172                        0.7809422       0.2596899         0.3483973
## 173                        0.6835719       0.1805378         0.7803185
##     perc_non_college_jobs perc_low_wage_jobs
## 168             0.2782434         0.08716058
## 169             0.5193282         0.05145295
## 170             0.4748271         0.13746574
## 171             0.5087367         0.15915810
## 172             0.5483503         0.19906500
## 173             0.1245406         0.02858310
```

```
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
##  Min.   :0.5743         Min.   :0.0000
##  1st Qu.:0.7741         1st Qu.:0.2090
##  Median :0.8319         Median :0.2862
##  Mean   :   Inf         Mean   :0.2874
##  3rd Qu.:0.8974         3rd Qu.:0.3623
##  Max.   :   Inf         Max.   :0.5518
##                         NA's   :1
##  perc_employed_fulltime_yearround perc_unemployed   perc_college_jobs
##  Min.   :0.5857                   Min.   :0.04383   Min.   :0.0633
##  1st Qu.:0.7009                   1st Qu.:0.15899   1st Qu.:0.2974
##  Median :0.7484                   Median :0.19723   Median :0.4160
##  Mean   :0.7476                   Mean   :0.21140   Mean   :0.4478
##  3rd Qu.:0.7896                   3rd Qu.:0.25229   3rd Qu.:0.6170
##  Max.   :1.0000                   Max.   :1.00000   Max.   :0.8383
##                                                     NA's   :1
##  perc_non_college_jobs perc_low_wage_jobs
##  Min.   :0.08278       Min.   :0.00000
##  1st Qu.:0.27995       1st Qu.:0.06957
##  Median :0.42020       Median :0.10857
##  Mean   :0.41498       Mean   :0.11481
##  3rd Qu.:0.52756       3rd Qu.:0.15353
##  Max.   :0.85364       Max.   :0.36566
##  NA's   :1             NA's   :1
```

What are the different categories of college majors?

```r
table(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                   Interdisciplinary
##                                    7                                   1
##                 Law & Public Policy                   Physical Sciences
##                                    5                                  10
##            Psychology & Social Work                      Social Science
##                                    9                                   9
```

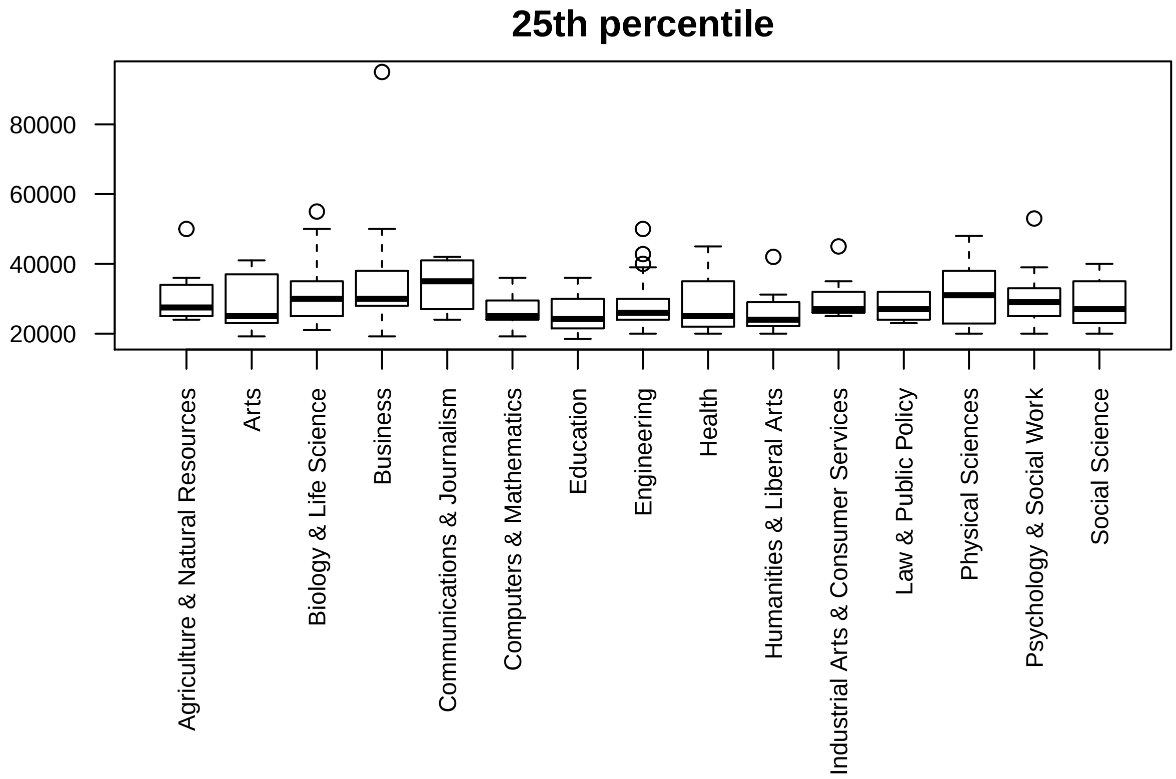
Only one major falls into the "Interdisciplinary" category. We certainly cannot estimate an effect for this category, so we will remove it.

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

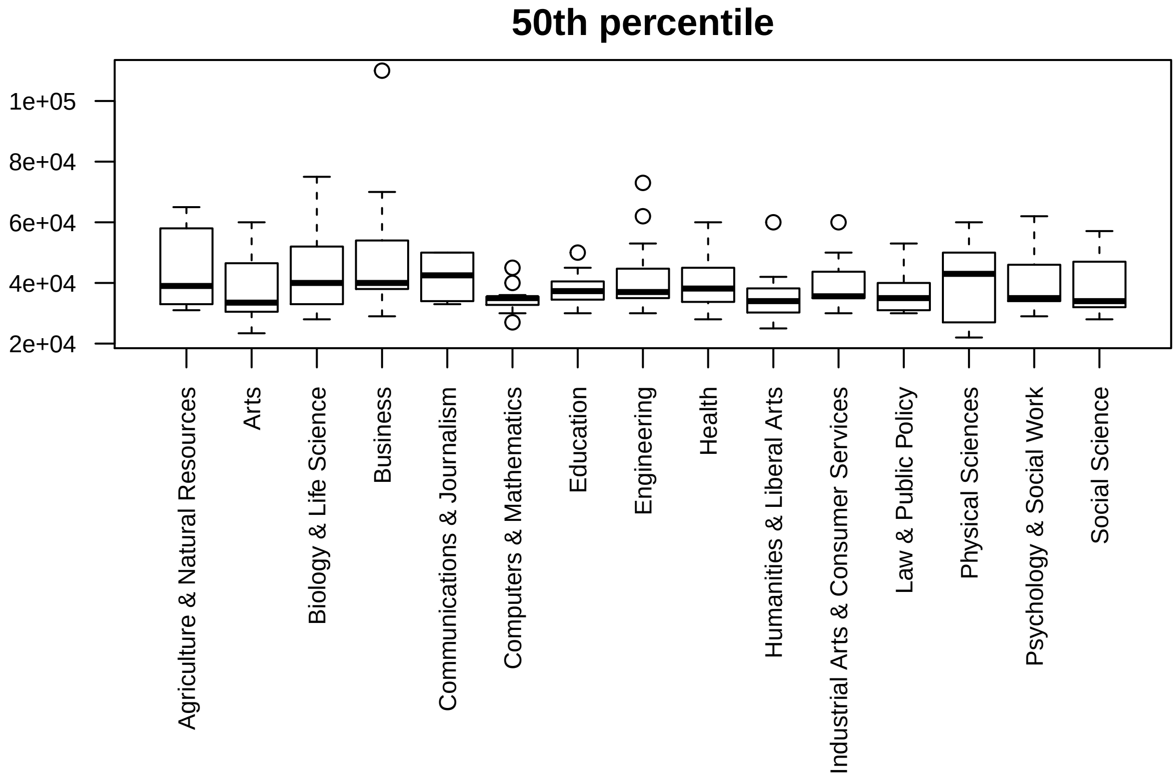
```
##
##     Agriculture & Natural Resources                                Arts
##                                  10                                   8
##              Biology & Life Science                            Business
##                                  14                                  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
```

Median income in a category is a useful measure because it indicates an income level that is typical in that category. Later we will used median income as our outcome measure in linear regression, but it is useful to look at plots of how the three different income measures (25th, 50th, and 75th percentile of income among those reporting income in the survey) vary across categories.

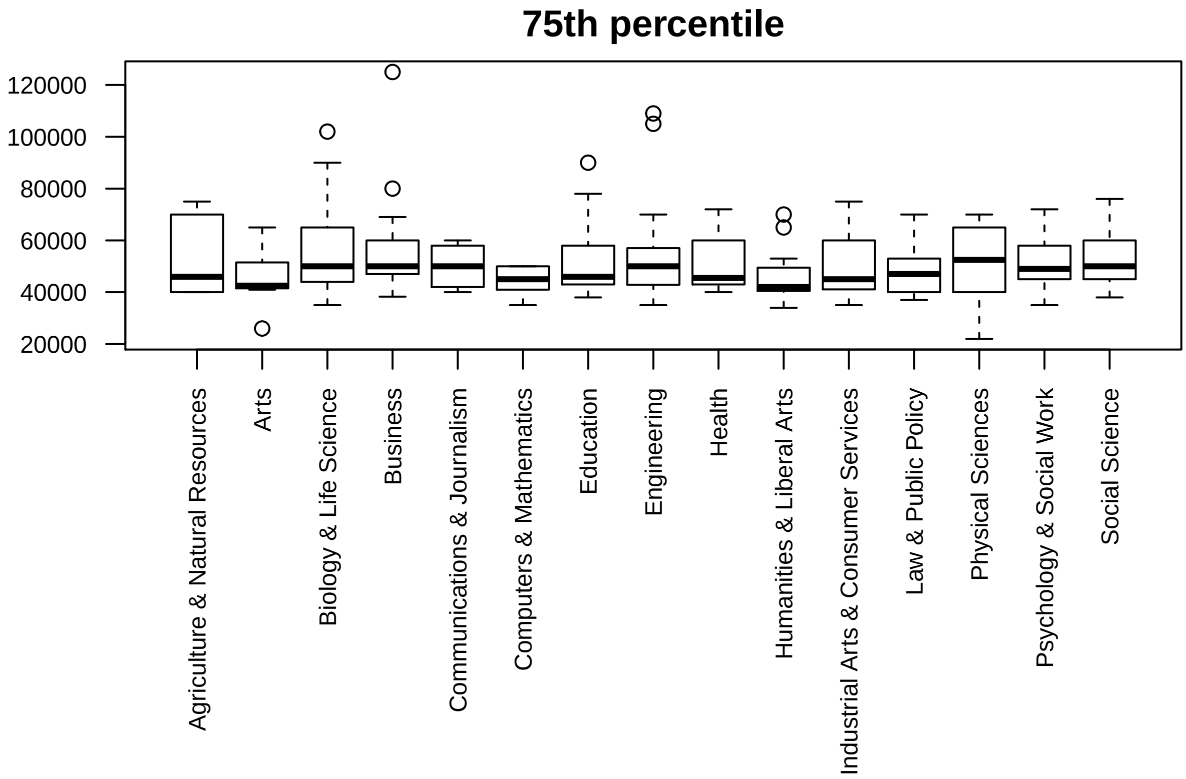
```r
par(mar = c(13,4.5,2,0.5))
boxplot(p25th ~ major_category, data = college, main = "25th percentile", las = 2, cex.axis = 0.75)
```



```
boxplot(median ~ major_category, data = college, main = "50th percentile", las = 2, cex.axis = 0.75)
```



```
boxplot(p75th ~ major_category, data = college, main = "75th percentile", las = 2, cex.axis = 0.75)
```



There doesn't seem to be considerable variation across categories for either of these three measures, so we will stick with using the median as our outcome measure.

Linear regression

Because the income information is defined to pertain to full-time, year-round workers, we will look at other characteristics: namely, gender effects and effects related to type of job (jobs requiring a college degree and jobs that are low-wage service positions). We can fit this linear model and view inference results with the following commands:

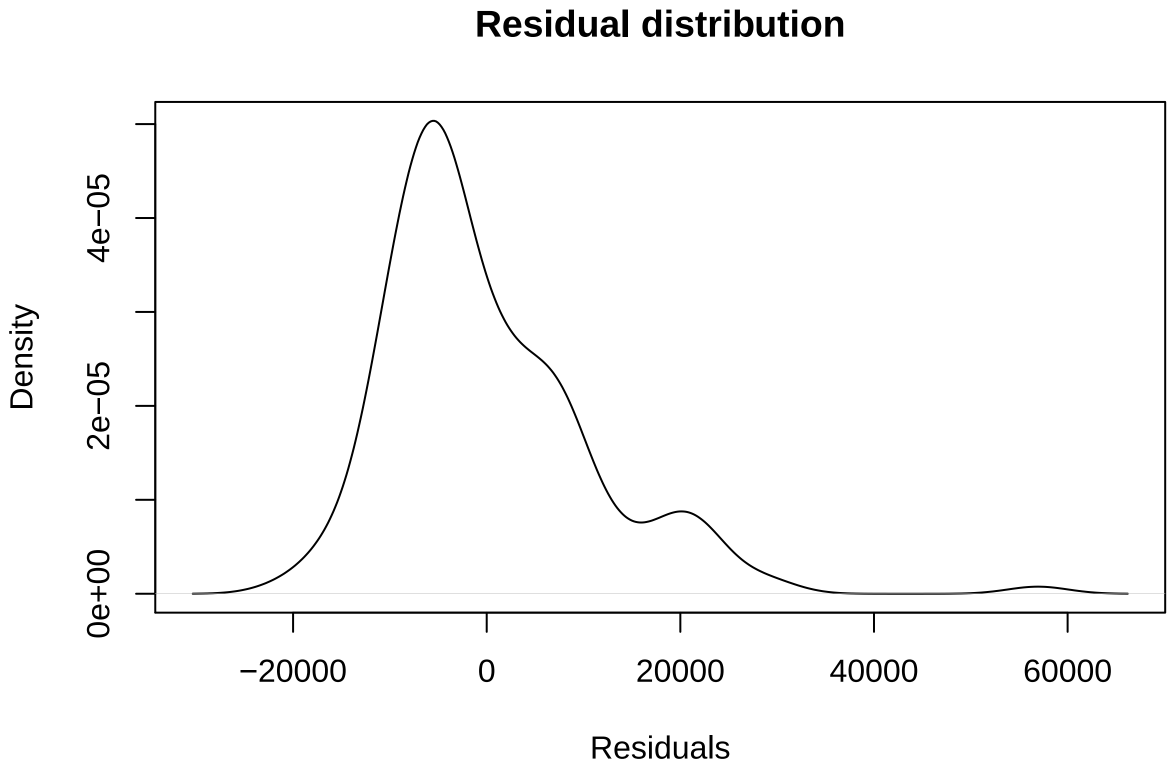
```
lmfit <- lm(median ~ major_category+perc_women+perc_college_jobs+perc_low_wage_jobs, data = college)
summary(lmfit)
```

```
##
## Call:
## lm(formula = median ~ major_category + perc_women + perc_college_jobs +
##     perc_low_wage_jobs, data = college)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -21088  -6917  -3211   5965  56932
##
## Coefficients:
##                                                    Estimate Std. Error
## (Intercept)                                         49491.3     5647.2
## major_categoryArts                                  -5429.2     5435.1
## major_categoryBiology & Life Science                  707.2     4772.0
## major_categoryBusiness                               5540.4     4846.5
## major_categoryCommunications & Journalism           -2776.6     6761.1
## major_categoryComputers & Mathematics              -9607.2     5129.6
## major_categoryEducation                            -5140.6     4591.3
## major_categoryEngineering                          -2931.4     4185.3
## major_categoryHealth                               -3700.9     4880.2
## major_categoryHumanities & Liberal Arts            -9022.7     4711.1
## major_categoryIndustrial Arts & Consumer Services  -2731.0     5604.6
## major_categoryLaw & Public Policy                  -5612.6     6374.9
## major_categoryPhysical Sciences                    -3268.4     5109.4
## major_categoryPsychology & Social Work             -5102.7     5340.0
## major_categorySocial Science                       -3059.1     5287.2
## perc_women                                         -5011.5     3958.1
## perc_college_jobs                                  -7841.4     5283.5
## perc_low_wage_jobs                                  2108.6    16018.8
##                                                   t value Pr(>|t|)
## (Intercept)                                         8.764 3.36e-15 ***
## major_categoryArts                                 -0.999   0.3194
## major_categoryBiology & Life Science                0.148   0.8824
## major_categoryBusiness                              1.143   0.2548
## major_categoryCommunications & Journalism          -0.411   0.6819
## major_categoryComputers & Mathematics              -1.873   0.0630 .
## major_categoryEducation                            -1.120   0.2646
## major_categoryEngineering                          -0.700   0.4847
## major_categoryHealth                               -0.758   0.4494
## major_categoryHumanities & Liberal Arts            -1.915   0.0573 .
## major_categoryIndustrial Arts & Consumer Services  -0.487   0.6268
## major_categoryLaw & Public Policy                  -0.880   0.3800
## major_categoryPhysical Sciences                    -0.640   0.5233
## major_categoryPsychology & Social Work             -0.956   0.3408
## major_categorySocial Science                       -0.579   0.5637
## perc_women                                         -1.266   0.2074
## perc_college_jobs                                  -1.484   0.1398
## perc_low_wage_jobs                                  0.132   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
```

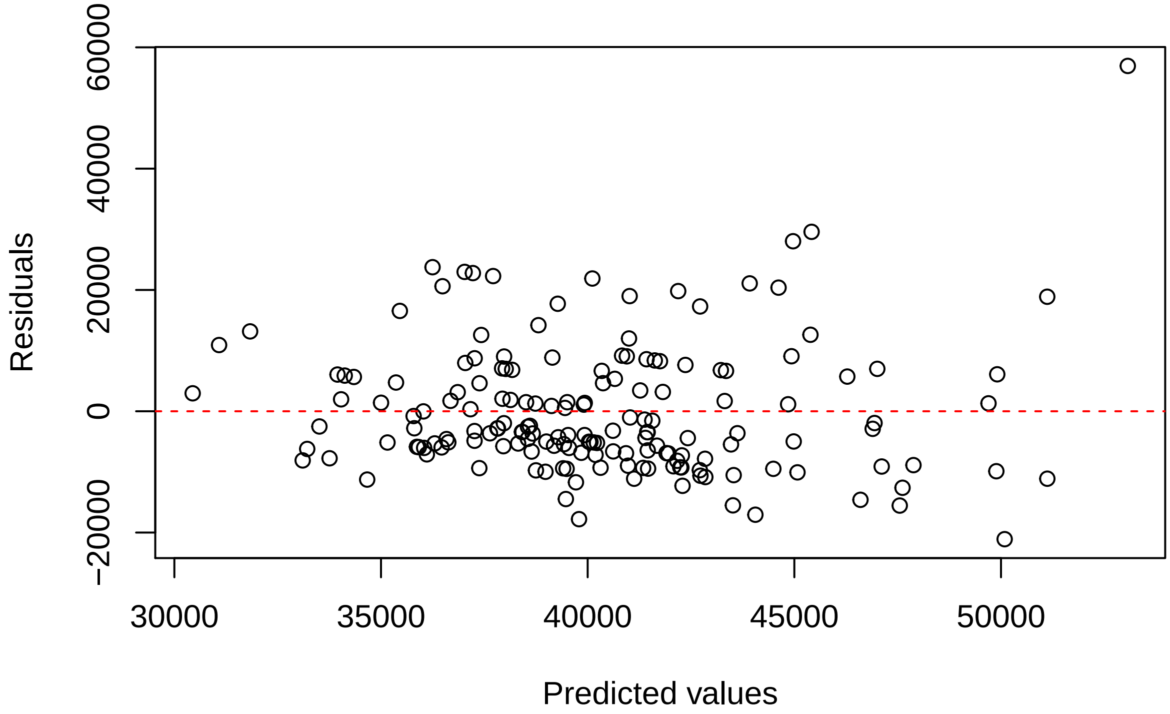
Holding constant gender distributions and skill category distributions, we don’t see much effect of major category (reference category is Agriculture & Natural Resources - how can we tell?). In particular, it is important to consider the multiple hypothesis testing issue here as we have many different major categories. Considering this, we really don’t see much of a category effect on income. Further the F-statistic results at the bottom do not suggest that the variables included have an impact on median income.

We can check if regression assumptions are met with diagnostic plotting:

```
resid <- residuals(lmfit)
fitted <- fitted.values(lmfit)
plot(density(resid), xlab = "Residuals", ylab = "Density", main = "Residual distribution")
```



```r
plot(fitted, resid, xlab = "Predicted values", ylab = "Residuals")
abline(h = 0, col = "red", lty = "dashed")
```



Normality assumptions don't seem far off (approximately), and heteroskedasticity doesn't seem to be an issue. Perhaps there is one outlier, but that is unlikely to have changed the overall results too much.

Overall there doesn't seem to be an effect of college major category on median income in this study.

FYI: A slightly neater display of the inference results for the coefficients can be obtained with the `tidy` function in the `broom` package:

```
tidy(lmfit)
```

```
##                                                  term   estimate std.error
## 1                                         (Intercept) 49491.2558  5647.217
## 2                                  major_categoryArts -5429.2005  5435.121
## 3                major_categoryBiology & Life Science   707.2026  4771.954
## 4                              major_categoryBusiness  5540.3971  4846.530
## 5           major_categoryCommunications & Journalism -2776.6432  6761.149
## 6               major_categoryComputers & Mathematics -9607.2267  5129.631
## 7                             major_categoryEducation -5140.5741  4591.264
## 8                           major_categoryEngineering -2931.4357  4185.267
## 9                                major_categoryHealth -3700.8590  4880.207
## 10            major_categoryHumanities & Liberal Arts -9022.6778  4711.130
## 11 major_categoryIndustrial Arts & Consumer Services -2730.9600  5604.552
## 12                major_categoryLaw & Public Policy -5612.5815  6374.887
## 13                  major_categoryPhysical Sciences -3268.3799  5109.439
## 14           major_categoryPsychology & Social Work -5102.6824  5340.003
## 15                     major_categorySocial Science -3059.1373  5287.175
## 16                                       perc_women -5011.5115  3958.062
## 17                                perc_college_jobs -7841.4095  5283.453
## 18                               perc_low_wage_jobs  2108.6219 16018.828
##     statistic      p.value
## 1   8.7638314 3.360373e-15
## 2  -0.9989106 3.194152e-01
## 3   0.1481998 8.823802e-01
## 4   1.1431679 2.547545e-01
## 5  -0.4106762 6.818845e-01
## 6  -1.8728883 6.299166e-02
## 7  -1.1196424 2.646207e-01
## 8  -0.7004178 4.847302e-01
## 9  -0.7583406 4.494137e-01
## 10 -1.9151832 5.733484e-02
## 11 -0.4872753 6.267607e-01
## 12 -0.8804206 3.800124e-01
## 13 -0.6396749 5.233406e-01
## 14 -0.9555580 3.408029e-01
## 15 -0.5785958 5.637127e-01
## 16 -1.2661529 2.073828e-01
## 17 -1.4841448 1.398277e-01
## 18  0.1316340 8.954468e-01
```