The Gender Pay Gap

Conducting a Multivariate Analysis of the 2017 National Longitudinal Survey of Youth

Ben Christensen

12/16/2020

The average wages of men and women in the National Longitudinal Survey of Youth (NLSY) differ significantly. One explanation is gender discrimination. Here are other possible explanations:

- Women may be more likely to stay home with children, losing work experience and forgoing pay
- Women may be less likely to request a pay raise due to psychological differences
- Women may be more likely to choose occupations that have lower incomes
- Average height differences between men and women may influence the difference in incomes because tall people could be more likely to be promoted

Data processing and summarization

Gender and Income

Income The dependent variable is income. Those who did not receive income from a job are given a value of -4 (Valid Skip) for the dependent variable. For this reason, I replace Income with 0 for every person who did not receive an income in the previous year. This decision is further discussed in the final discussion portion of the paper.

Other people are unsure of the exact income they earned. These were asked to estimate which category of income they fell in:

1. $1 - $5,000
2. $5,001 - $10,000
3. $10,001 - $25,000
4. $25,001 - $50,000
5. $50,001 - $100,000
6. $100,001 - $250,000
7. More than $250,000

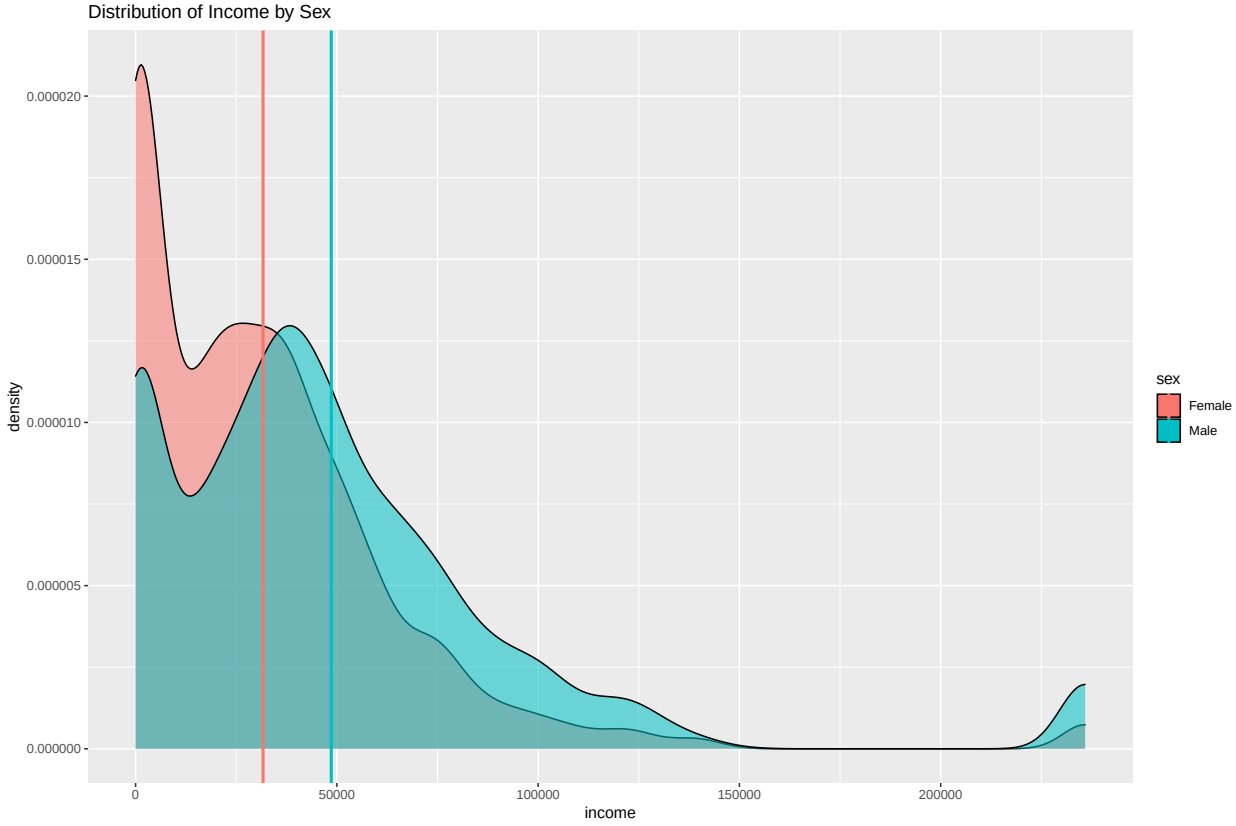
I use the median income for each category as the income estimate and replace missing incomes with this estimate when possible. The income variable is top-coded at $149,000. Anyone with a higher value than the top code is assigned the average income of everyone above the top-code: $235884. For those in the 6th and 7th categories of estimated income, the median is greater than $235884. So I set these to $235884 to match the other observations that surpassed the top-code. Those remaining with a negative value for income are assigned an income of NA and removed from the data set. That leaves 6634 observations in the data.

Gender Every person in the data set has a value for sex, either 1 for Male or 2 for Female. I recode the variable so each observation is 'Male' or 'Female' instead of 1 or 2.

The key question in this analysis is how income varies across sex. Using a t-test, the gender pay gap in the data is $16945.02 (favoring men) with a 95% confidence interval of ($14963.31 to $18926.73).

sex	average.income	count
Female	31665	3411
Male	48610	3223

The following density plot shows the distribution of income by sex. The mean income for each sex is shown by a vertical line. Both distributions are skewed right with higher average income for men than women. Nearly twice as many women have near-zero incomes. About twice as many men have top-coded incomes. For almost every income level above women's average income, more men than women earn at that level.



Other Variables

I renamed variables to more meaningful names as follows:

YINC-1700_2017 -> income

KEY!SEX -> sex

PSTRAN_GPA.01_PSTR -> GPA

CV_BIO_CHILD_HH_2007 -> bio.children.2007

CV_BIO_CHILD_HH_2009 -> bio.children.2009

CV_BIO_CHILD_HH_2011 -> bio.children.2011

CV_BIO_CHILD_HH_2015 -> bio.children.2015

CV_HH_NET_WORTH_P_1997 -> HH.net.worth.1997

CV_HGC_BIO_DAD_1997-> bio.dad.degree

CV_HGC_BIO_MOM_1997 -> bio.mom.degree

CV_HGC_RES_DAD_1997 -> res.dad.degree

CV_HGC_RES_MOM_1997 -> res.mom.degree

CVC_SAT_MATH_SCORE_2007_XRND -> math.SAT

CVC_SAT_VERBAL_SCORE_2007_XRND -> verbal.SAT

CV_HIGHEST_DEGREE_1112_2011 -> highest.degree

CV_YTH_REL_HH_CURRENT_1997 -> parents

CV_MARSTAT_COLLAPSED_2017 -> marriage.status

YEMP_OCCODE-2002.01_2017 -> occupation

YTEL-52~000001_2007 -> agreeable.1

YTEL-52~000002_2007 -> agreeable.2

YTEL-52~000003_2007 -> agreeable.3

YTEL-52~000004_2007 -> agreeable.4

YSAQ-000A000001_2011 -> height.feet

YSAQ-000A000002_2011 -> height.inches

Detailed information about these variables is shown in the next section.

Methodology

Variable Exploration

We need to find variables that could plausibly cause income variation and that are highly correlated with gender in the data set. This will enable a discussion on whether other characteristics cause the income variation across gender. Then we can see if controlling for these variables changes the effect gender has on income.

Numeric Variables The following are numeric variables that correlate highly with income – for which we may find causal relationships plausible (correlations shown in parentheses):

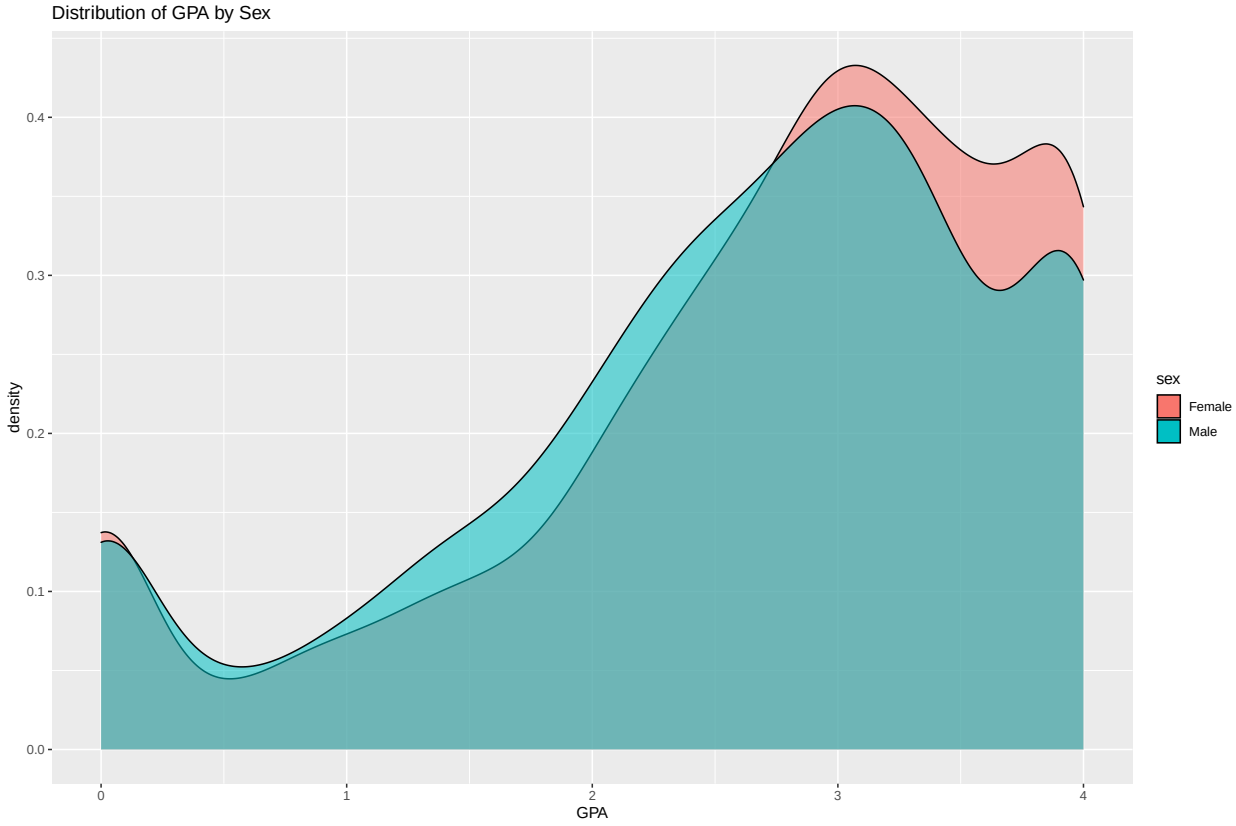
- GPA (0.26)
- Number of biological children (0.21)
- Household net worth in 1997 (0.2)
- Parent's degree (0.18)

The following numeric variable has a low correlation but may still be of interest:

- Height (0.04)

GPA The GPA variable has two implied decimal places, so a 2.75 GPA is written as '275'. For this reason, I divide the variable by 100. Those with a GPA higher than 4.0 were assigned a value of -3, so I replace the -3 with 4.00. 175 of these observations were Female and 148 were Male. Other negative values are reassigned as NA.

There are 3352 non-missing observations for GPA.

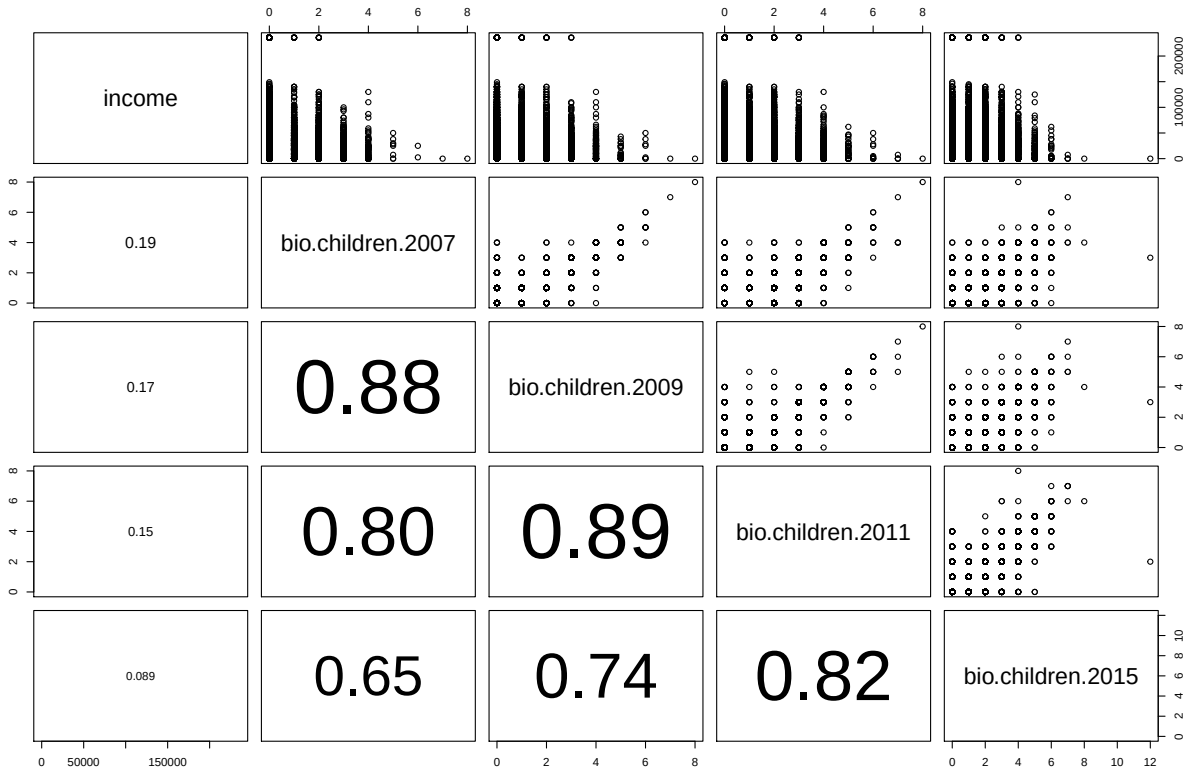


Number of Children There are four variables that measure the number of biological children the respondent has in the household. These measure the same thing but were recorded at different times: 2007, 2009, 2011, and 2015. Those without any biological children are coded as -4, meaning those with a value of 0 have children, just not living with them. For our purposes, these two cases can be treated the same. It is plausible that each would have more time for a full-time job than respondents that have children living with them. So I recode each -4 to a 0. Other negative values are re-coded to NA.

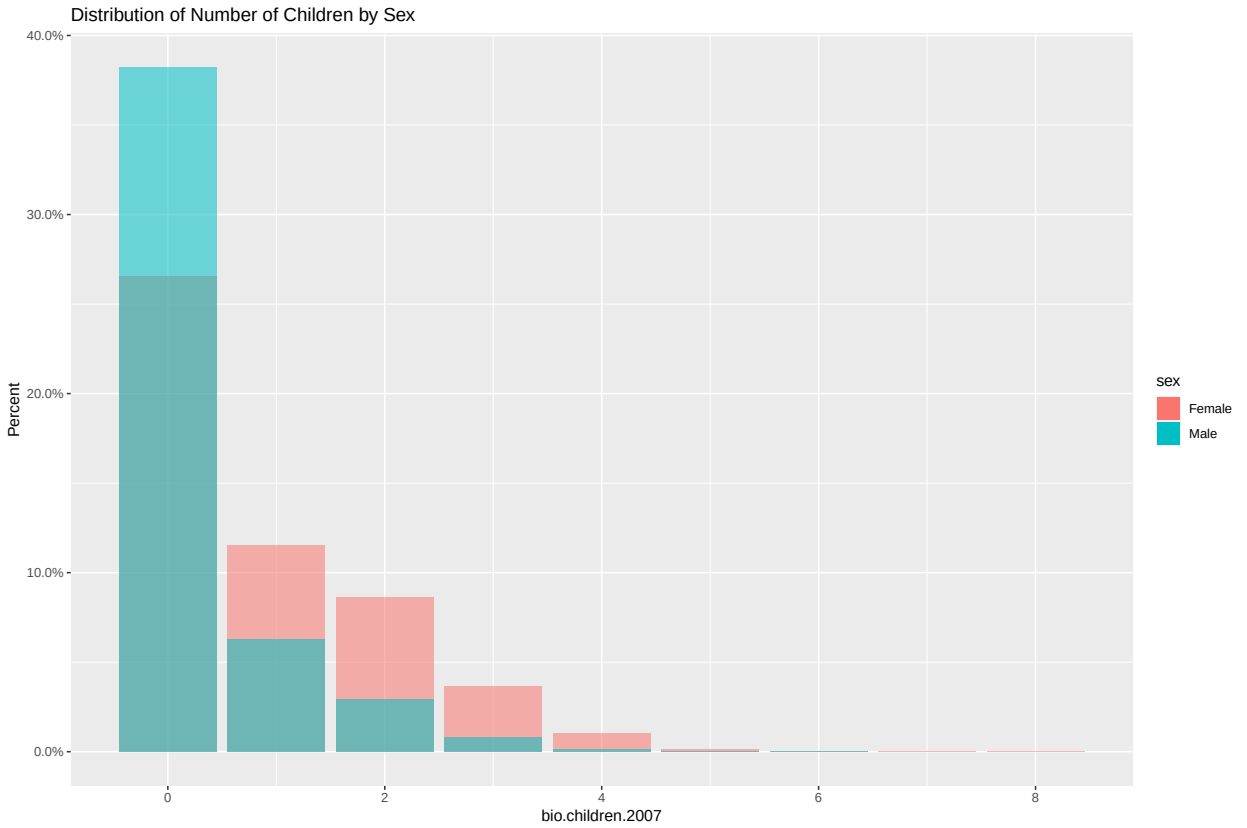
There are 6133, 6265, 6271, and 6259 non-missing observations for biological children for 2007, 2009, 2011, and 2015, respectively.

Now, for variable choice I've constructed a pairs plot to show the correlation between each variable and income.

The variables are highly correlated, with the 2007 measure most correlated with income. For that reason, I use bio.children.2007 and exclude the others from analysis.



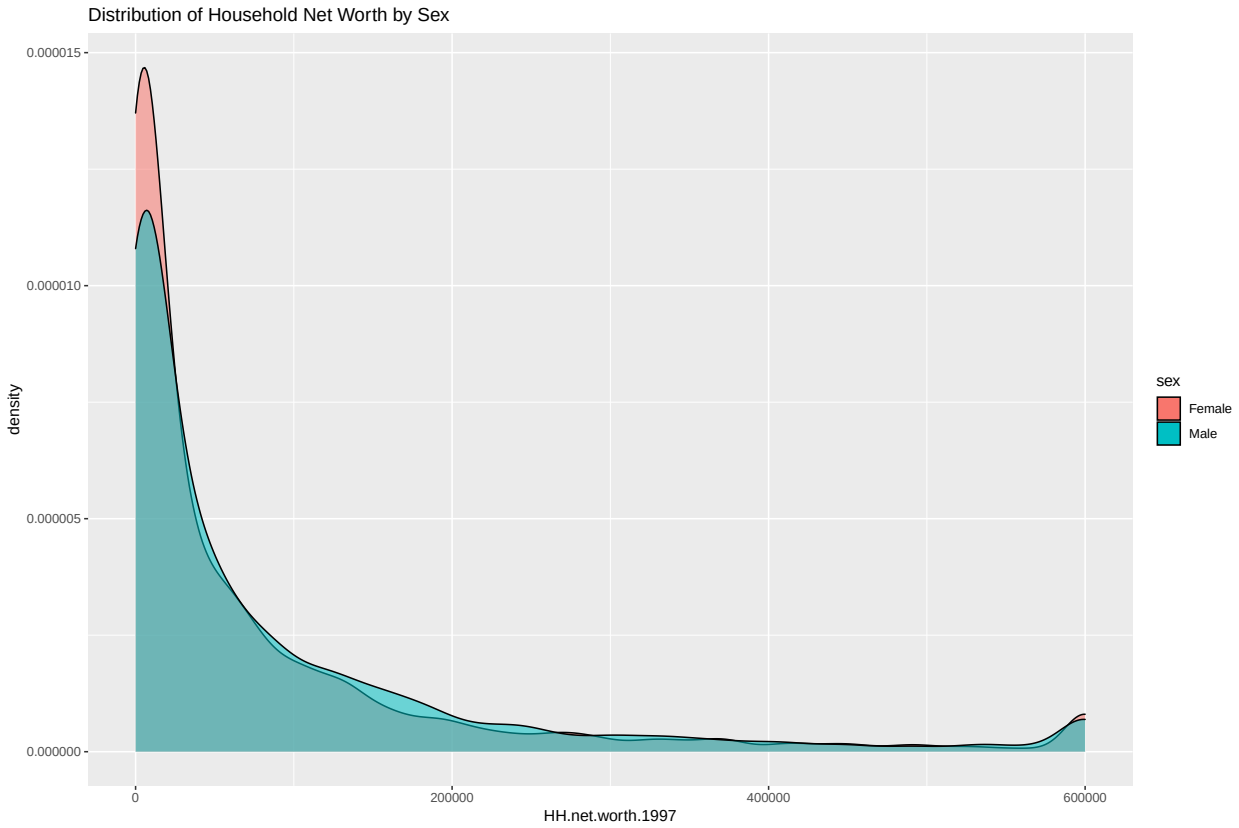
The following graph shows that a larger percent of men than women in the data set have 0 children, suggesting this variable is a good candidate for linear regression.



Household Net Worth 1997 This variable is top-coded at $600,000. 1627 have skip-codes of -3 or -4. I recode valid skips of -4 to Household net worth of 0. The other skip-code is re-labeled as NA.

There are 5406 non-missing observations for this variable.

The following graph shows the distribution of Household net worth in 1997 by gender. Other than a larger percent of women growing up in low-net worth households, the variation across gender is minimal. However, I would still like to see if this variable has some effect on income when included in linear regression as an interactive variable with gender.



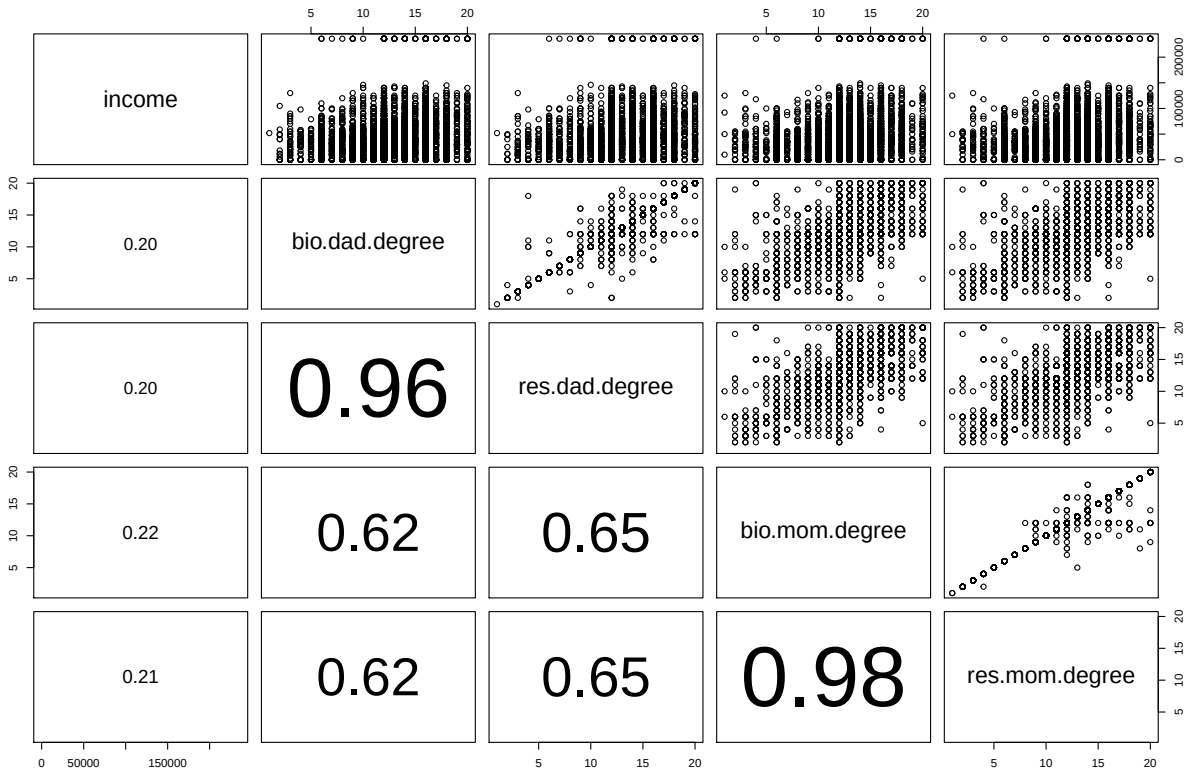
Parents' Education Levels There are four variables that measure a respondent's parents' education levels. They ask the same question, but they ask it of the biological father and mother and the residential father and mother. I mark valid skips, invalid skips, and 'UNGRADED' as NA. The valid skip entry for residential father or mother means that parent is not present in the household which may be important information, but this is at least partially captured by the parents variable. These variables take on a finite number of values, but each step represents one additional year of education, so I include it as a numerical variable.

0. NONE
1. 1ST GRADE
2. 2ND GRADE
3. 3RD GRADE
4. 4TH GRADE
5. 5TH GRADE
6. 6TH GRADE
7. 7TH GRADE
8. 8TH GRADE
9. 9TH GRADE
10. 10TH GRADE
11. 11TH GRADE
12. 12TH GRADE
13. 1ST YEAR COLLEGE
14. 2ND YEAR COLLEGE
15. 3RD YEAR COLLEGE
16. 4TH YEAR COLLEGE
17. 5TH YEAR COLLEGE

18. 6TH YEAR COLLEGE
19. 7TH YEAR COLLEGE
20. 8TH YEAR COLLEGE OR MORE

There are 5284 non-missing observations for each of these variables.

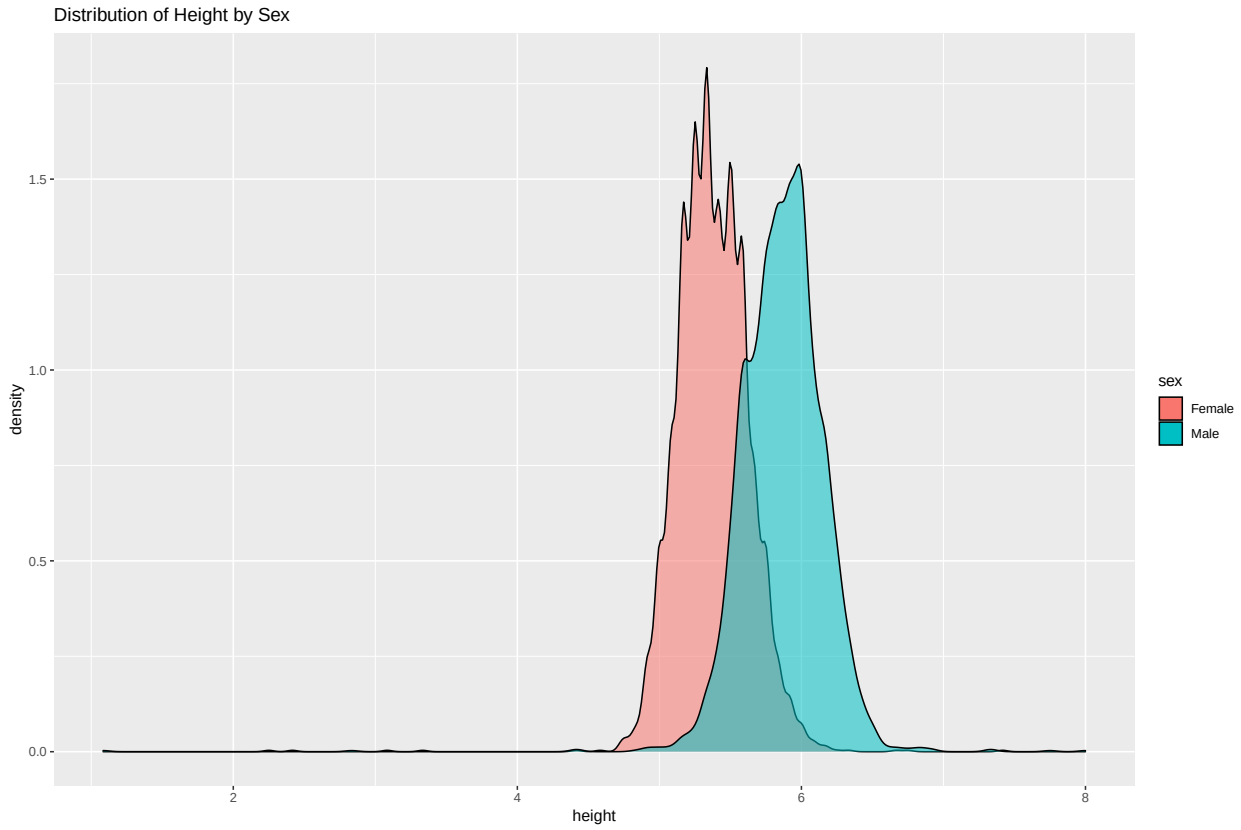
For variable choice I've constructed a pairs plot to show the correlation between the variables. The biological and residential versions of each are highly correlated. Even the correlations across father and mother's degrees are high. I'll include only the bio.mom.degree variable because it is most correlated with income.



Height The height variable is split into two variables: height in feet and additional height in inches. Meaning a respondent that is 5'3" would answer the 5 for the first question and 3 for the second. I combine the variables to get the full height and convert it to feet. 484 of the 489 respondents who did not give a response for the inches question also did not give a response for the feet question. I mark them all as NA.

There are 6142 non-missing observations for height.

The following plot shows men in the data set are taller than women which makes this a good candidate for linear regression.



Categorical Variables The following are categorical variables that are statistically significant predictors of income for which we may find causal relationships plausible (analysis of variance P-values shown in parentheses):

- Math SAT score (9.21e-124)
- Verbal SAT score (6.66e-119)
- Highest Degree earned (as of 2011) (1.85e-102)
- Parent Makeup (1.28e-43)
- Marriage Status (2.21e-09)
- Occupation (0.000179)

The following variable has a lower p-value, but may be of interest:

- Proxy for agreeableness (0.1)

SAT Scores The Math and Verbal SAT scores are grouped into 6 categories:

1. 200 - 300
2. 301 - 400
3. 401 - 500
4. 501 - 600

5. 601 - 700
6. 701 - 800

Those with invalid skip for the question were assigned -3 and those with a valid skip for the question were assigned -4. Both of these values may be interesting. A respondent who did not take the SAT (valid skip) may not intend to attend college (or they took a different exam). Those with an invalid skip may have refused to answer the question because their score was low. For these possibilites, I leave those codes in the data set instead of setting them to NA.

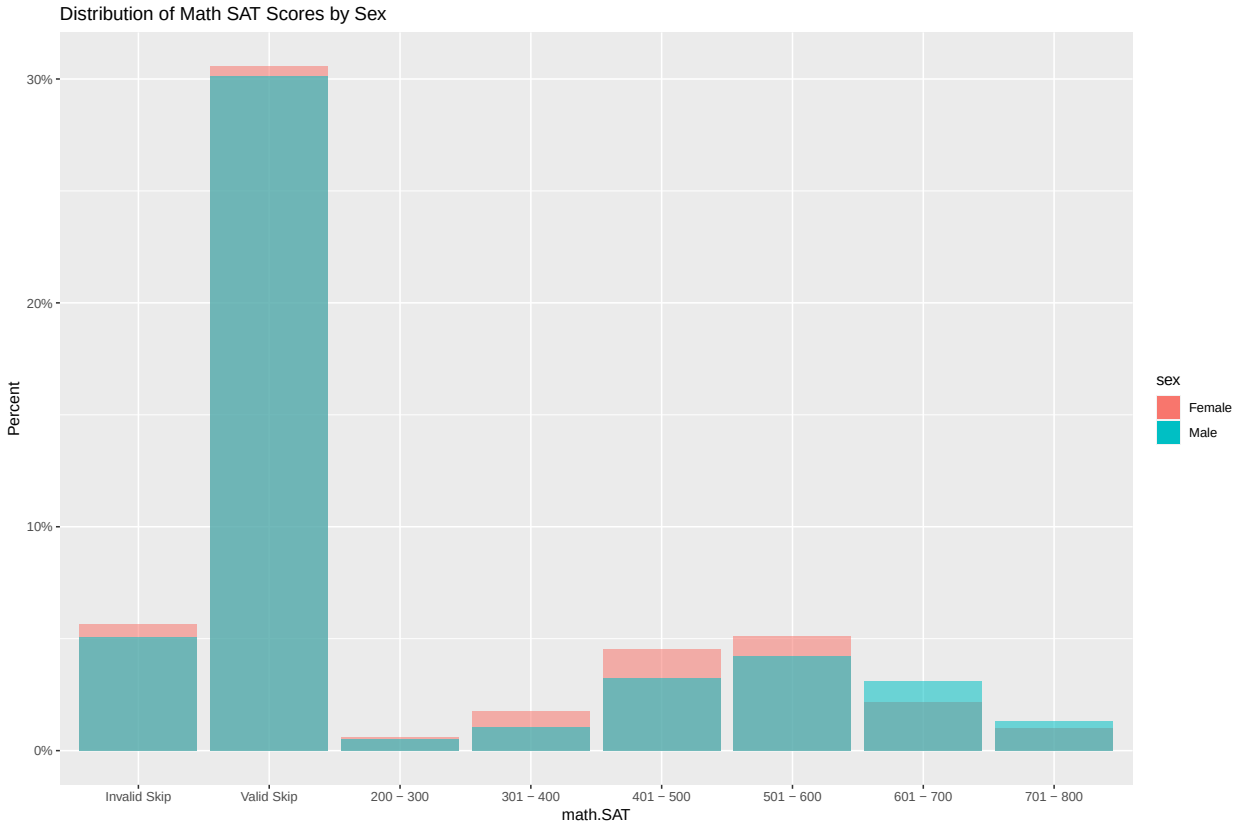
For math SAT scores, 1898 respondents gave a valid response, 710 were an invalid skip, 4026 and were a valid skip.

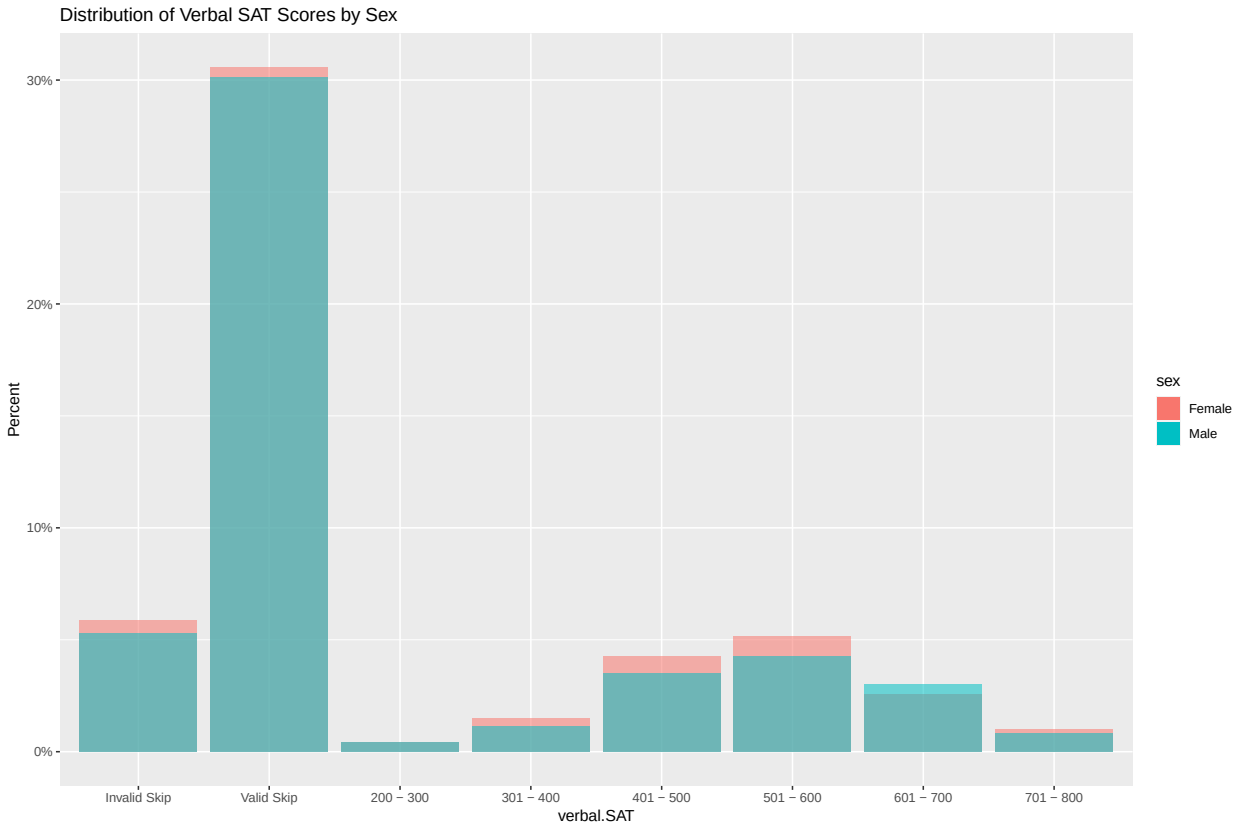
For verbal SAT scores, 1867 respondents gave a valid response, 741 were an invalid skip, 4026 and were a valid skip.

The number of valid responses are each larger than the 1563 available for ACT scores.

I recode the variables to string representations of the categories.

In the following two graphs we see that Math SAT scores seem to have a little more consistent gender variation than Verbal SAT scores. For this reason, I only include Math SAT scores in my final analysis.

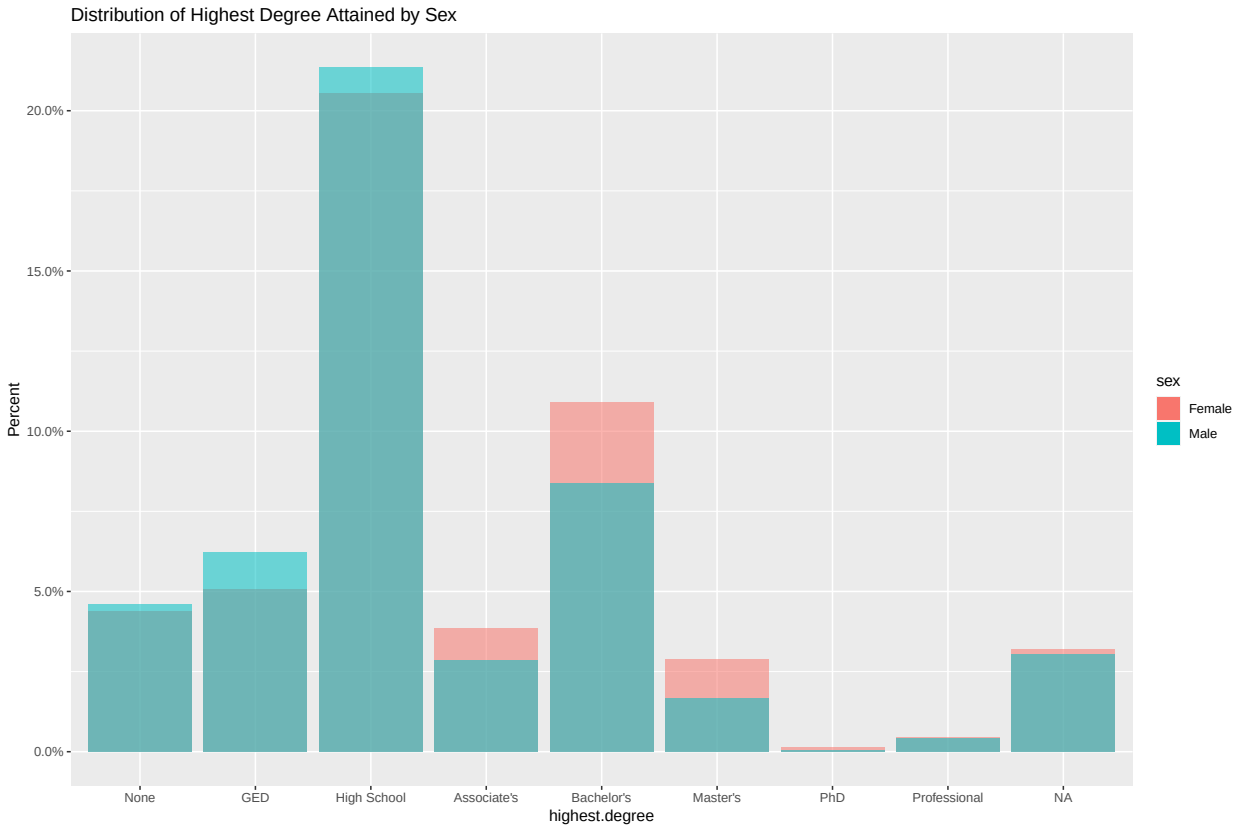




Highest Degree earned (as of 2011) This variable categorizes 'degree earned' as follows:

0. None
1. GED
2. High school diploma (Regular 12 year program)
3. Associate/Junior college (AA)
4. Bachelor's degree (BA, BS)
5. Master's degree (MA, MS)
6. PhD
7. Professional degree (DDS, JD, MD)

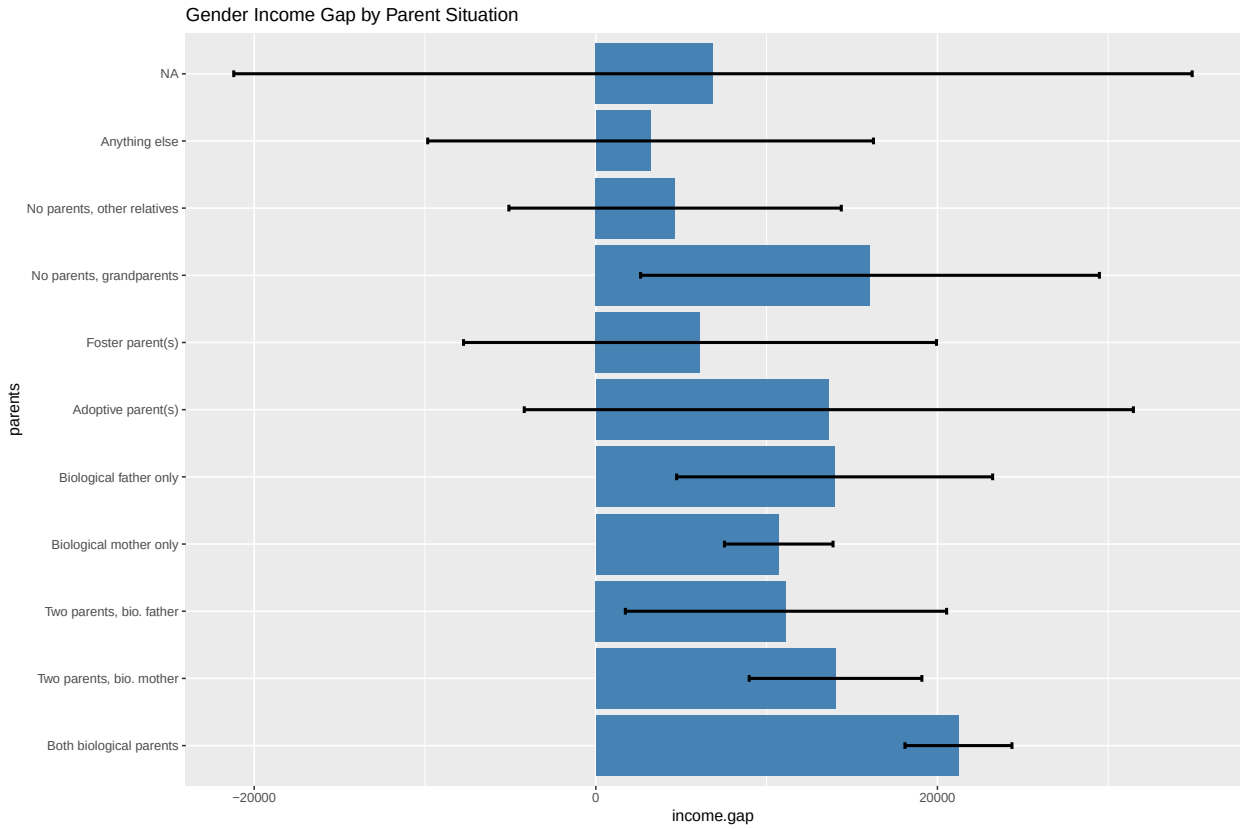
54 respondents were invalid skip. This does not seem large enough to justify leaving it as its own category in the data so I mark invalid skips as NA along with the 54 respondents that were not interviewed. I recode the rest to string values that represent their categories.



Parent Makeup This variable shows the relationship respondents had to the parent figures or guardians of their household in 1997. It has only 24 invalid skips and no other skips. These I code as NA. The other observations are grouped by the following categories. I recode the variable to string representations of the categories.

1. Both biological parents
2. Two parents, biological mother
3. Two parents, biological father
4. Biological mother only
5. Biological father only
6. Adoptive parent(s)
7. Foster parent(s)
8. No parents, grandparents
9. No parents, other relatives
10. Anything else

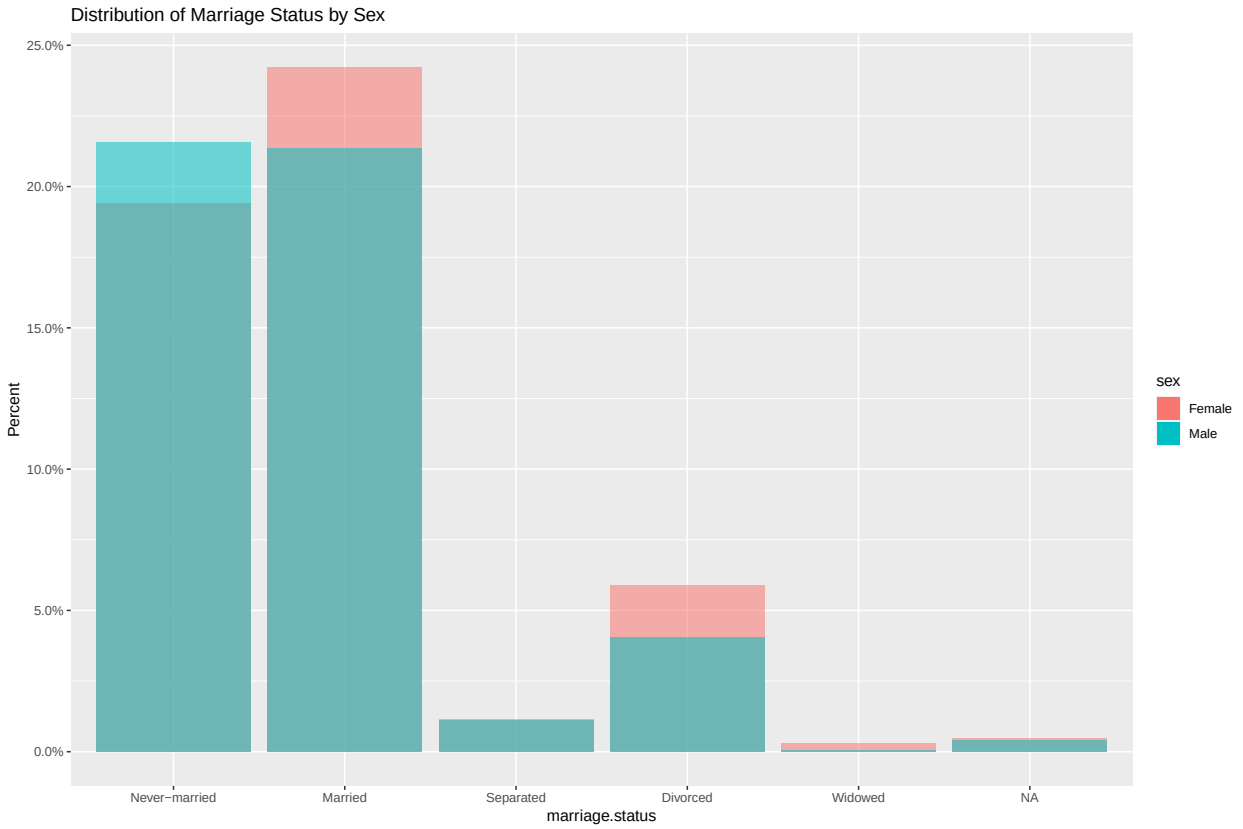
The following graph shows variation in the gender wage gap by parent makeup. Lower and upper bounds on the 95% confidence interval of this two-variable t-test are shown as black lines. Only a few groups have statistically different gender wage gaps: it is larger for individuals with both biological parents than those with only a biological mother and larger than those raised by parents other than relatives. This differences make it a good candidate for linear regression.

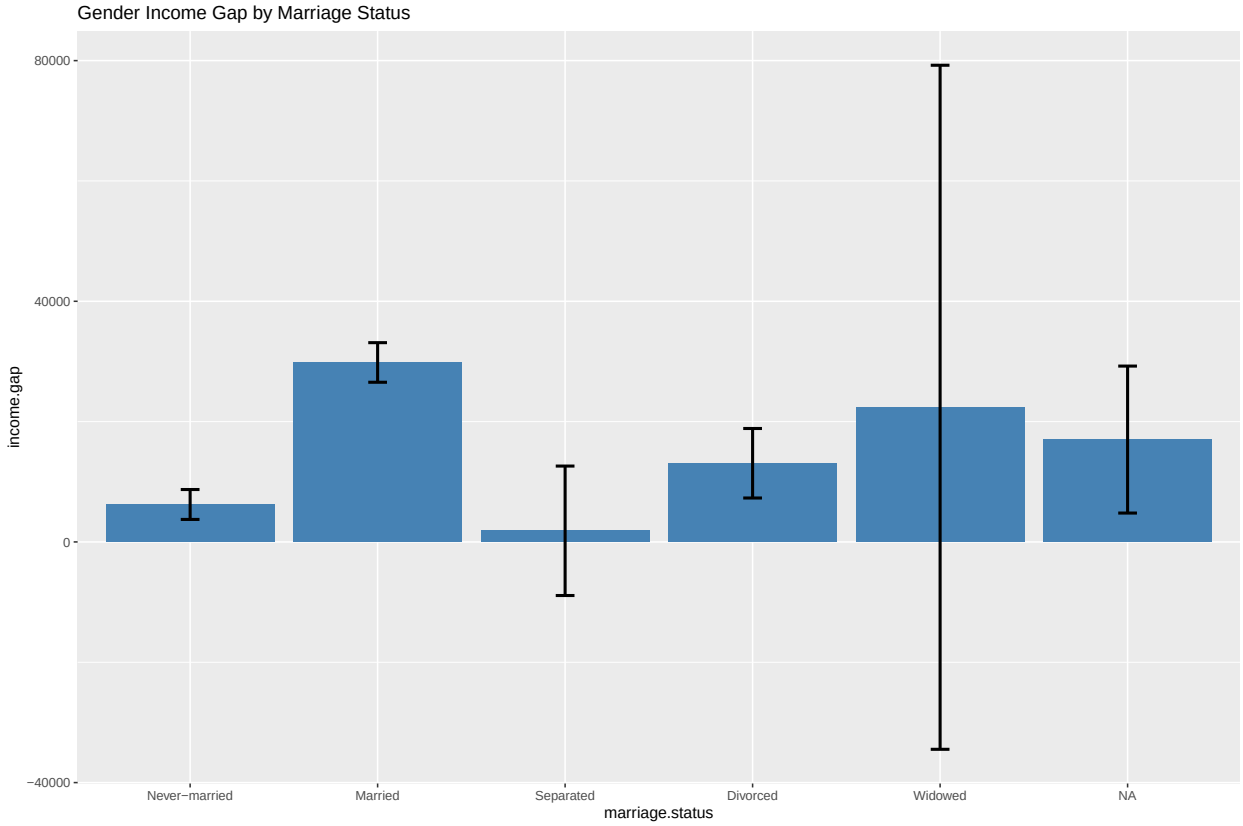


Marriage Status This variable records the marital status of the respondent as follows. Only 59 are invalid skips, and 0 were not interviewed. I code all of these as NA. Then I recode the variable to string representations of the categories.

0. Never-married
1. Married
2. Separated
3. Divorced
4. Widowed

There is some variation in marriage status distribution by gender, but the variation in wage gap by marital status is stronger evidence that this variable is a good candidate for linear regression. The error bars are very tight around the estimates for the gender wage gaps of married and never-married people.


Distribution of Marriage Status by Sex
25.0%
20.0%
15.0%
10.0%
5.0%
0.0%
Percent
Never−married
Married
Separated
Divorced
Widowed
NA
marriage.status
sex
Female
Male


Gender Income Gap by Marriage Status
80000
40000
0
−40000
income.gap
Never−married
Married
Separated
Divorced
Widowed
NA
marriage.status

Occupation There are 2002 Census Occupation Codes. They fall into the following categories. To these I add a category "NONE" for those that were a Valid Skip (-4). These eventually get re-coded to "OTHER" as explained further on. Respondents with any other negative value were re-assigned NA. I replace each code with the name of the category to which it belongs.

10 TO 430 : EXECS, ADMINS, MANAGERS

500 TO 950 : MANAGEMENT RELATED

1000 TO 1240: MATH, COMPUTER SCIENTISTS

1300 TO 1530: ENGINEERS, ARCHITECTS, SURVEYORS

1540 TO 1560: ENGINEERING, RELATED TECHNICIANS

1600 TO 1760: PHYSICAL SCIENTISTS

1800 TO 1860: SOCIAL SCIENTISTS, RELATED

1900 TO 1960: LIFE, PHYSICAL, SOCIAL SCIENCE TECH

2000 TO 2060: COUNSELORS, SOCIAL, RELIGIOUS

2100 TO 2150: LAWYERS, JUDGES, LEGAL SUPPORT

2200 TO 2340: TEACHERS

2400 TO 2550: EDUCATION, TRAINING, LIBRARY

2600 TO 2760: ENTERTAINERS, PERFORMERS, SPORTS

2800 TO 2960: MEDIA, COMMUNICATION

3000 TO 3260: HEALTH DIAGNOSIS AND TREATING

3300 TO 3650: HEALTH CARE TECHNICAL, SUPPORT

3700 TO 3950: PROTECTIVE SERVICE

4000 TO 4160: FOOD PREPARATIONS, SERVING

4200 TO 4250: CLEANING, BUILDING SERVICE

4300 TO 4430: ENTERTAINMENT ATTENDANTS, RELATED

4500 TO 4650: PERSONAL CARE, SERVICE

4700 TO 4960: SALES, RELATED

5000 TO 5930: OFFICE, ADMINISTRATIVE SUPPORT

6000 TO 6130: FARMING, FISHING, FORESTRY

6200 TO 6940: CONSTRUCTION TRADES, EXTRACTION

7000 TO 7620: INSTALLATION, MAINTENANCE, REPAIR

7700 TO 7750: PRODUCTION, OPERATING

7800 TO 7850: FOOD PREPARATION

7900 TO 8960: SETTER, OPERATORS, TENDERS

9000 TO 9750: TRANSPORTATION, MATERIAL MOVING

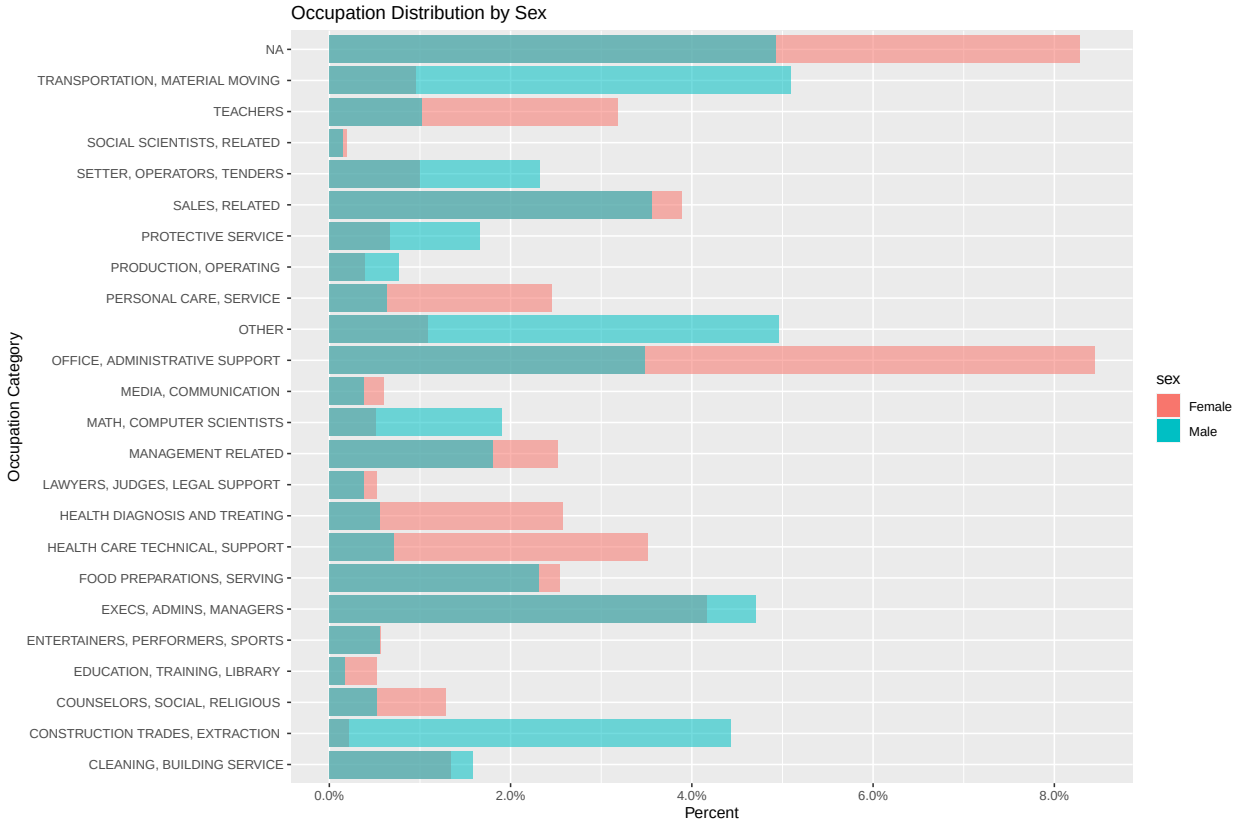
9800 TO 9840: MILITARY SPECIFIC OCCUPATIONS

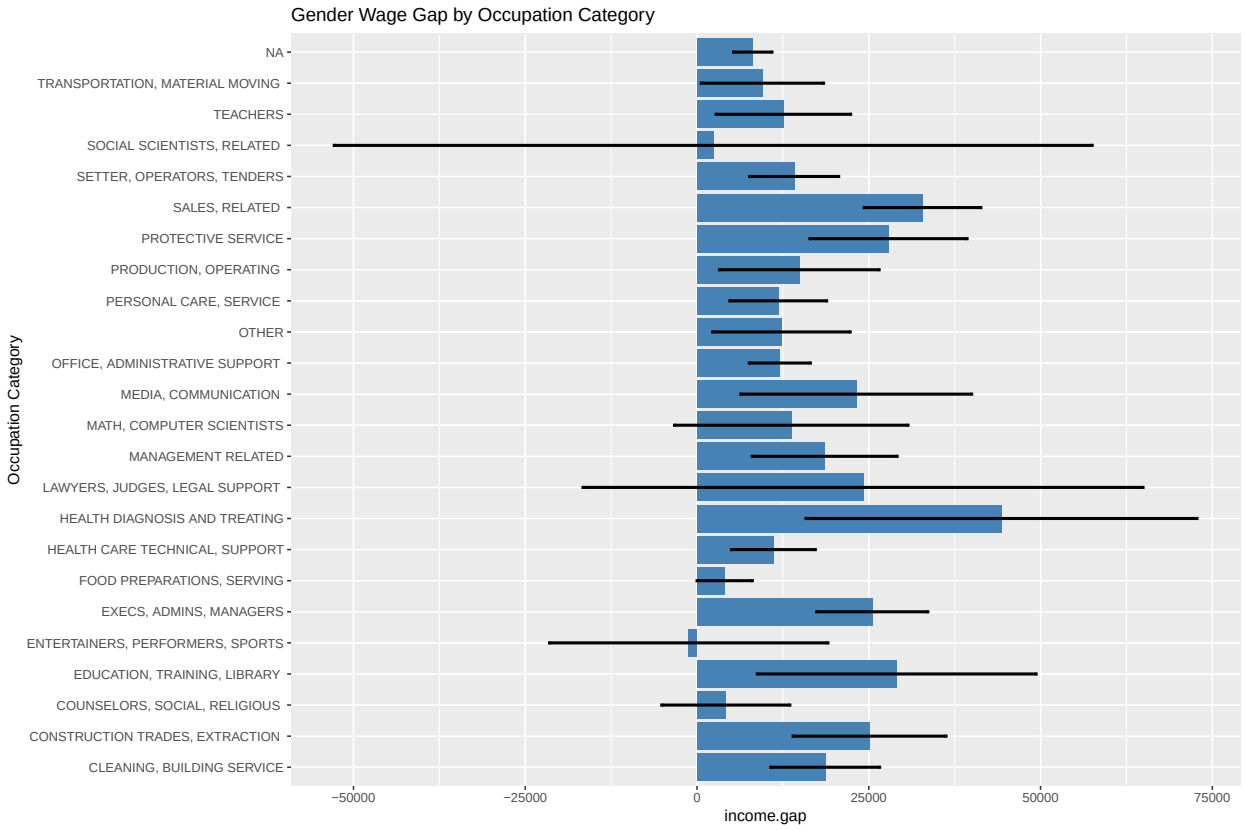
9950 TO 9990: ACS SPECIAL CODES

Before making further changes to the data there were was a category with an extremely large wage gap: "Life, Physical, and Social Science Technicians". I discovered that there were only 4 men in this category

and 1 woman. To avoid potentially misleading situations like this, I removed categories that had fewer than 10 men or 10 women and re-classified those respondents as 'OTHER'.

It is clear that women are more likely to work in some occupation categories than others. Also, there is significant gender wage gap variation for some categories (i.e. FOOD PREPARATIONS, SERVING vs. SALES, RELATED) making this a good candidate for linear regression.





Proxy for agreeableness There are four variables in the dataset where respondents were asked to express how much they agree or disagree with a given statement. I wanted to see if these questions impacted the effect gender had on income variation because they seem linked to the psychological "Big Five" personality dimension of agreeableness. This dimension is roughly approximate to how conflict-avoidant a person is.

Each statement had 4604 valid skip respondents. This may warrant exclusion in the final analysis. For now I recode all missing values as NA. These are the levels of agreement respondents may choose from. I recode the numbers to the string representation of each response.

0. Strongly Disagree
1. Disagree
2. Neither Agree nor Disagree
3. Agree
4. Strongly Agree

Looking at the distribution of responses, only the second and fourth statements had at least 10 respondents of each gender in every category, so I limit my analysis to these two variables. For each of these statements I expect agreeable people to disagree more often (using the rough approximation that compassion and agreeableness are the same). Based on my limited understanding of psychology, my hypothesis is that women are more agreeable than men.

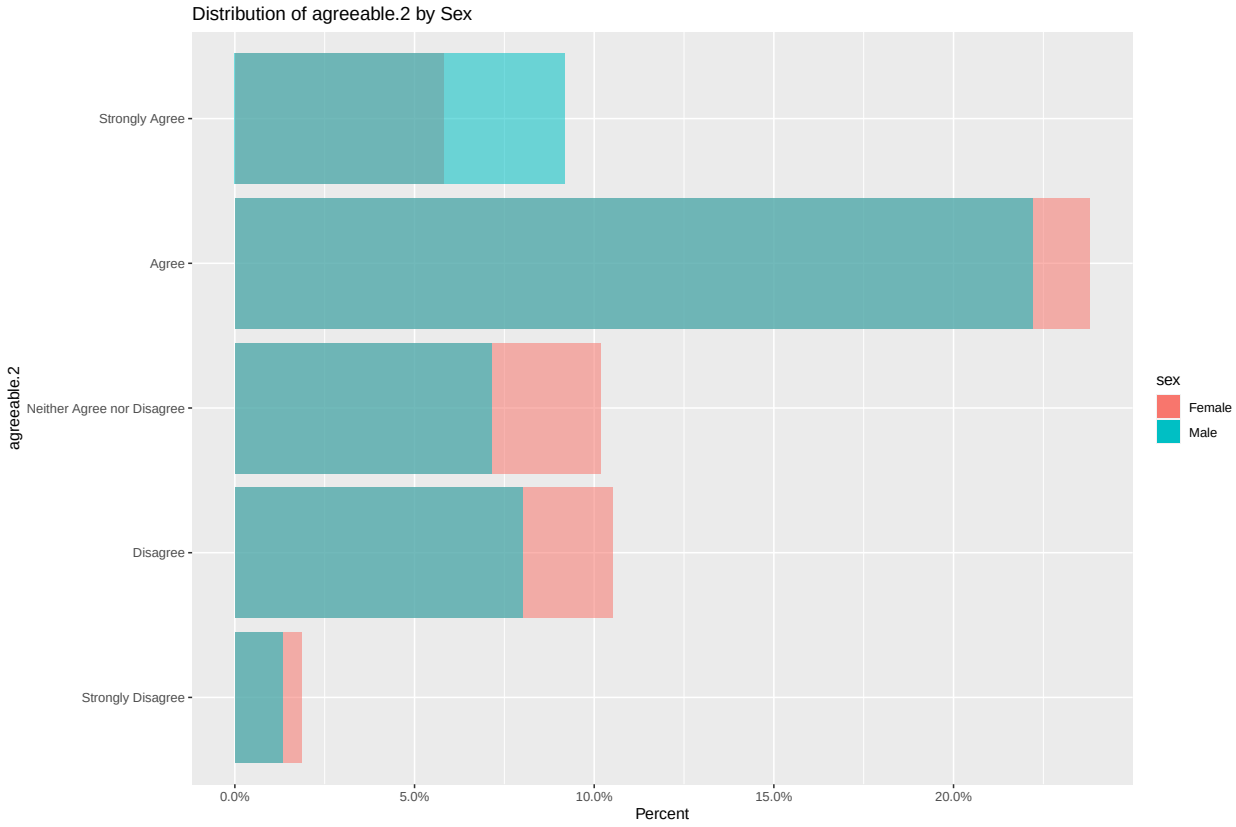
Statement 2:

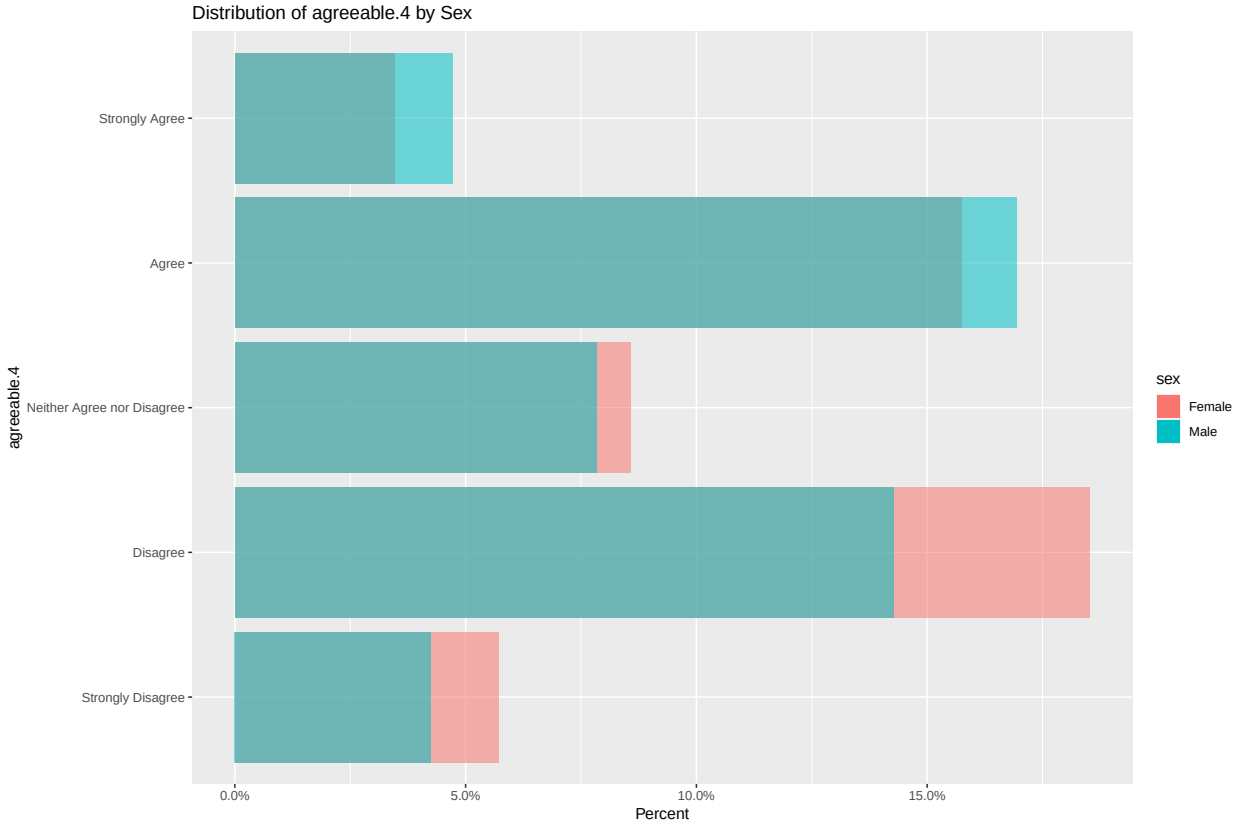
"Those in need have to learn to take care of themselves and not depend on others."

Statement 4:

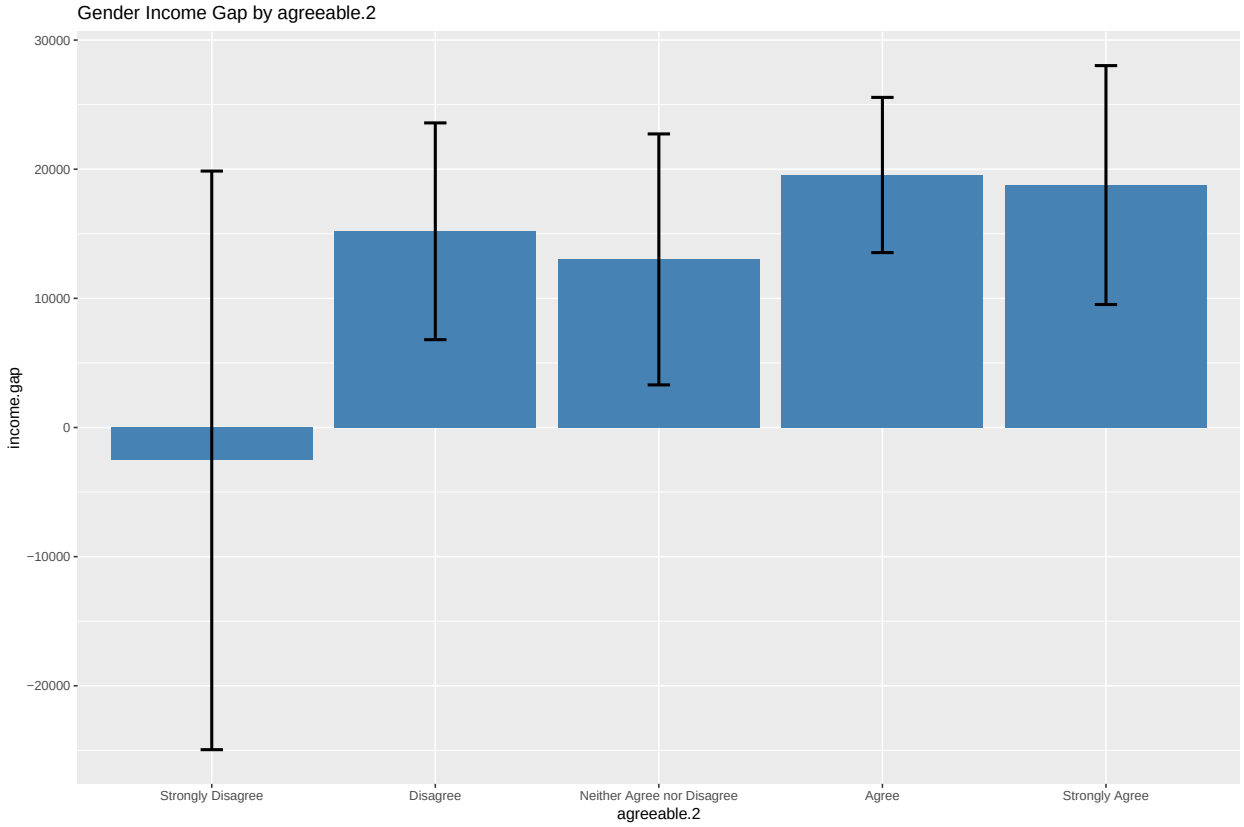
"These days people need to look after themselves and not overly worry about others."

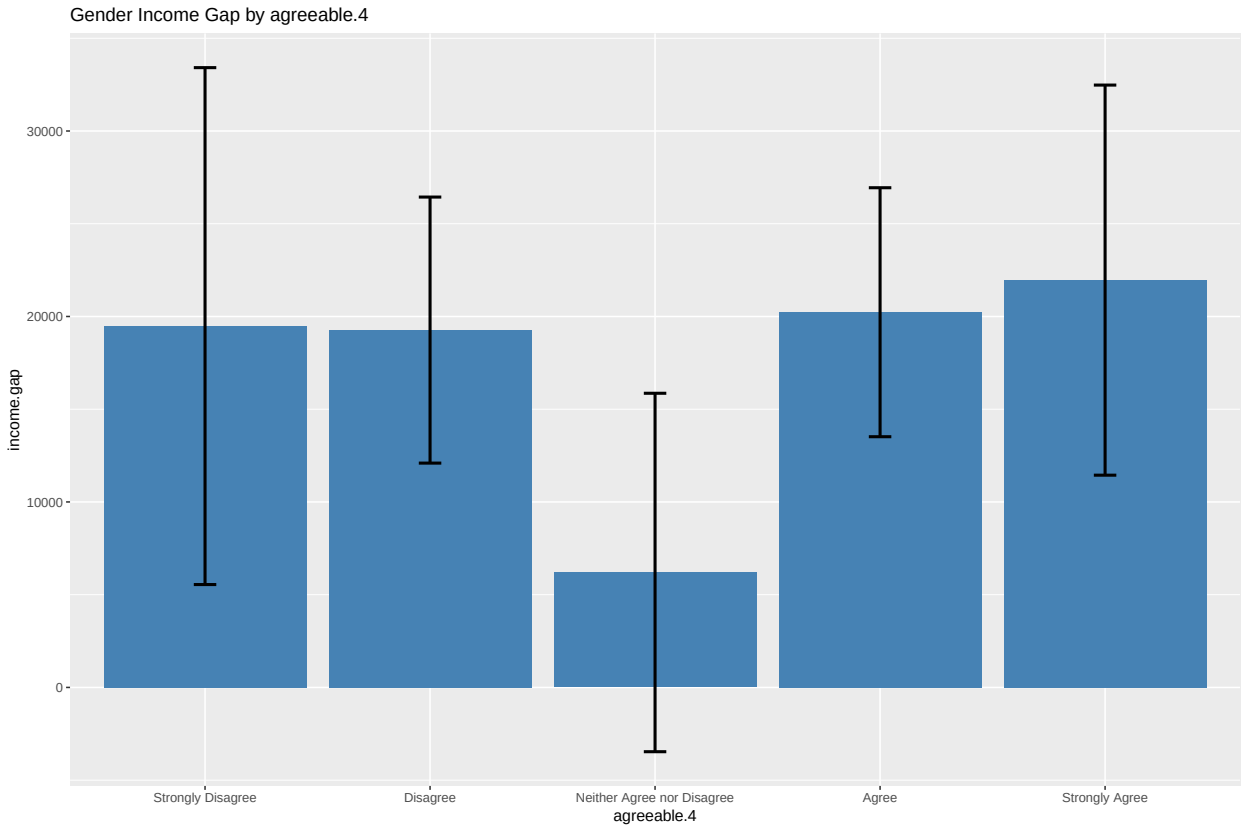
As we can see in the following graphs, men are more likely to “Strongly Agree” with these statements.





Now, let's look at the wage gap across each variable. Instead of providing evidence that 'disagreeableness' lessens the gender wage gap, these seem to suggest that it has no effect on the gender wage gap or perhaps even the opposite effect. Additionally, all of the error bars overlap. We will see more clearly what effect the variables have in the final analysis with linear regression.





Findings

Now for the linear regression analysis. At this point there are 6634 observations in the data set. After running a linear regression including the previously discussed variables, the proxy for agreeableness has no statistically significant categories and requires the omission of 5121 observations. The GPA variable is statistically significant, but it requires the omission of 3282 observations. Each variable caused a reduction in adjusted R-squared from 0.35 to 0.33 for the agreeableness variable and to 0.31 for GPA. This persuaded me to remove these variables. Only 2839 observations are deleted due to missingness in the resulting model.

I wanted to test if the following variables had interactive effects with gender. I updated the model with each interactive variable (i.e. bio.children.2007 * sex) and ran an ANOVA test. The third variable, marriage.status had a significant P-value for the ANOVA test, suggesting it added predictive power to the model, so I include it as an interaction term in the final model.

- bio.children.2007 (0.13)
- bio.mom.degree (0.11)
- marriage.status (3.59×10^{-12})
- HH.net.worth.1997 (0.18)

Regressing income on only gender gives the following results. The adjusted R-squared is 0.04 and the estimated effect of gender on income is 16945.02 which matches the results of the t-test we ran at the beginning of our analysis.

```
##
## Call:
## lm(formula = income ~ sex, data = nlsy)
```

```
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -48610 -29165  -6665  16335 204219
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    31666        699    45.3   <2e-16 ***
## sexMale        16945       1003    16.9   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 40800 on 6632 degrees of freedom
## Multiple R-squared:  0.0413, Adjusted R-squared:  0.0411
## F-statistic:  285 on 1 and 6632 DF,  p-value: <2e-16
```

Our final model improves on this simple approach by increasing the adjusted R-squared to 0.36. The new estimate for the effect of gender on income is 6907.76. That estimate means, all other model variables held equal, men are predicted to earn 6907.76 more than women. It is still significant at the 0.01 level.

Other statistically significant variables include:

- 1997 household net worth
- The highest category of math SAT scores
- All but one category of the highest.degree variable
- The category of living with mom and a step-parent
- Many of the occupation categories
- The male and married interaction

Most of the interpretations are straightforward. Interpreting categorical variable coefficients follows the pattern established for gender. Numeric variables have a different interpretation. Take as an example 1997 household net worth. Its coefficient means, all else held equal, an increase of $100,000 in the HH.net.worth.1997 variable results in a predicted increase in respondent income of $1936.56.

For the interaction between Male and marriage status, the interpretation is as follows: all else held equal, married men are predicted to earn $6907.76(for being Male) + $17133.35(for being male and married) = $24041.11 more than married women.

```
##
## Call:
## lm(formula = income ~ sex + bio.children.2007 + HH.net.worth.1997 +
##     bio.mom.degree + height + math.SAT + highest.degree + parents +
##     marriage.status + occ + sex:marriage.status, data = nlsy)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -141153  -18293   -3170   12739  218934
##
## Coefficients:
##                                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)                        2.20e+03   1.18e+04    0.19  0.85169
## sexMale                            6.91e+03   2.10e+03    3.29  0.00102 **
## bio.children.2007                 -4.93e+02   6.57e+02   -0.75  0.45336
## HH.net.worth.1997                  1.94e-02   4.55e-03    4.26  2.1e-05 ***
```

```
## bio.mom.degree                          1.27e+02   2.11e+02    0.60  0.54705
## height                                  6.65e+02   2.09e+03    0.32  0.75074
## math.SATValid Skip                     -7.37e+02   1.84e+03   -0.40  0.68884
## math.SAT200 - 300                      -1.68e+03   6.37e+03   -0.26  0.79144
## math.SAT301 - 400                      -7.89e+02   3.59e+03   -0.22  0.82605
## math.SAT401 - 500                       1.48e+03   2.58e+03    0.57  0.56608
## math.SAT501 - 600                       5.06e+03   2.51e+03    2.01  0.04437 *
## math.SAT601 - 700                       5.40e+03   2.95e+03    1.83  0.06725 .
## math.SAT701 - 800                       1.19e+04   3.92e+03    3.05  0.00233 **
## highest.degreeGED                       3.39e+03   2.58e+03    1.32  0.18818
## highest.degreeHigh School               7.30e+03   2.25e+03    3.25  0.00118 **
## highest.degreeAssociate's               1.46e+04   2.92e+03    5.00  6.0e-07 ***
## highest.degreeBachelor's                2.64e+04   2.72e+03    9.72  < 2e-16 ***
## highest.degreeMaster's                  3.28e+04   3.53e+03    9.29  < 2e-16 ***
## highest.degreePhD                       4.06e+04   1.15e+04    3.52  0.00044 ***
## highest.degreeProfessional              9.79e+04   6.39e+03   15.32  < 2e-16 ***
## parentsTwo parents, bio. mother        -4.72e+03   1.80e+03   -2.62  0.00876 **
## parentsTwo parents, bio. father        -7.04e+03   4.27e+03   -1.65  0.09948 .
## parentsBiological mother only          -1.36e+03   1.35e+03   -1.01  0.31392
## parentsBiological father only          -5.82e+03   3.49e+03   -1.67  0.09515 .
## parentsAdoptive parent(s)              -8.32e+03   1.01e+04   -0.83  0.40927
## parentsFoster parent(s)                -1.11e+04   9.67e+03   -1.15  0.25099
## parentsNo parents, grandparents        -2.96e+03   4.32e+03   -0.69  0.49245
## parentsNo parents, other relatives     -9.96e+03   7.52e+03   -1.33  0.18523
## parentsAnything else                   -4.96e+03   8.66e+03   -0.57  0.56708
## marriage.statusMarried                  1.53e+03   1.70e+03    0.90  0.36621
## marriage.statusSeparated                9.90e+03   5.09e+03    1.95  0.05153 .
## marriage.statusDivorced                 3.89e+03   2.56e+03    1.52  0.12851
## marriage.statusWidowed                 -1.05e+03   9.70e+03   -0.11  0.91340
## occCONSTRUCTION TRADES, EXTRACTION      1.46e+04   4.00e+03    3.65  0.00027 ***
## occCOUNSELORS, SOCIAL, RELIGIOUS        3.99e+03   4.94e+03    0.81  0.41856
## occEDUCATION, TRAINING, LIBRARY        -3.63e+03   6.51e+03   -0.56  0.57776
## occENTERTAINERS, PERFORMERS, SPORTS     1.09e+04   5.87e+03    1.85  0.06394 .
## occEXECS, ADMINS, MANAGERS              3.02e+04   3.69e+03    8.19  3.6e-16 ***
## occFOOD PREPARATIONS, SERVING           2.09e+03   3.92e+03    0.53  0.59319
## occHEALTH CARE TECHNICAL, SUPPORT       9.05e+03   4.04e+03    2.24  0.02524 *
## occHEALTH DIAGNOSIS AND TREATING        3.36e+04   4.38e+03    7.68  2.1e-14 ***
## occLAWYERS, JUDGES, LEGAL SUPPORT       2.62e+04   6.92e+03    3.79  0.00015 ***
## occMANAGEMENT RELATED                   2.34e+04   4.09e+03    5.72  1.2e-08 ***
## occMATH, COMPUTER SCIENTISTS            2.82e+04   4.64e+03    6.07  1.4e-09 ***
## occMEDIA, COMMUNICATION                 3.98e+03   6.09e+03    0.65  0.51349
## occOFFICE, ADMINISTRATIVE SUPPORT       9.55e+03   3.51e+03    2.72  0.00655 **
## occOTHER                                1.58e+04   3.85e+03    4.11  4.1e-05 ***
## occPERSONAL CARE, SERVICE              -2.47e+03   4.34e+03   -0.57  0.56886
## occPRODUCTION, OPERATING                1.81e+04   5.46e+03    3.30  0.00096 ***
## occPROTECTIVE SERVICE                   2.20e+04   4.54e+03    4.84  1.3e-06 ***
## occSALES, RELATED                       1.82e+04   3.67e+03    4.96  7.3e-07 ***
## occSETTER, OPERATORS, TENDERS           7.78e+03   4.18e+03    1.86  0.06288 .
## occSOCIAL SCIENTISTS, RELATED           1.72e+04   8.73e+03    1.97  0.04918 *
## occTEACHERS                             2.70e+03   4.20e+03    0.64  0.52019
## occTRANSPORTATION, MATERIAL MOVING      8.79e+03   3.78e+03    2.32  0.02024 *
## sexMale:marriage.statusMarried          1.71e+04   2.35e+03    7.30  3.6e-13 ***
## sexMale:marriage.statusSeparated       -7.17e+03   7.38e+03   -0.97  0.33160
## sexMale:marriage.statusDivorced         4.82e+03   3.85e+03    1.25  0.21041
```

```
## sexMale:marriage.statusWidowed       2.80e+04   2.57e+04    1.09  0.27446
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 33100 on 3736 degrees of freedom
##   (2839 observations deleted due to missingness)
## Multiple R-squared:  0.367,  Adjusted R-squared:  0.357
## F-statistic: 37.3 on 58 and 3736 DF,  p-value: <2e-16
```

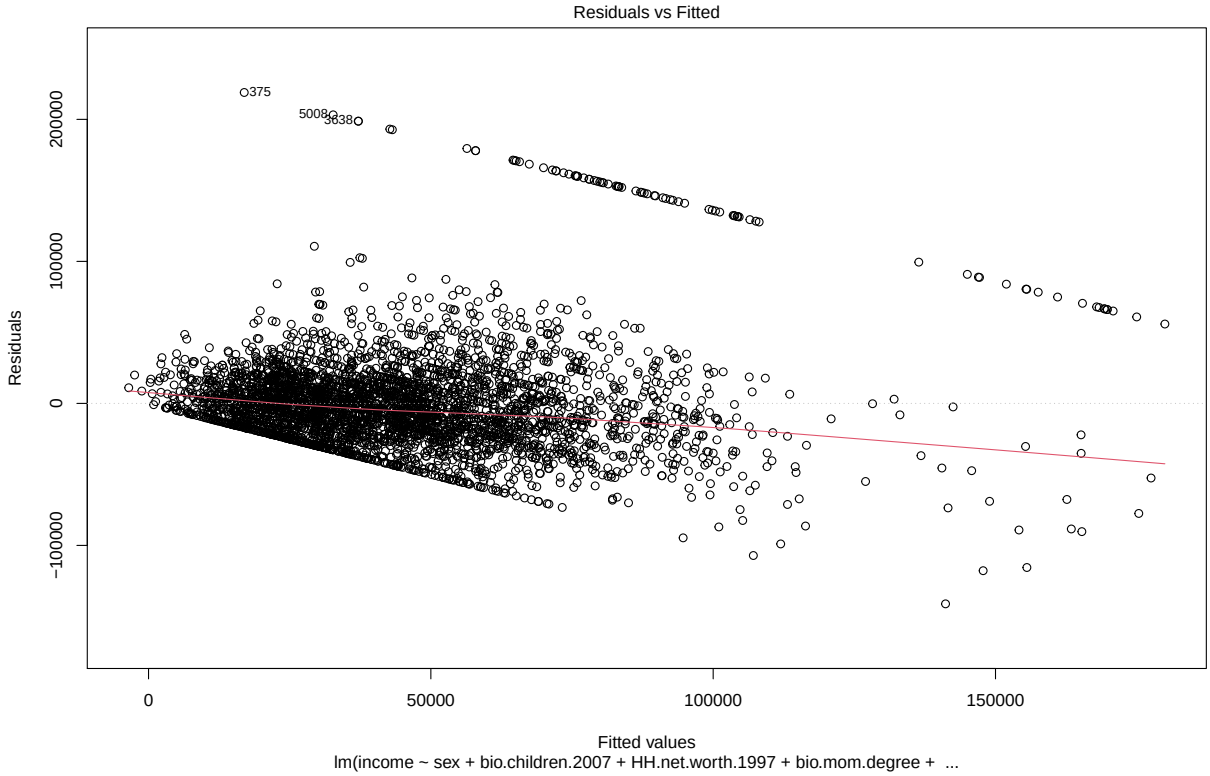
The following plots show potential issues with the linear regression.

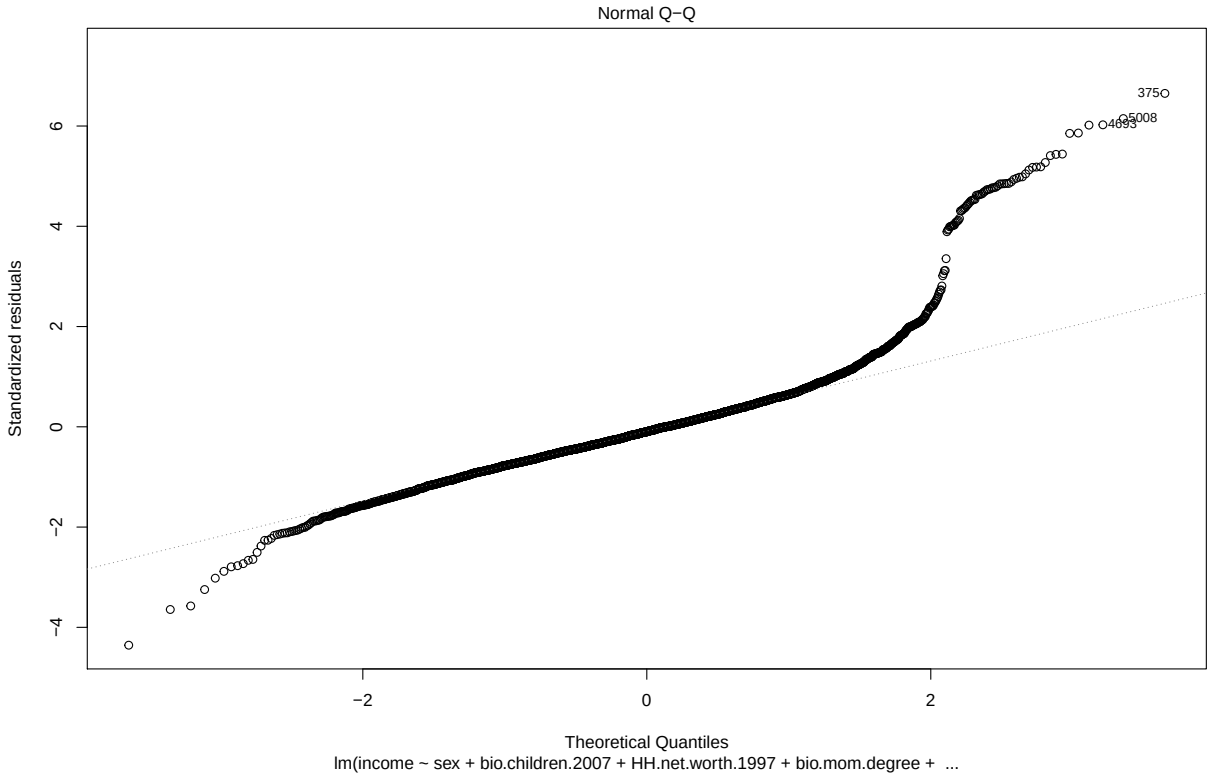
The first plot shows close to linear relationship between fitted values and residuals. I am not concerned by its output.

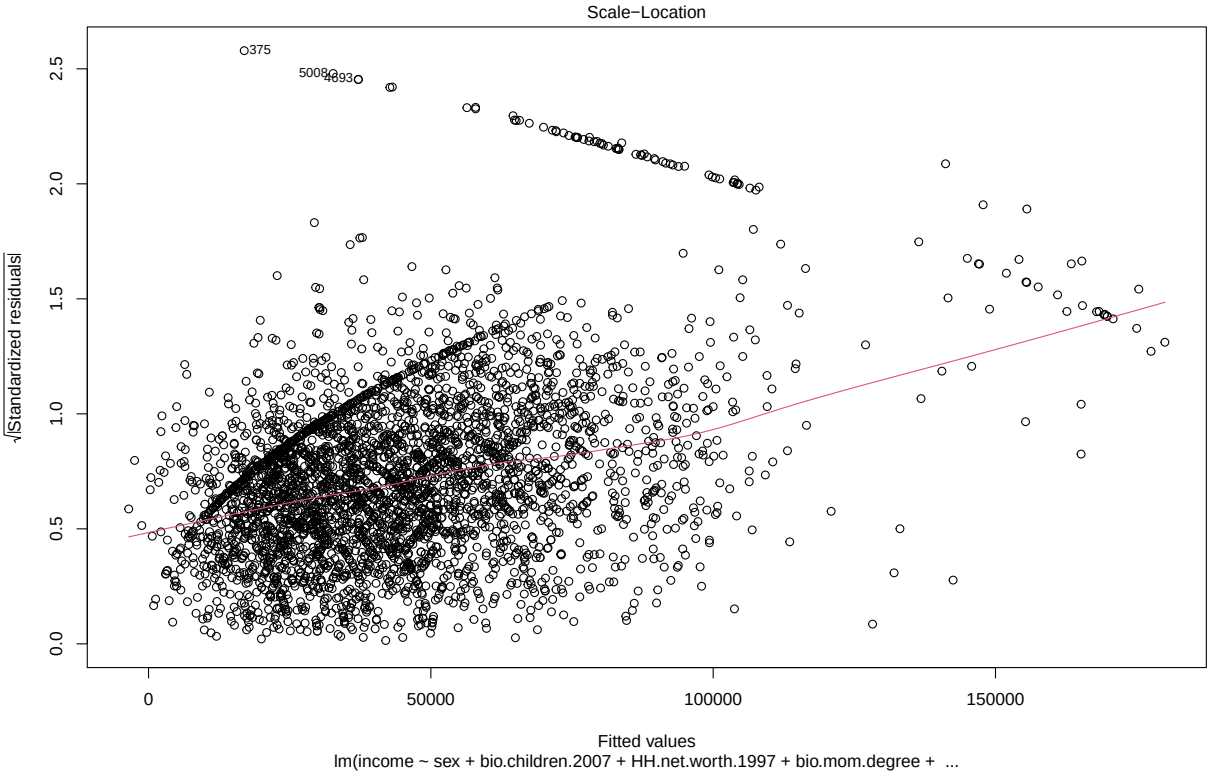
The second is a Q-Q plot which shows that many of the residuals follow a normal distribution, but larger residuals are much larger than would be expected from a normal distribution. This is likely due to the extremely high wages of the top 2% of earners. This suggests removing those earners may improve the reliability of the model.

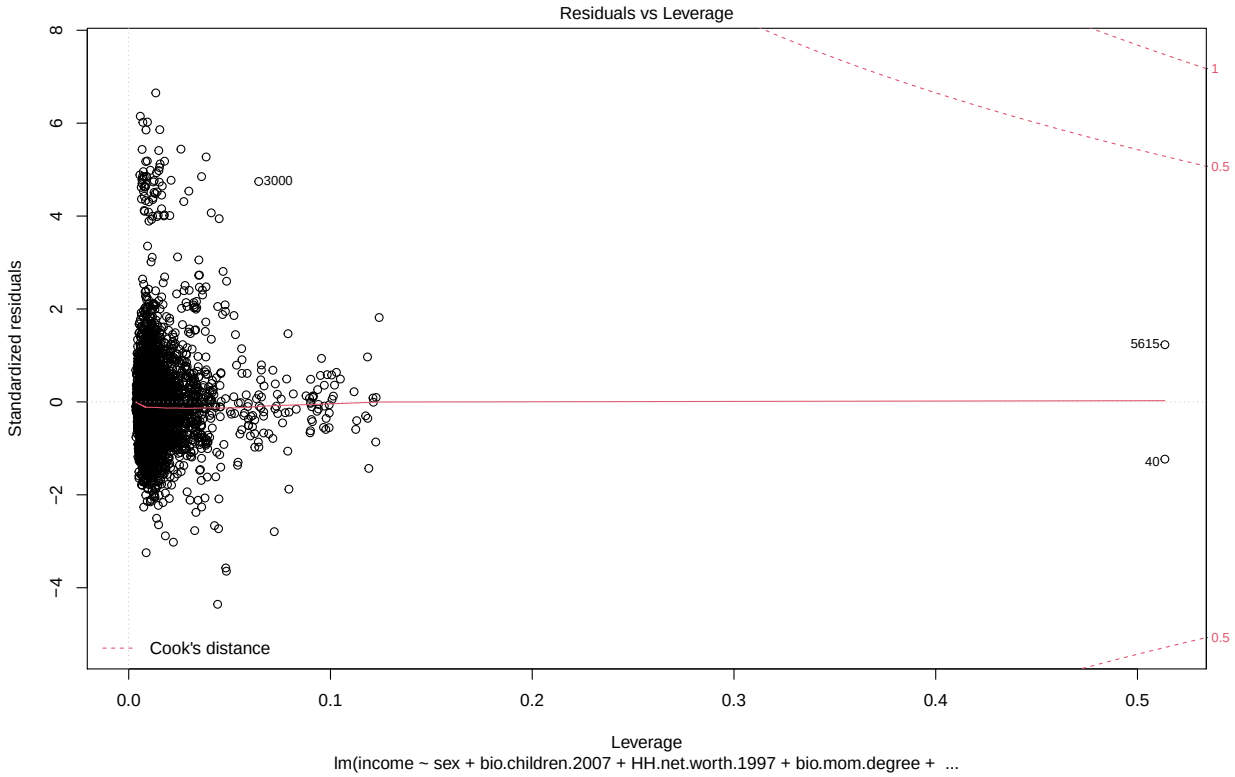
The third, Scale-Location plot shows a positive relationship which is worrisome. Perhaps removing high earners may correct this as well.

The final plot shows no outliers have undue influence on the model.

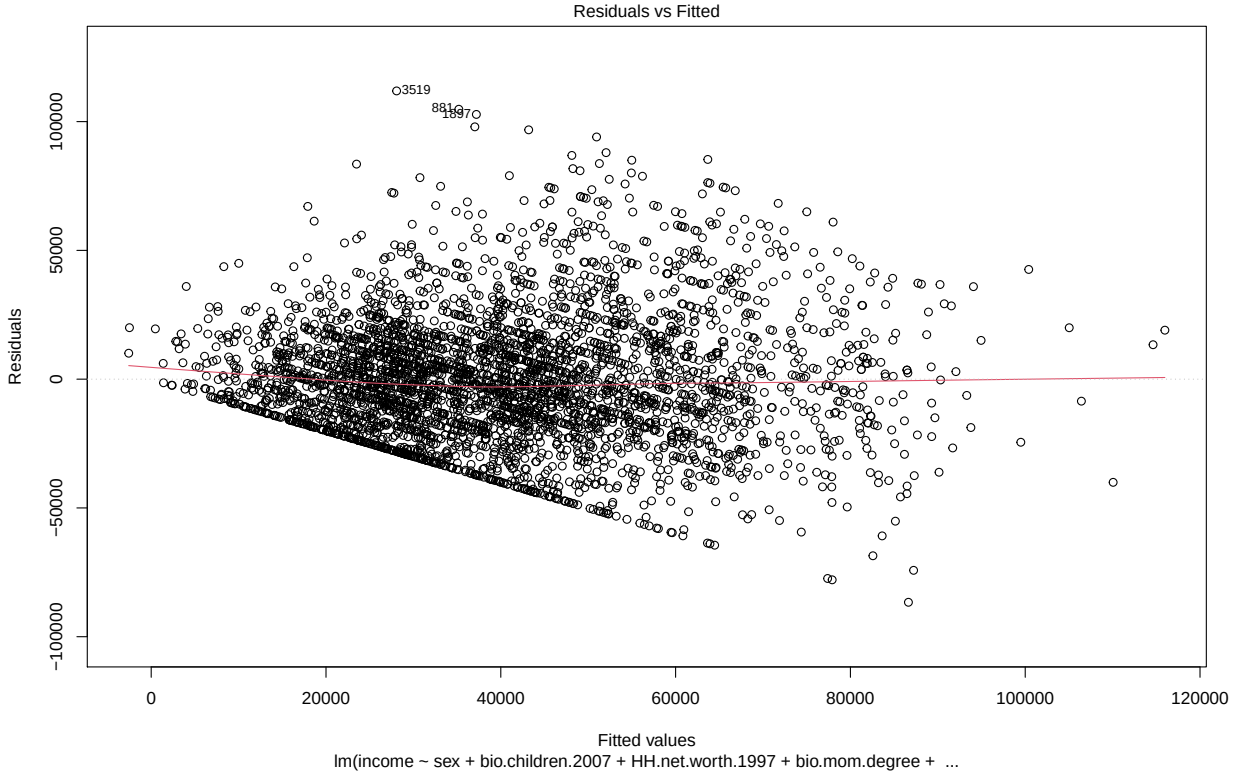


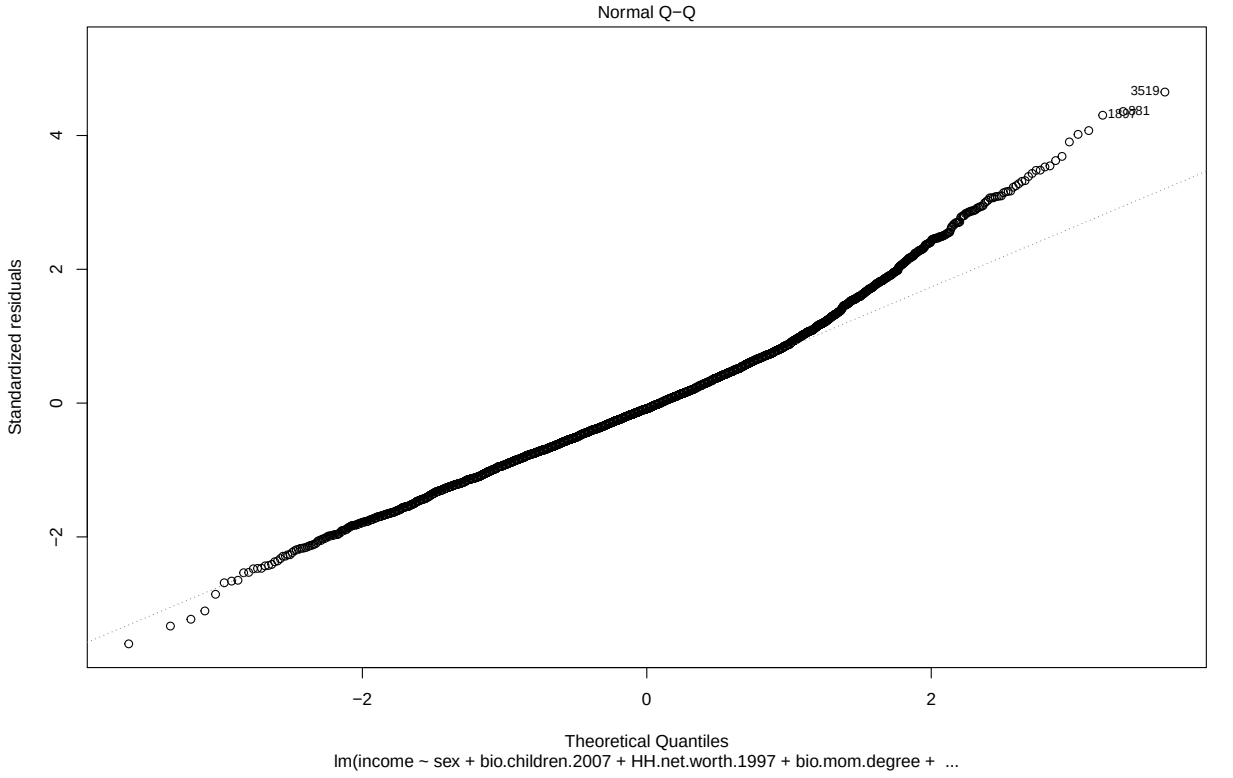

Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(income ~ sex + bio.children.2007 + HH.net.worth.1997 + bio.mom.degree + ...
375
5008
4693

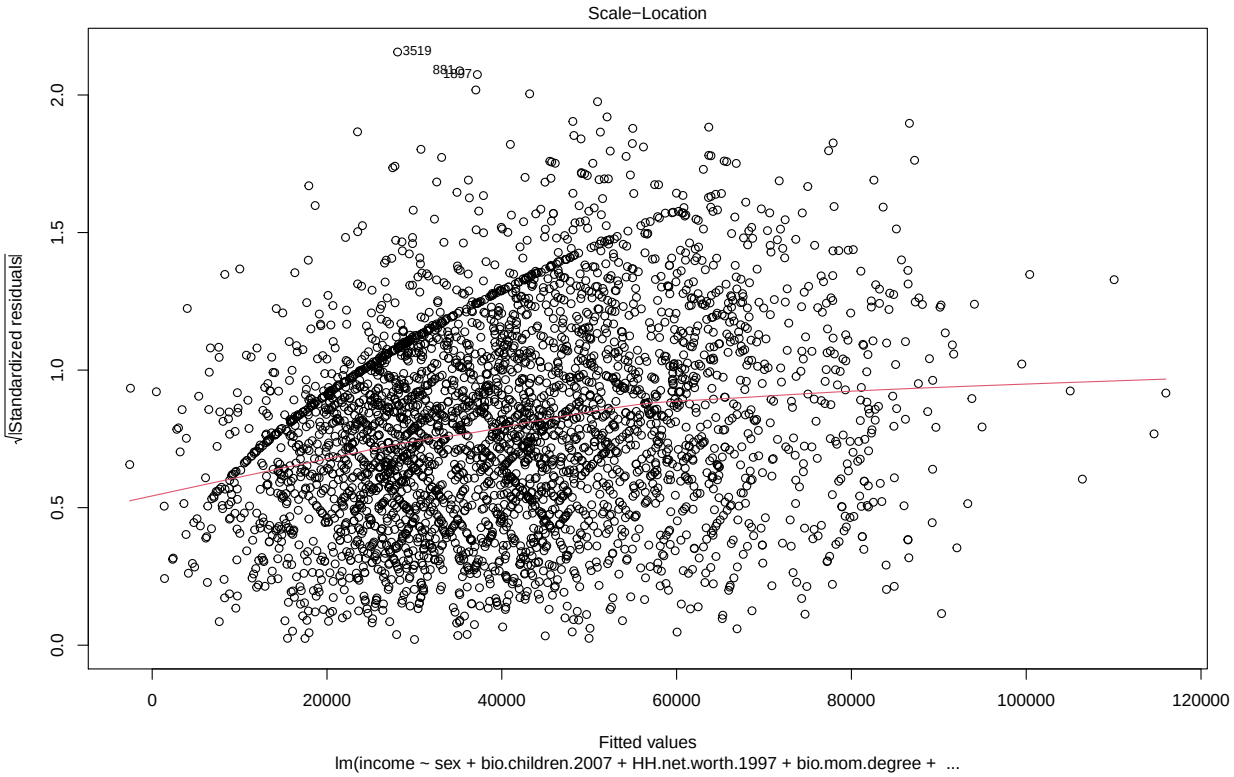

Scale−Location
√|Standardized residuals|
Fitted values
lm(income ~ sex + bio.children.2007 + HH.net.worth.1997 + bio.mom.degree + ...
375
5008
4693

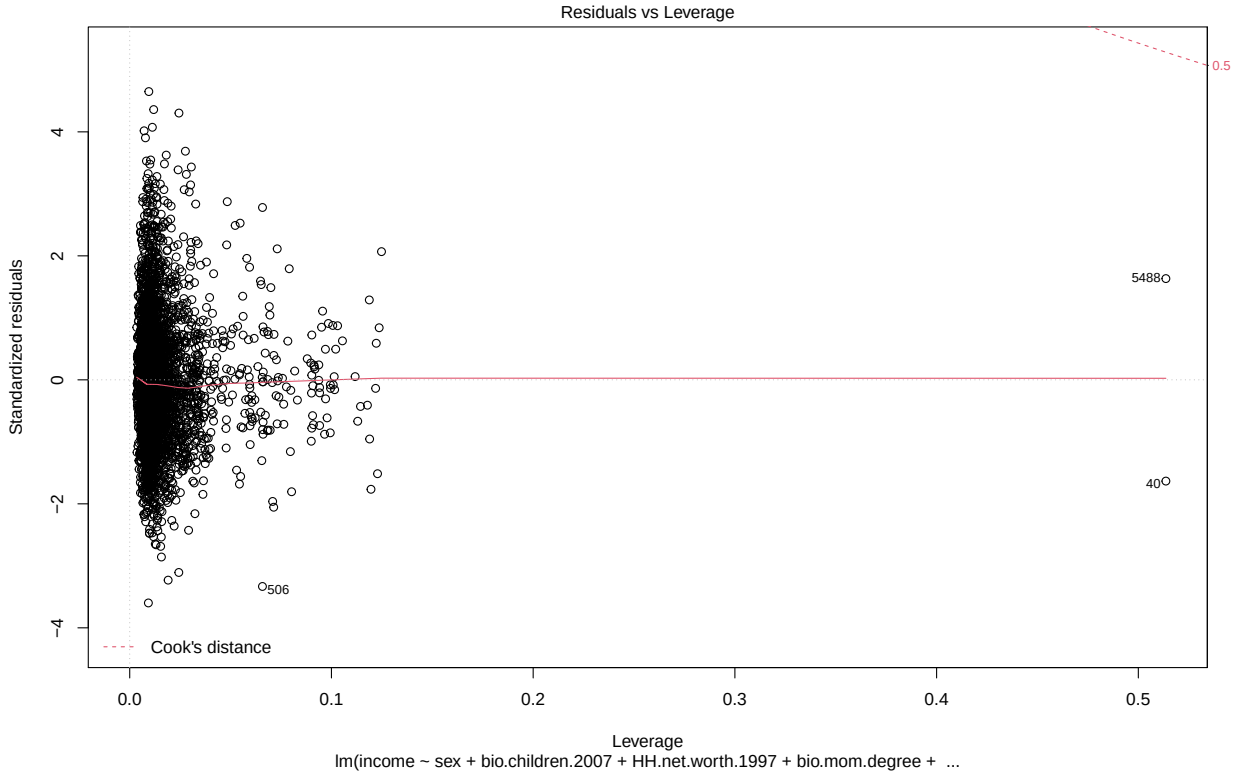


As can be seen in the following plots, removing the top 2% of earners improves the normality of residuals and the linear relationship between residuals and fitted values. After this model change, the highest category of Math-SAT scorers is no longer statistically significant, but all other previously significant variables remain significant. Notably, the estimate for the gender income gap reduces from \$6907.76 to \$3392.73. But, 74.07% of the top 2% of earners are male, so removing those earners will necessarily bias downward the gender income gap. So, despite issues presented by previous plots, I conclude using the results of the model that includes the top 2% of earners.


Residuals vs Fitted
3519
881
1897
Residuals
100000
50000
0
−50000
−100000
0
20000
40000
60000
80000
100000
120000
Fitted values
lm(income ~ sex + bio.children.2007 + HH.net.worth.1997 + bio.mom.degree + ...


Normal Q−Q
3519
881
1897
Standardized residuals
4
2
0
−2
−2
0
2
Theoretical Quantiles
lm(income ~ sex + bio.children.2007 + HH.net.worth.1997 + bio.mom.degree + ...


Scale−Location
√|Standardized residuals|
3519
8810
1997
0.0
0.5
1.0
1.5
2.0
0
20000
40000
60000
80000
100000
120000
Fitted values
lm(income ~ sex + bio.children.2007 + HH.net.worth.1997 + bio.mom.degree + ...


Residuals vs Leverage
Standardized residuals
0.5
5488
40
506
Cook's distance
−4
−2
0
2
4
0.0
0.1
0.2
0.3
0.4
0.5
Leverage
lm(income ~ sex + bio.children.2007 + HH.net.worth.1997 + bio.mom.degree + ...

Discussion

After controlling for multiple variables, the effect of gender on income declined from $16945.02 to $6907.76. This suggests that there was omitted variable bias. The persistence of the positive effect of gender is evidence of gender discrimination in the workplace. I attempted to explain the variation by using height, occupation choice, agreeableness, and desire to stay home with children. Occupation choice had a significant impact in the model and height did not. Married men were estimated to make more than married women which may be evidence for women being more likely to stay home with children. The question I used as a proxy for agreeableness did not have a statistically significant impact and I eventually omitted it from the analysis because it had a large number of missing values.

For future analysis I recommend finding a better proxy for the psychological dimension of agreeableness. An in-depth analysis of the big 5 psychological dimensions may be beneficial as well.

I included the incomes of those who did not work in my analysis. My model in part is attempting to predict who chooses not to work (income of 0), and in part predicting each individual's income. It may be more reasonable to fit two models to answer the two questions. A better analysis would use number of years worked as a predictor and then compare only the incomes of working individuals. That variable, however, was not included in the data set.

I'm convinced there is more to the story behind the gap in average income for men and women because of my analysis, but I do not believe that the final effect I found is the accurate estimate of the income loss women experience due to discrimination. Variables that better measure work experience and psychological differences may be needed to find a more accurate estimate.