Gender Wage Inequality in STEM

Lydia Gibson, Sara Hatter & Ken Vu

April 28, 2022

Introduction

Do we choose our career path based on gender-based social roles or based on top salary? Although many countries, such as China, have incorporated women into their labor power to become a powerful economy[1], women still choose careers that are more in sync to gender stereotype.

Undoubtedly, personality characteristics associated with women, are sympathy, kindness, warmth, and reflect a concern about other people. However, the traits associated to men are achievement orientation and ambitiousness, and concern about accomplishing tasks. These characteristics are very noticeable in the stereotypical association of men in the worker role and women in the family role[2].

More schools are encouraging girls to enter STEM programs and provided them with many resources to succeed in these types of careers. Despite these efforts, women tend to choose career where the median pay is lower.

Data Description

The data was obtained from the American Community Survey 2010-2012 Public Use Microdata Series and has been already subsetted to only concern STEM majors (particularly with an interest in women majoring in STEM). For each row in the data set (which represents one major), there's a collection of details and statistics about the major, such as the type of major (i.e. Engineering, Health Science, etc), the proportion of women in the sample of individuals working in that particular field, and other relevant pieces of information.

Data set

- Link to data set: https://github.com/fivethirtyeight/data/blob/master/college-majors/women-stem.csv

The dimensions of the data set are 76 rows (`Major`) by 9 columns.

Variables

- `Median`: Median earnings of full-time, year-round workers
- `Rank`: Rank by median earnings
- `Major_code`: Major code, FO1DP in ACS PUMS
- `Major`: Major description
- `Major_category`: Category of major from Carnevale et al
- `Total`: Total number of people with major
- `Men`: Male graduates
- `Women`:Female graduates
- `ShareWomen`: Women as share of total

Research Question and Goals

Our research question tries to find associations within STEM college majors that influence median wages. Our goals are to explore the data for STEM college majors and to create a predictive model for median wages.

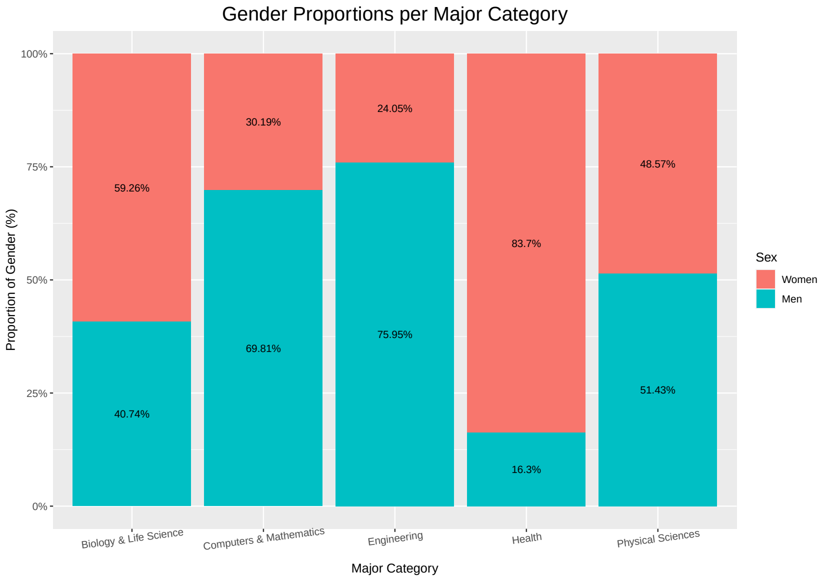
Research Question:

What associations exist within STEM college majors that have an effect on median wages?

Goals:

- To explore the data for STEM college majors.
- To create a predictive model for median wage.

Stacked Bar chart: Gender Proportions per Major Category



Exploratory Data Analysis

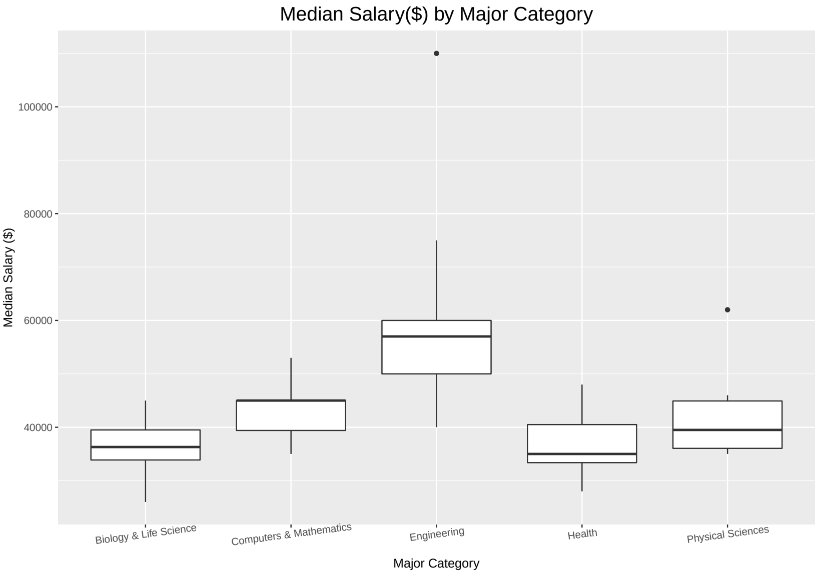
`Median` wage of the individual majors ranged from $\$26,000$ for Zoology to $\$110,000$ for Petroleum Engineering ($Mdn = \$44350, M = \$46118$) .

We have set `Major_category` as a factor with the following levels:

- [1]"Biology & Life Science"
- [2]"Computers & Mathematics"
- [3]"Engineering"
- [4]"Health"
- [5]"Physical Sciences"

so that we can further distinguish the variation of share of women within major categories and the median wages each major category earns.

Box Plot: Median Wage by Major Category



Test differences between major categories

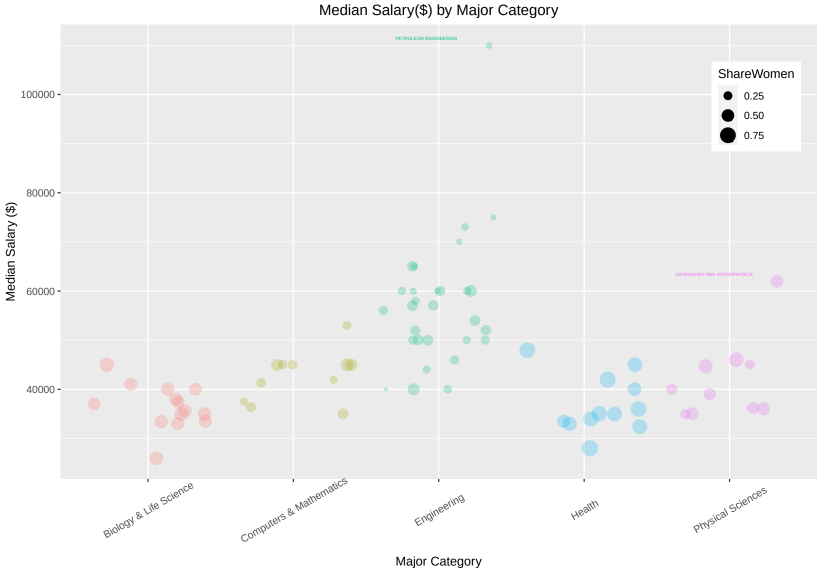
Based on our boxplot, we noticed there may be a significant difference between median wage by major category so we ran an ANOVA to test our hypothesis:

$H_0 : \alpha_1 = \alpha_2 = \alpha_3 = \alpha_4 = \alpha_5 = 0$

$H_A : \alpha_i \neq 0, i = 1, 2..., 5$

Based on our one-way ANOVA, we reject the null hypothesis and concluded that there are statistically significant differences in Median Wages between Major Categories ($F(4, 71) = [16.71], p = [0.0000001013]$).

Jitter plot: Median Wage by Major Category

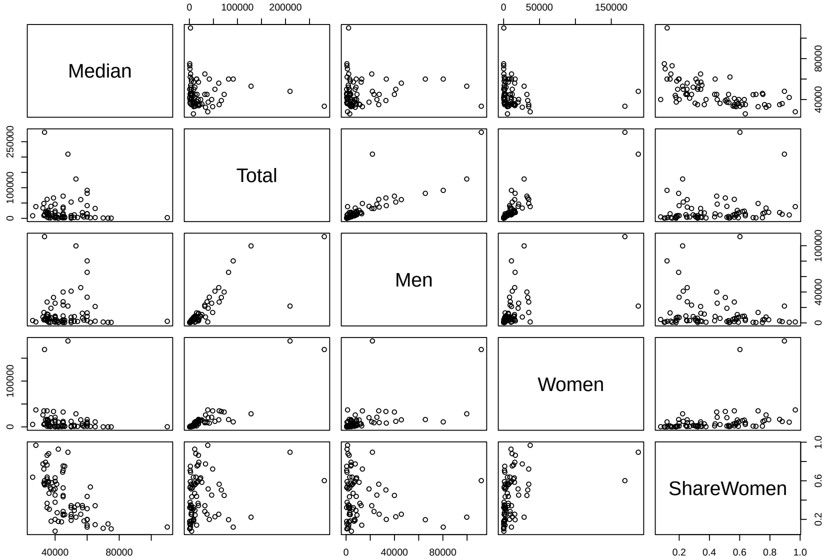


Further Cleaning

- For our analysis, we also removed the columns `Major_code` and `Rank` as they aren't relevant predictors for our purposes.

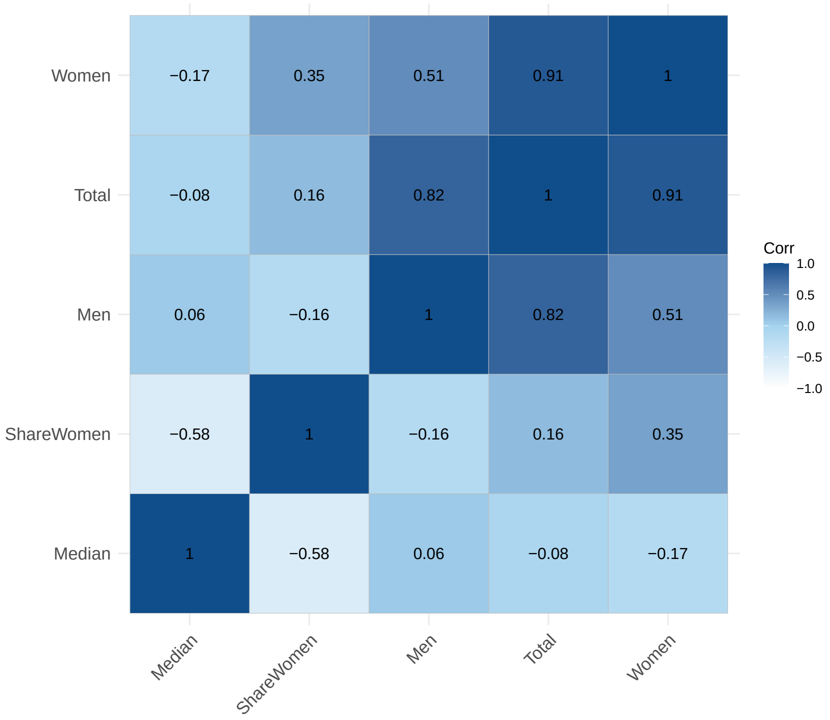
```
##   Major_category Total   Men Women ShareWomen Median
## 1    Engineering  2339  2057   282  0.1205643 110000
## 2    Engineering   756   679    77  0.1018519  75000
## 3    Engineering   856   725   131  0.1530374  73000
## 4    Engineering  1258  1123   135  0.1073132  70000
## 5    Engineering  2573  2200   373  0.1449670  65000
## 6    Engineering 32260 21239 11021  0.3416305  65000
```

Scaterplot Matrix



Scatterplot Matrix Insights

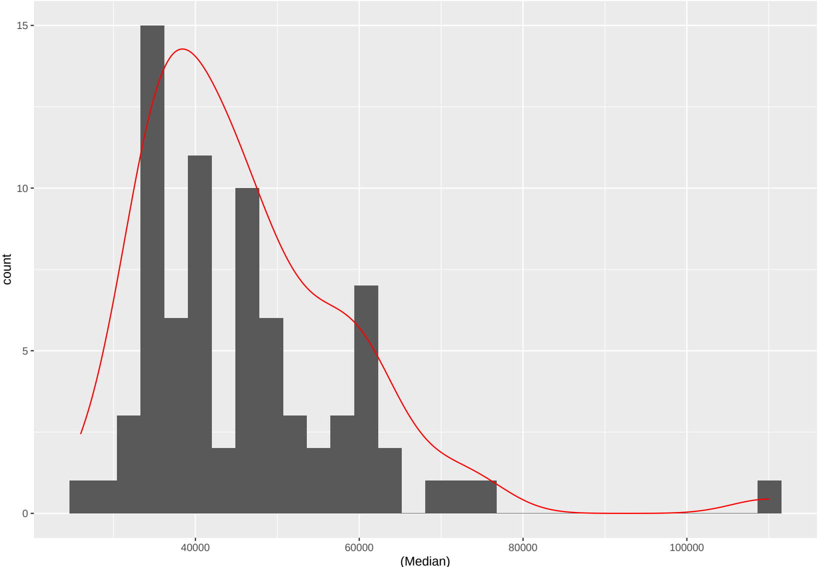
- As expected, there seems to be a negative association between `ShareWomen` and `Median`. This is one of the main motivators for our research.
- There may be an issues of multicollinearity between `Total`, `Men`, `Women` and `ShareWomen`, so we will run some analyses to assess which of these predictors could be removed from our model. To address this, we will run a correlation matrix.



Women
Total
Men
ShareWomen
Median
−0.17
0.35
0.51
0.91
1
−0.08
0.16
0.82
1
0.91
0.06
−0.16
1
0.82
0.51
−0.58
1
−0.16
0.16
0.35
1
−0.58
0.06
−0.08
−0.17
Median
ShareWomen
Men
Total
Women
Corr
1.0
0.5
0.0
−0.5
−1.0

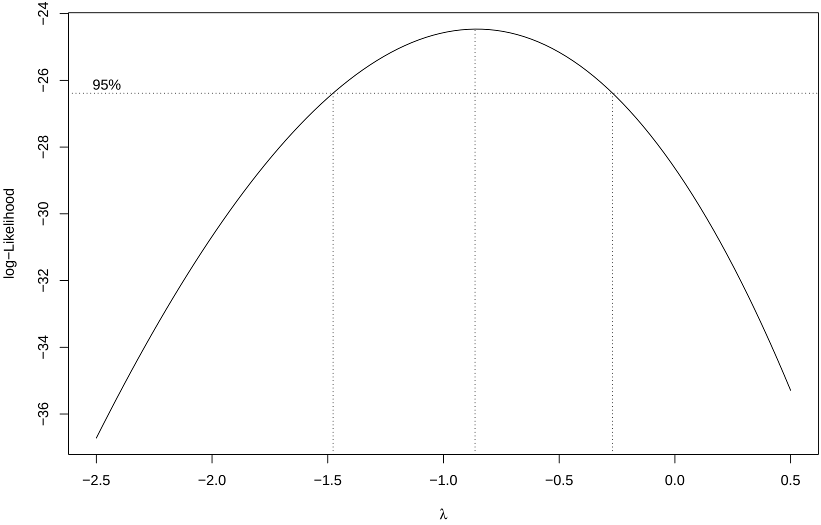
Methods and Results: Checking Assumptions

Before beginning our analysis, we began by exploring the normality within our response variable, `Median`.



Box Cox

We notices that there was some skewing, so we decided to do a Box-Cox test to see if a transformation is necessary.



Box-Cox Summary output

```
## bcPower Transformation to Normality
##    Est Power Rounded Pwr Wald Lwr Bnd Wald Upr Bnd
## Y1   -0.8569          -1      -1.4598       -0.254
##
## Likelihood ratio test that transformation parameter is e
##  (log transformation)
##                             LRT df      pval
## LR test, lambda = (0) 8.338064  1 0.0038823
##
## Likelihood ratio test that no transformation is needed
##                             LRT df             pval
## LR test, lambda = (1) 41.68169  1 0.00000000010741
```

Our rounded power is -1 so we will do an inverse transformation of the response `Median`. However, model interpretability may be difficult.

Building Predicitive Model

We started with the full additive model but it removed to many variables so we decided switched to a model with interactions.

```
Step:  AIC=-1896.41
(Median^(-1)) ~ Major_category

                 Df  Sum of Sq        RSS     AIC
<none>                         9.7008e-10 -1896.4
- Major_category  4 1.3021e-09 2.2722e-09 -1839.7

Call:
lm(formula = (Median^(-1)) ~ Major_category, data = dat2[-c(2)])

Residuals:
        Min          1Q      Median          3Q         Max 
-0.000009108 -0.000001730  0.000000071  0.000001982  0.000010570 

Coefficients:
                                         Estimate    Std. Error t value Pr(>|t|)    
(Intercept)                          0.0000278915  0.0000009879  28.233  < 2e-16 ***
Major_categoryComputers & Mathematics -0.0000041904  0.0000014893  -2.814  0.00633 ** 
Major_categoryEngineering            -0.0000096922  0.0000012029  -8.057 1.31e-11 ***
Major_categoryHealth                 -0.0000001474  0.0000014541  -0.101  0.91955    
Major_categoryPhysical Sciences      -0.0000033268  0.0000015304  -2.174  0.03306 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.000003696 on 71 degrees of freedom
Multiple R-squared:  0.5731,	Adjusted R-squared:  0.549 
F-statistic: 23.83 on 4 and 71 DF,  p-value: 1.611e-12
```

Building Predicitive Model w/ Interaction

Since the additive model removed all but one predictor, we reran the model with interactions

Running step-wise to reduce the model's AIC

```
Step:  AIC=-1896.41
(Median^(-1)) ~ Major_category

                 Df  Sum of Sq        RSS     AIC
<none>                         9.7008e-10 -1896.4
- Major_category  4 1.3021e-09 2.2722e-09 -1839.7

Call:
lm(formula = (Median^(-1)) ~ Major_category + Men + Women + ShareWomen +
    Men:ShareWomen, data = dat2[-c(2)])

Residuals:
          Min            1Q        Median            3Q           Max
-0.0000090859 -0.0000022392 -0.0000000436  0.0000018485  0.0000107030

Coefficients:
                                        Estimate Std. Error t value Pr(>|t|)
(Intercept)                            2.648e-05  2.667e-06   9.928 8.57e-15 ***
Major_categoryComputers & Mathematics -3.192e-06  1.877e-06  -1.701   0.0937 .
Major_categoryEngineering             -8.453e-06  1.884e-06  -4.488 2.90e-05 ***
Major_categoryHealth                   4.561e-07  1.775e-06   0.257   0.7980
Major_categoryPhysical Sciences       -3.010e-06  1.572e-06  -1.915   0.0598 .
Men                                   -6.676e-11  4.375e-11  -1.526   0.1318
Women                                 -5.069e-11  3.036e-11  -1.669   0.0997 .
ShareWomen                             1.909e-06  4.222e-06   0.452   0.6527
Men:ShareWomen                         2.748e-10  1.488e-10   1.846   0.0693 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.000003694 on 67 degrees of freedom
Multiple R-squared:  0.5976,    Adjusted R-squared:  0.5495
F-statistic: 12.44 on 8 and 67 DF,  p-value: 9.644e-11
```

Test significance of predictor `Women`

```
Analysis of Variance Table

Model 1: (Median^(-1)) ~ Major_category + Men + ShareWomen + Men:ShareWomen
Model 2: (Median^(-1)) ~ Major_category
  Res.Df        RSS Df   Sum of Sq      F Pr(>F)
1     68 9.5244e-10
2     71 9.7008e-10 -3 -1.7636e-11 0.4197 0.7394
```

Given $p = 0.7394 > \alpha = 0.05$, we fail to reject H_0 (`Women` is not a significant predictor). Thus, we can remove the predictor `Women`.

Getting the reduced final model

```
Call:
lm(formula = (Median^(-1)) ~ Major_category + Men + ShareWomen +
    Men:ShareWomen, data = dat2[-c(2)])

Residuals:
          Min            1Q        Median            3Q           Max
-0.0000092133 -0.0000020260  0.0000001303  0.0000021737  0.0000106200

Coefficients:
                                       Estimate Std. Error t value Pr(>|t|)
(Intercept)                           2.710e-05  2.676e-06  10.128 3.24e-15 ***
Major_categoryComputers & Mathematics -3.442e-06  1.895e-06  -1.816   0.0737 .
Major_categoryEngineering            -8.866e-06  1.892e-06  -4.687 1.38e-05 ***
Major_categoryHealth                 -3.988e-07  1.722e-06  -0.232   0.8176
Major_categoryPhysical Sciences      -3.090e-06  1.592e-06  -1.941   0.0564 .
Men                                  -4.140e-11  4.157e-11  -0.996   0.3228
ShareWomen                            1.084e-06  4.248e-06   0.255   0.7993
Men:ShareWomen                        8.965e-11  1.006e-10   0.891   0.3759
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.000003743 on 68 degrees of freedom
Multiple R-squared:  0.5808,    Adjusted R-squared:  0.5377
F-statistic: 13.46 on 7 and 68 DF,  p-value: 9.025e-11
```

$$Y^{-1} = 2.71 \cdot 10^{-5} - 3.441 \cdot 10^{-6} x_1 - 8.87 \cdot 10^{-6} x_2 - 3.991 \cdot 10^{-7} x_3 - 3.09 \cdot 10^{-6} x_4 - 4.14 \cdot 10^{-11} x_5 + 1.08 \cdot 10^{-6} x_6 + 8.97 \cdot 10^{-11} x_5 \cdot x_6$$

Predictive power

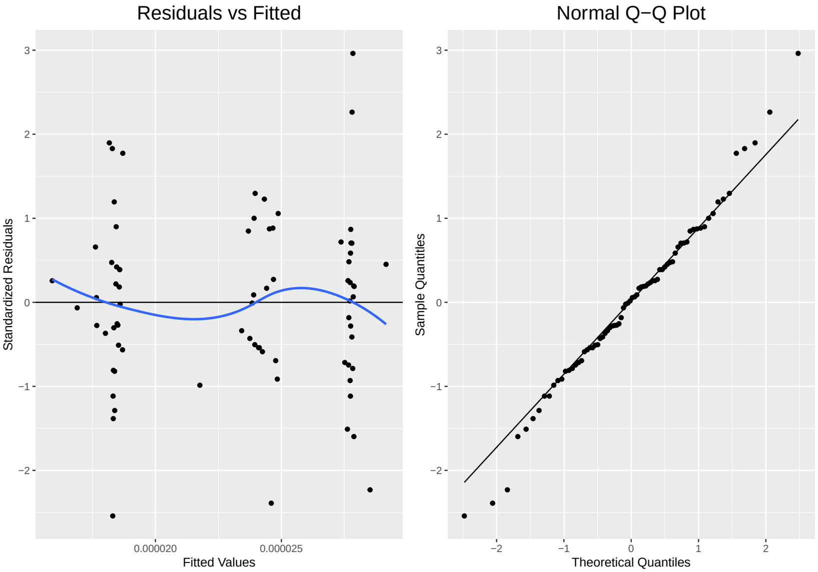
Here we do a prediction interval for `Median`$^{-1}$ for Statistics & Decision Sciences then take the inverse so that our response is in our original units.

```
##          fit      lwr      upr
## 1 41240.31 61595.08 30997.01
```

Looking at the actual `Median` for Statistics & Decision Sciences, we see that the actual response is within our prediction interval of (30997,61595).

Major	Major Category	Men	Share Women	Median
STATISTICS AND DECISION SCIENCE	Computers & Mathematics	2960	0.5265	45000

Model Diagnostics



Model Diagnostics (Numeric Tests)

- Verifying constant variance ($\alpha = 0.05$)

```
##
##  studentized Breusch-Pagan
##  test
##
## data:  lm_reduced
## BP = 3.2776, df = 7,
## p-value = 0.8582
```

- Verifying normality of residuals ($\alpha = 0.05$)

```
##
##  Shapiro-Wilk normality
##  test
##
## data:  rstandard(lm_reduced)
## W = 0.98673, p-value =
## 0.6165
```

Multicollinearity (VIF)

```
## Major_categoryComputers & Mathematics
##                                   2.41
##            Major_categoryEngineering
##                                   4.58
##                 Major_categoryHealth
##                                   2.14
##       Major_categoryPhysical Sciences
##                                   1.57
##                                    Men
##                                   4.20
##                             ShareWomen
##                                   5.21
##                         Men:ShareWomen
##                                   4.19
```

Conclusion

In conclusion

- There is an association with gender and median wage of STEM majors.
- We can predict the median wage of STEM majors based on the major category, total number of men in the major and total proportion of women in the major.
- We all should have majored in Petroleum Engineering!

Further Research

- If we had sex disaggregated data for median wage, we could see the difference in median wage by gender for each major.
- If we had time series data, we could then see how median wage changes with an influx of women and/or exodus of men from a given major.
- Since we only looked at STEM majors, it would be interesting to see if these same variables (`Major_category`, `Men`, `ShareWomen`) are associated with median wage for all majors.

Bibliography

Etaugh, Claire A., and Judith S. Bridges. *Women's Lives: A Psychological Exploration.* 3rd ed., Pearson, 2013.

Kristof, Nicholas D. *Half the Sky: Turning Oppression into Opportunity for Women Worldwide.* Three Rivers Press, 2010.

Code Appendix

For supplementary R script, visit

- https://github.com/lgibson7/Gender-Wage-Inequality-in-STEM