Gender Wage Inequality in STEM

Lydia Gibson, Sara Hatter & Ken Vu

STAT 632, California State University East Bay, Spring 2022

Contents

I. Introduction

Do gender-based social roles or top salary impact our choices of career paths? Although many countries, such as China, have incorporated women into their labor force and developed strong economies as a result, women still tend to choose careers that align more with gender stereotypes.[2] Undeniably, the personality characteristics often associated with women are sympathy, kindness, and warmth, which all reflect a sense of concern towards other people. On the other hand, the traits frequently associated with men are success and ambition, which are concerned more with accomplishing tasks. These characteristics manifest themselves in the stereotypical association of men with the worker role and women with the family role.[1]

In response to this gender bias, more schools are encouraging girls to enter STEM programs in addition to providing them with various resources to help them succeed in these types of careers. However, despite these efforts, women still tend to choose careers where the median pay is lower. Thus, our research question aims to find associations within STEM college majors that influence their median wages. Our goals are to explore the data for STEM college majors and to create a predictive model for median wages.

II. Data Description

a. Data Overview

The data was obtained from the American Community Survey 2010-2012 Public Use Microdata Series and is a subset which only contains STEM majors. The data dimensions are **76 rows** (STEM majors) by **9 columns** (variables). The variables are: `Rank`, `Major_code`, `Major`, `Major_category`, `Total`, `Men`, `Women`, `ShareWomen`, and `Median`. Below are descriptions of each variable.

- `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

b. Exploratory Data Analysis

For the purpose of exploratory data analysis, `Major_category` was set as a factor to explore the variation of the share of women within major categories and the median wages for those major categories. Within the STEM majors, median wage ranges from $\$26,000$ for Zoology to $\$110,000$ for Petroleum Engineering, with median wage's median = \$44350 and its mean = \$46118. Through a stacked barplot of gender proportions per major category, we see that the biggest proportion of women chose fields related to *Health* and the biggest proportion of men chose fields related to *Engineering* (Figure 1). This is congruent with the gender roles and personality characteristics associated with women and men.

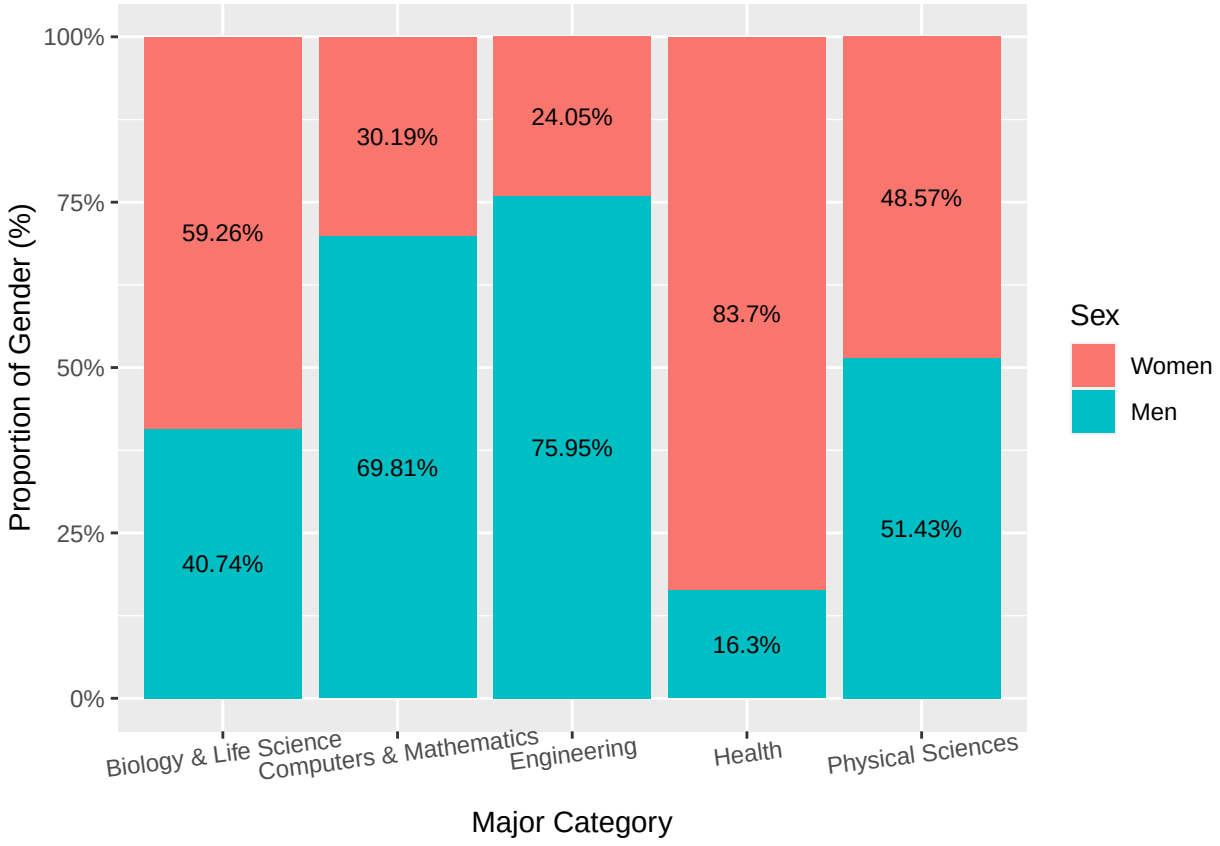

Figure 1: Gender proportions per major category.

Side-by-side boxplots of each major category were then generated to show descriptive statistics, such as the interquartile range, to help identify STEM majors which are outliers with regards to the `Median` variable (see Figure 2). From our jitter plot, we noticed that the *Engineering* major category contains an outlier, *Petroleum Engineering*, and another outlier, *Astronomy and Astrophysics*, can be found in the Physical

Sciences major category. As seen in Figure 2, *Petroleum Engineering* has a smaller proportion of women compared to men (`ShareWomen`= 0.121) with a median wage of $\$110,000$ (`Median`= 110000). *Astronomy and Astrophysics* has a roughly balanced proportion of women compared to men (`ShareWomen`= 0.536) with a median wage of $\$62,000$ (`Median`= 62000). These data visualizations illustrate that there may be a significant difference between median wage by major category, as well as an association between the proportion of women in the major and its median wage.

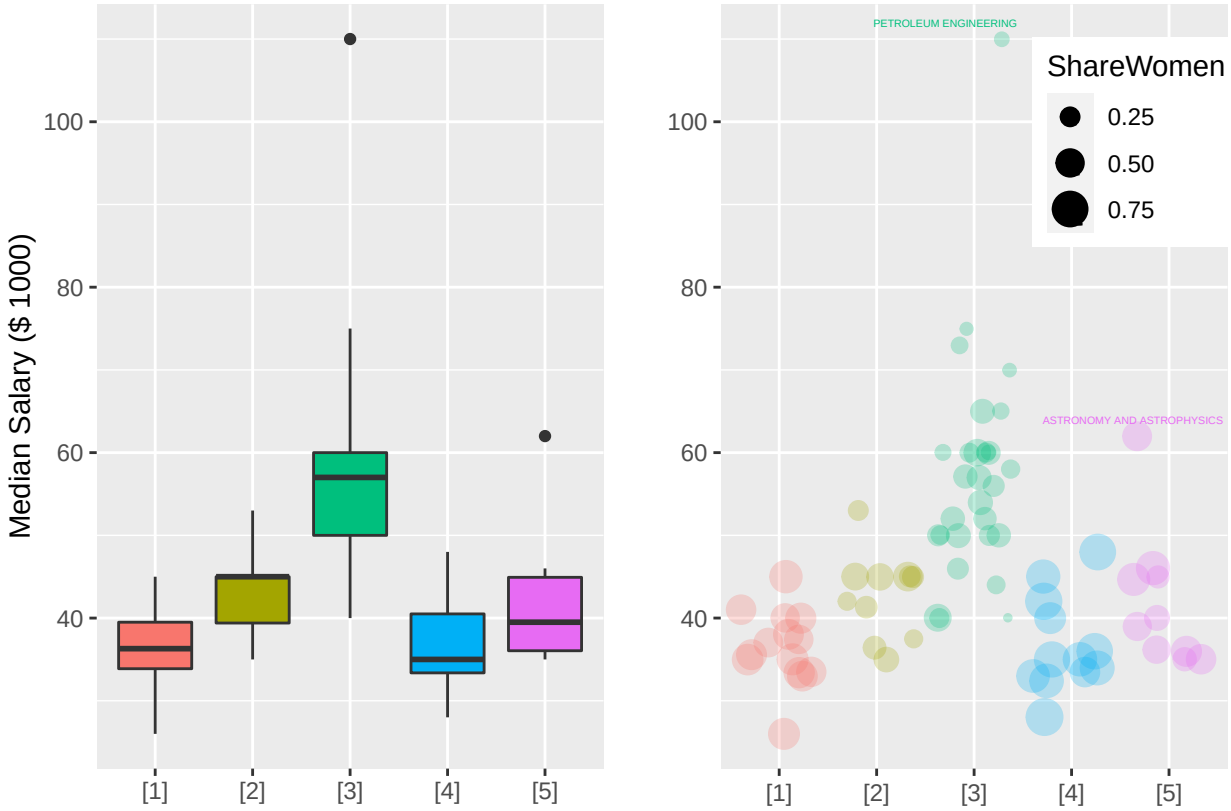

Figure 2: Side-by-side box plot (left) and jitter plot (right) of Median wage ($1000) by major category. *Major categories: [1]Biology & Life Science, [2]Computers & Mathematics, [3]Engineering, [4]Health, [5]Physical Sciences.*

Analysis of Variance (ANOVA) was done to test if there is a statistically significant difference between median wage for any of our five major categories. Based on our ANOVA, we rejected our null hypothesis, given $(F(4,71) = [16.7]; p = 0.00000001013 < \alpha = 0.05$, and concluded that there is a statistically significant difference between median wage per major category. Since they are irrelevant to our analysis, the columns `Major_code` and `Rank` were removed from our dataset. Then, we generated a scatterplot matrix which

revealed that there seems to be a negative association between `ShareWomen` and `Median`. The scatterplot matrix also supports our assumption that there may be an issue of multicollinearity among `Total`, `Men`, `Women`, and `ShareWomen`. This observation makes sense since the column `Total` is the sum of the columns `Men` and `Women`. Likewise, `ShareWomen` refers to the ratio of `Women` to `Total`.

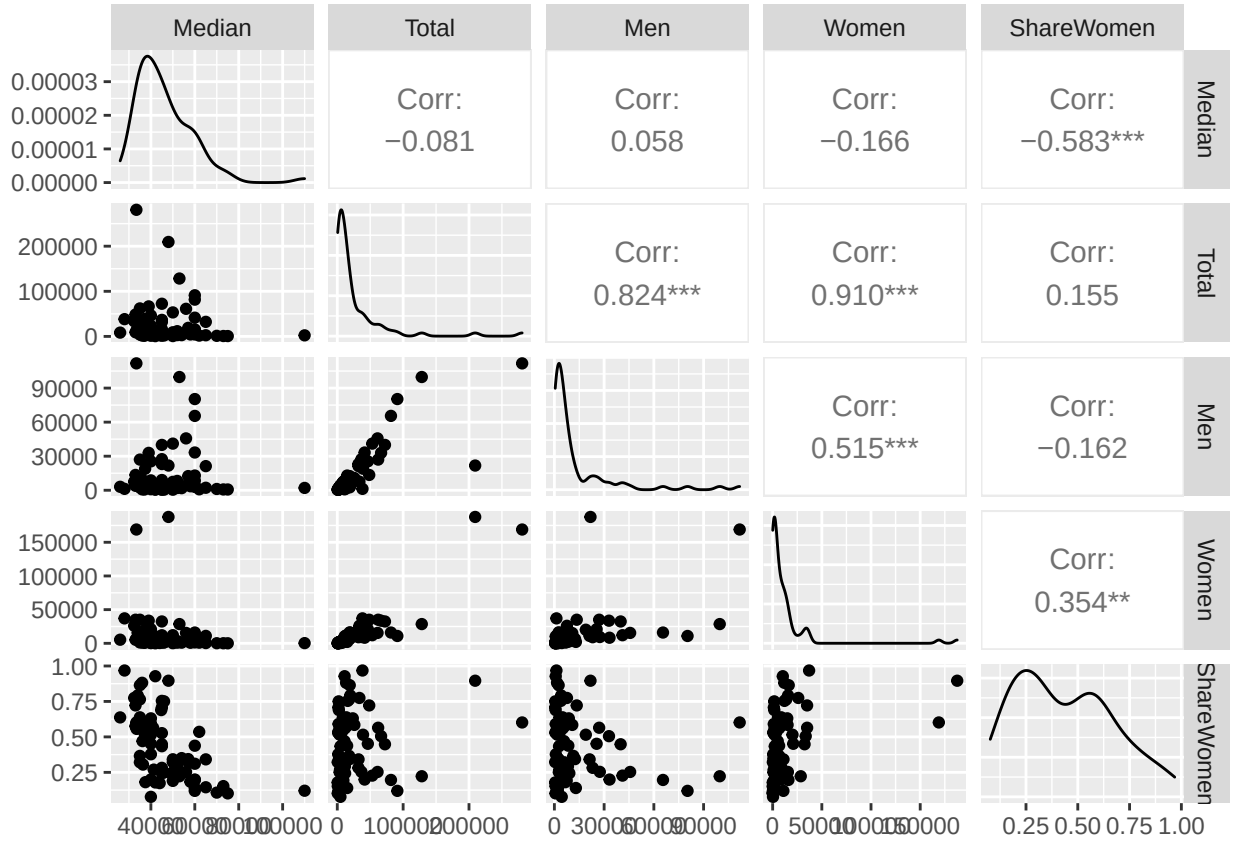

Figure 3: Scatterplot matrix.

c. Box Cox Transformation

We selected the column for median wage, `Median`, as our response variable. While checking normality, linearity, and constant variance, we noticed the data for `Median` shows some right skewing. Accordingly, a Box-Cox test was performed to see if a transformation was necessary (see Figure 4). The resulting rounded power was -1, suggesting that an inverse transformation of the response was required to help with right-skewedness. However, this transformation would later complicate the interpretability of the model.

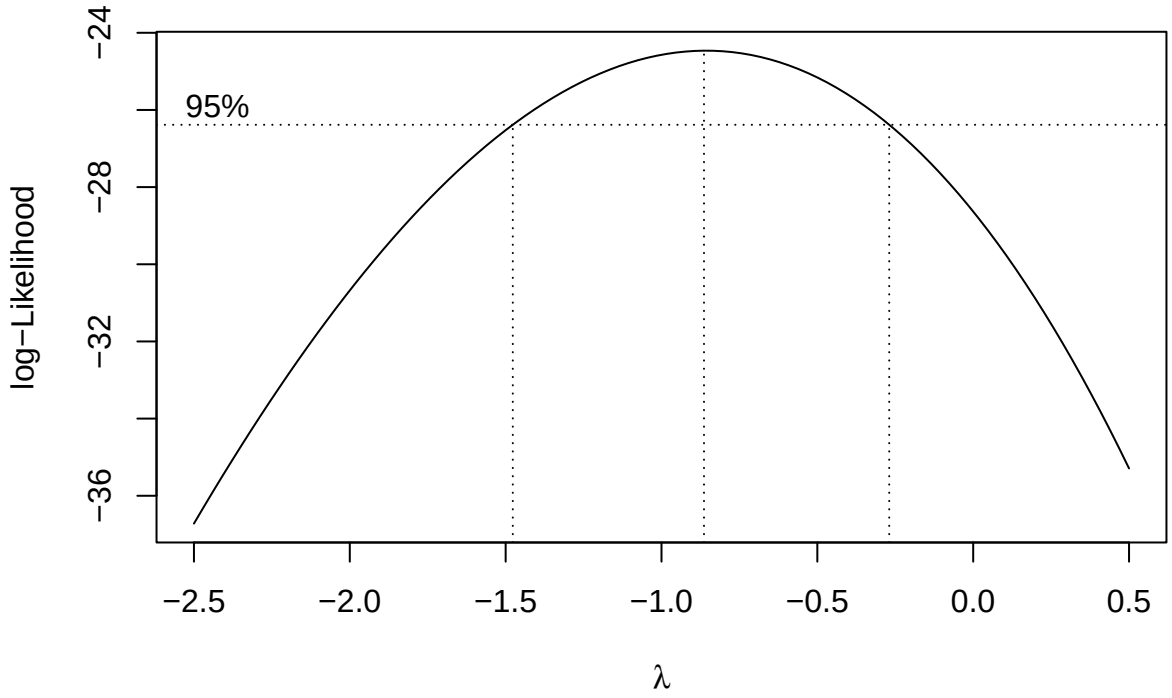

Figure 4: Box-Cox plot for `Median`.

III. Methods and Results

a. Model Fitting

$$[1] Y^{-1} = \beta_0 + \beta_{Major\_category} + \beta_{Total} + \beta_{Men} + \beta_{Women} + \beta_{ShareWomen} + \epsilon$$

Our full addictive model is described by equation [1]. Running this model through the step-wise function using AIC as our criterion, we ended up removing too many predictors; thus, it was decided to check for interactions to see if this new model would help with this issue. Then, another step-wise function was run to reduce the model's AIC. This process resulted in the removal of the predictor `Women` because the $p - value = 0.7394 > \alpha = 0.05$. The final reduced model is described by:

$$\widehat{Median}^{-1} = (2.71 \times 10^{-5}) - (3.44 \times 10^{-6}) \cdot Computers\&Mathematics - (8.87 \times 10^{-6}) \cdot Engineering -$$
$$(3.99 \times 10^{-7}) \cdot Health - (3.09 \times 10^{-6}) \cdot PhysicalSciences - (4.14 \times 10^{-11}) \cdot Men +$$
$$(1.08 \times 10^{-6}) \cdot ShareWomen + (8.98 \times 10^{-11}) \cdot Men : ShareWomen.$$

We achieved an adjusted R^2 score of 0.5377, which means that roughly 53.77% of the variation in the inverse of `Median` can be explained by the model. While the score is not too low, it does indicate that in practical settings, the model still needs improvement. We also noticed that the predictors `Men`, `ShareWomen`, and the interaction term `Men:ShareWomen` are not statistically significant at any significance level (given their p-values).

As noted earlier, model interpretability would be difficult here due to the nature of the transformation. For example, looking at the coefficient for the variable `Major_categoryEngineering`, it can be interpreted to mean that if the major being examined is in the *Engineering* category (and all other predictors would be held constant), the intercept would decrease by roughly 8.866×10^{-6} inverse dollars.

b. Model Diagnostics

To verify the results of the model, a plot of the standardized residuals against the model's fitted values was made in addition to a Q-Q plot of the standardized residuals. In Figure 5, it can be seen on the left that the standardized residuals do not appear to have any discernible relationship with the final model's predicted values. After confirming this interpretation with the studentized Breusch-Pagan test, given $p = 0.8582 > \alpha = 0.05$, it was concluded that the assumption of constant variance for this data set holds up.

As for the Q-Q plot, although some of the data points seem to deviate from the Q-Q line at the tail ends of the data distribution, the standardized residuals do seen to follow the Q-Q line fairly well. To confirm this finding, the Shapiro-Wilks test was used, with resulting $p = 0.6165 > \alpha = 0.05$. Therefore, it can be concluded that the standardized residuals follow a normal distribution, confirming our normality assumption.

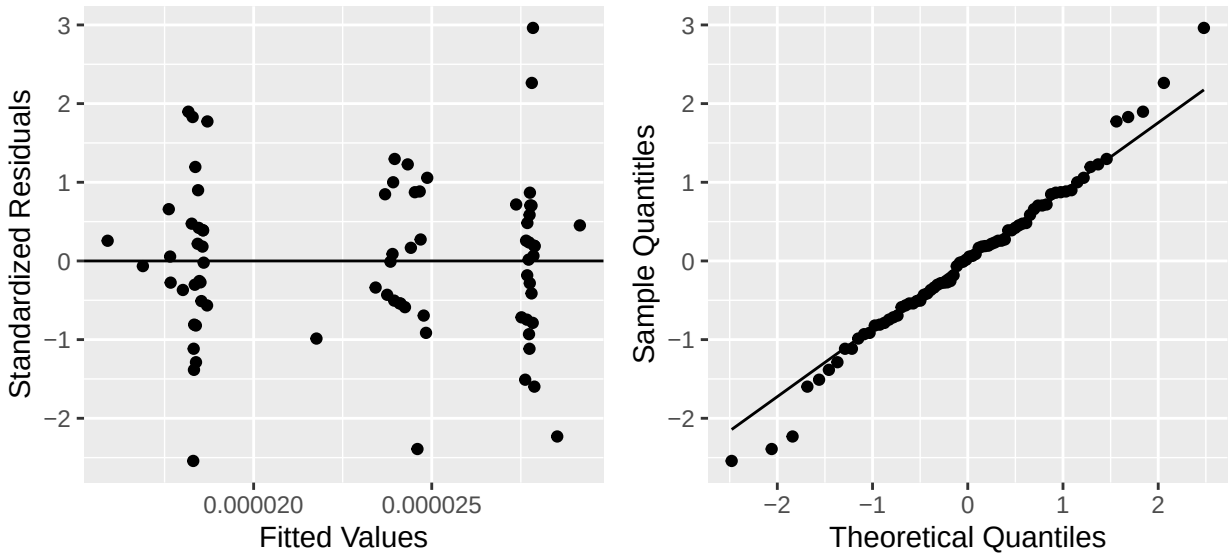

Figure 5: The residual plot (left) along with the Q-Q plot (right) for the final model.

c. Model Prediction

To confirm that the goal of creating a predictive model for median wages was achieved, 95% prediction intervals of $(\texttt{Median})^{-1}$ for selected majors were generated, as seen below in Table 1.

Major	Major Category	Men	Share Women	Median	Prediction Interval
Statistics & Decision Science	Computers & Mathematics	2960	0.5265	45000	(30997,61595)
Petroleum Engineering	Engineering	2057	0.1206	110000	(38461,94271)
Zoology	Biology & Life Sciences	3050	0.6373	26000	(28199,50201)
Astronomy & Astrophysics	Physical Sciences	832	0.5357	62000	(30976,61690)
Nursing	Health	21773	0.8960	48000	(27368,48764)

Table 1: Prediction intervals for the chosen majors.

Here, we find that the median wages for *Zoology*, *Astronomy & Astrophysics*, and *Petroleum Engineering* lie outside their prediction intervals, which makes sense given that they're outliers.

IV. Conclusion

a. Summary of Results

After the data analysis, the obtained results support that there is an association with gender and median wage in STEM majors. Since this is an association, we cannot assume that gender causes this difference. However, it would be interesting to see more research done in experimental research to find the causation of this discrepancy in median wage and gender.

Our final model has the capacity of predicting 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 tested our model on each major category, having the least success with majors having median wages at the extremes of the dataset or are outliers within their major category (i.e. Petroleum Engineering, Zoology, and Astronomy & Astrophysics).

The most 'important' conclusion from our research is that since Petroleum Engineering has the highest median salary in this data set (i.e. $\$110,000$), potential students should consider majoring in this field if they only care about the median salary. However, if more women go into this field, the median wage could potentially decrease.

b. Further Research

Despite the findings obtained, the data set was found to be too limited to get a thorough look at associations within STEM college majors that influence their median wages. For instance, if the data set was sex-disaggregated for median wage, it could be useful to see the difference in median wage by gender for each major. Another way to improve future research is to collect time series data so that analysis could be done to see how median wages change with an influx of women and/or exodus of men from a given major. Since this project only looked at STEM majors, it would also be interesting to see if these same variables (i.e. `Major_category`, `Men`, `ShareWomen`) are associated with the median wages for all majors.

Bibliography

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

2. 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