Gender Wage Inequality in STEM

Lydia Gibson, Sara Hatter and Ken Vu

Department of Statistics and Biostatistics, CSU East Bay

STAT 632 - Linear and Logistic Regression

Dr. Joshua Kerr

May 02, 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. 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 achievement orientation and ambitiousness, 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.

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 tries 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 has been already subsetted to only have 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. Within the STEM majors, median wage ranges from $\$26,000$ for Zoology and $\$110,000$ for Petroleum Engineering (Mdn = \$44350, M = \$46118).

The data dimensions are seventy-six majors in STEM fields (rows) and nine factors (columns), such as: `Rank`, `Major_code`, `Major`, `Major_category`, `Total`, `Men`, `Women`, `ShareWomen`, and `Median`. For this project, `Major_category` was set as a factor to explore the variation of share of women within major categories and the median wages each major category earns.

b. Exploratory Data Analysis

Through a stacked barplot of gender proportions per major category, the data showed 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 in addition to men.

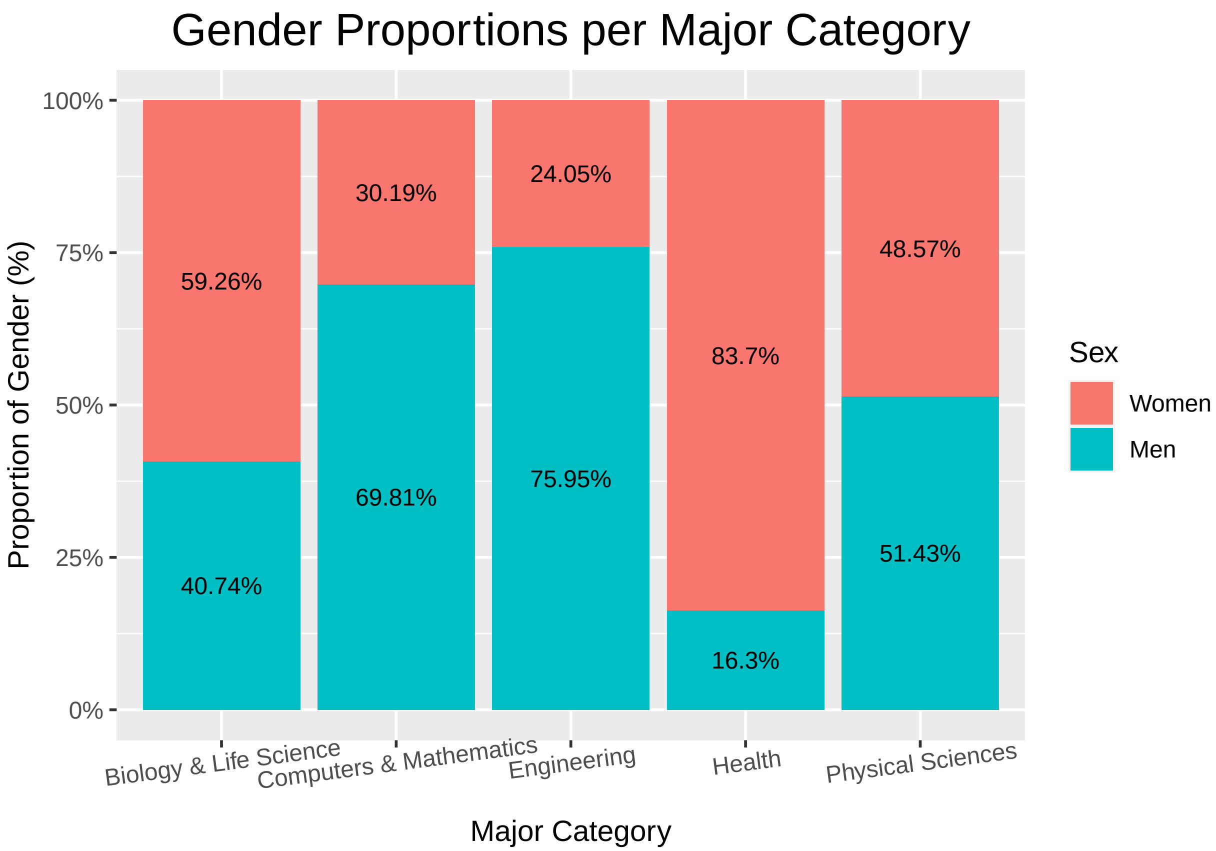


Figure 1: Gender proportions per major category.

Along with the stacked barplot, a boxplot was generated to help identify outliers in median wages within the major categories (see Figure 2).

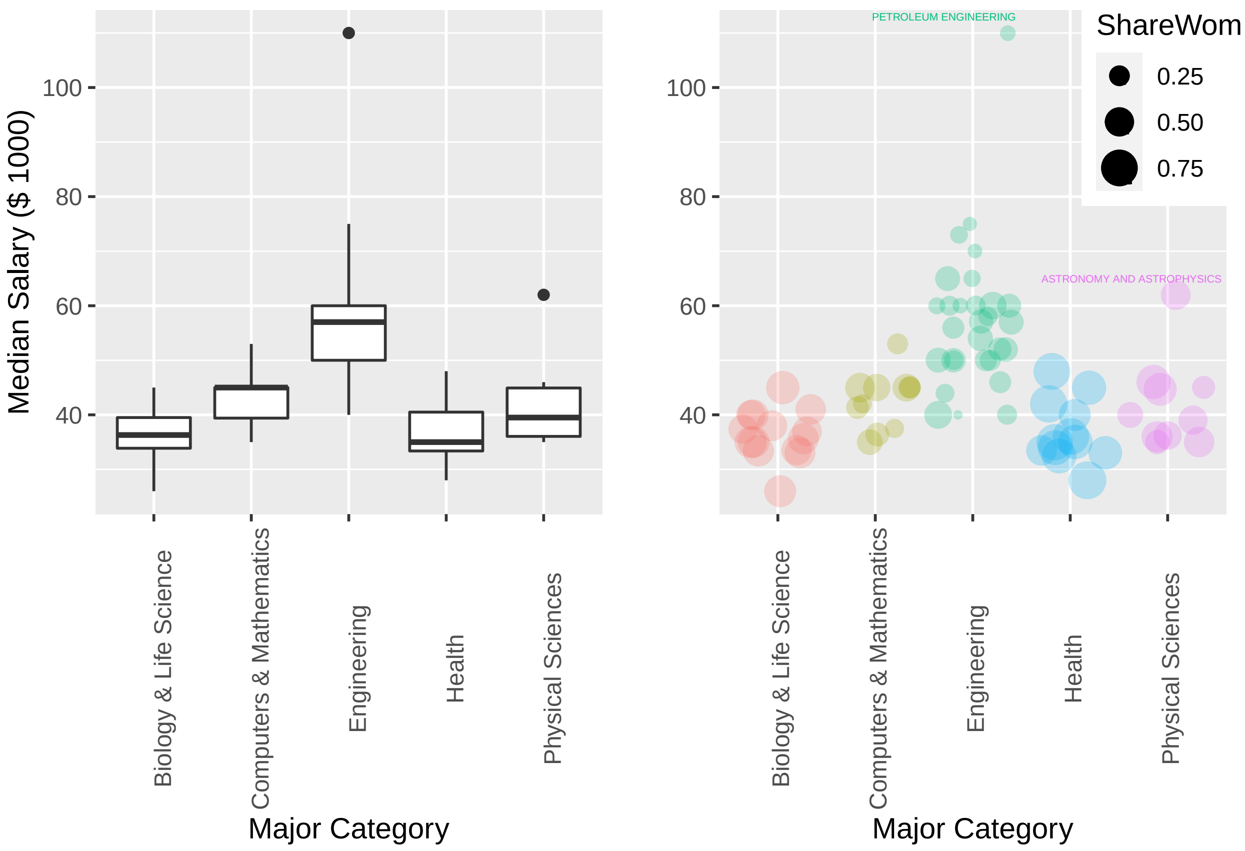


Figure 2: Median wage by major category.

Here, the boxplot confirmed that *Engineering* contains an outlier (i.e. "Petroleum Engineering") as well as another outlier for "Astronomy and Astrophysics". Both fields have less women compared to men. This procedure showed that there may be a significant difference between median wage by major category. Thus, ANOVA was done to test if there is actually at least one major category that its median wage is significantly different from the rest. The ANOVA test done supports the hypothesis with (F(4, 71) = [16.7], p-[0.00000001013]) $< \alpha = 0.05$.

Then, after removing the columns `Major_code` and `Rank` from our data (which were irrelevant to this project), a scatterplot matrix was created, which revealed that there seemed to be a negative association between `ShareWomen` and `Median` Also, the plot showed that there may be an issue of multicollinearity among `Total`, `Men`, `Women`, and `ShareWomen`. This observation makes sense since the column `Total` consists of the sum of the columns `Men` and `Women`. Likewise, `ShareWomen` refers to the proportion of women within the fields (see Figure 3).

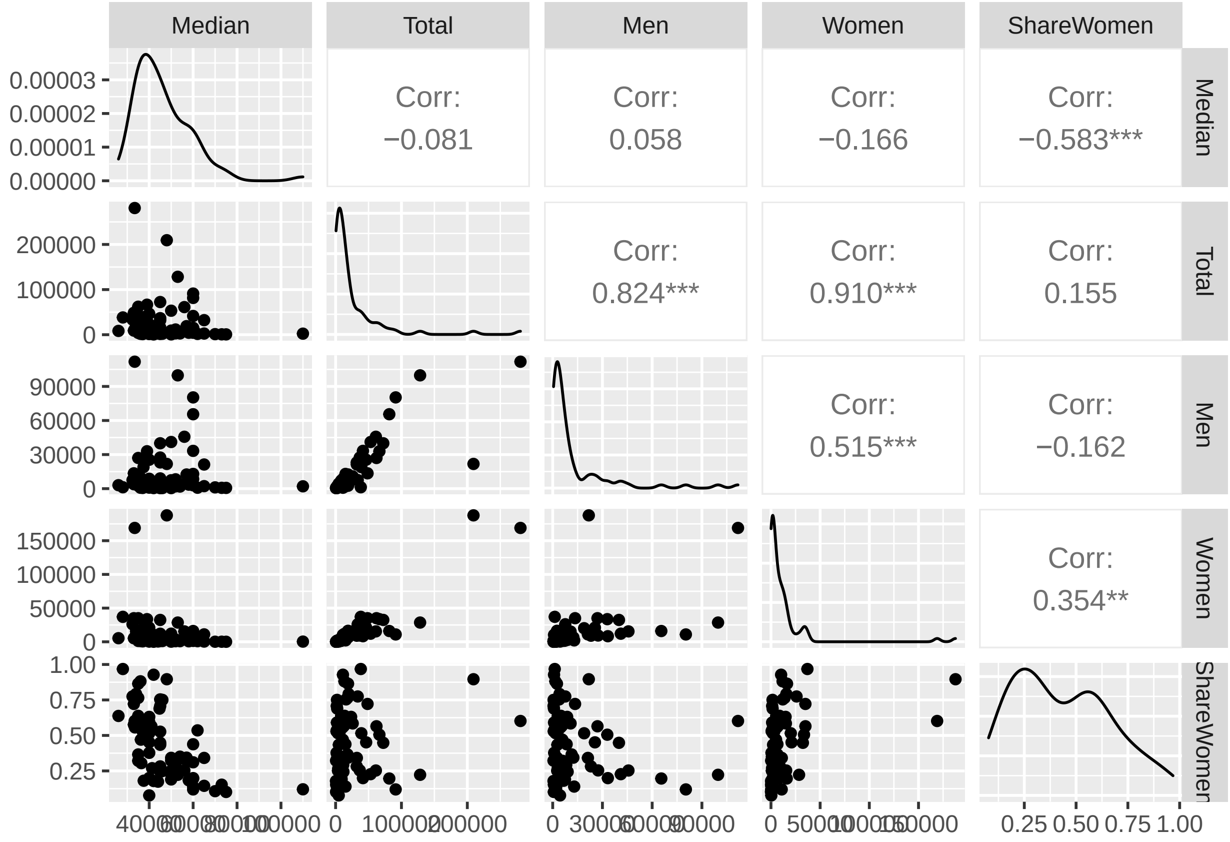


Figure 3: Scatterplot matrix for correlation insights.

c. Box Cox Transformation

The column for median wage `Median` was select as the response variable. During the process of checking normality, linearity, and constant variance, the data showed some skewing. Therefore, a Box-Cox test was performed to see if a transformation was necessary for `Median` (see Figure 4).

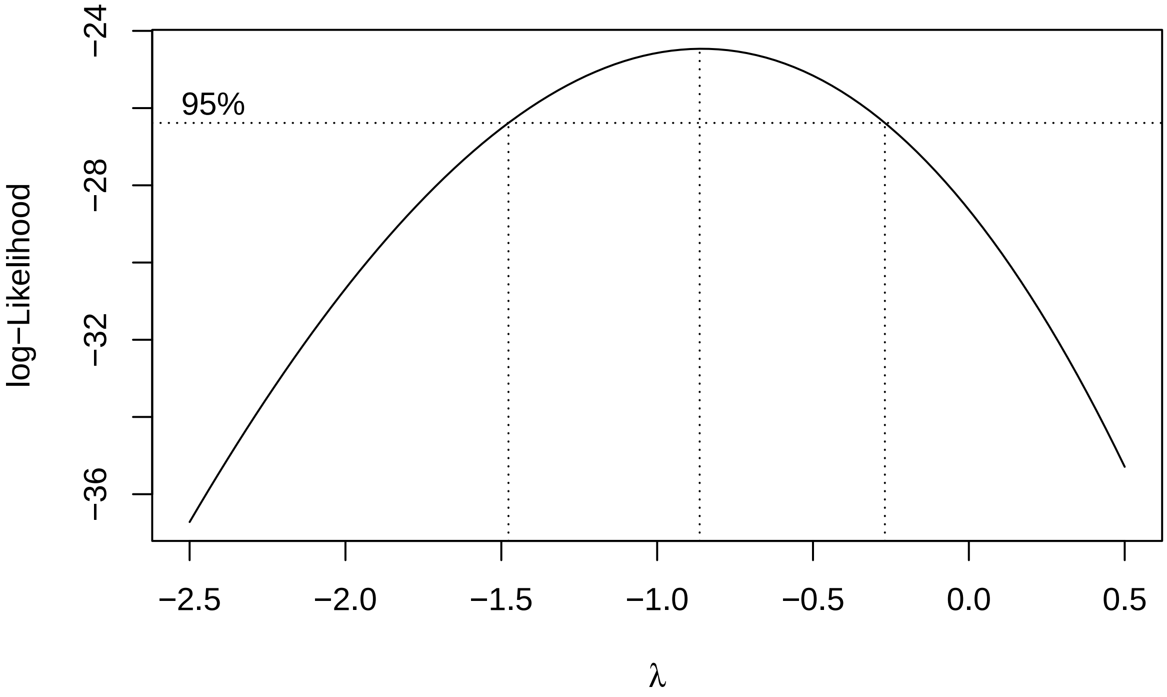


Figure 4: Box-Cox with a negative one as a power.

The resulting rounded power was `-1`, suggesting that an inverse transformation of the response `Median` wage is required to help with skewedness. However, this transformation would complicate the interpretability of the model.

III. Methods and Results

a. Model Fitting

The full addictive model is described by:

$$Y = \beta_0 + \beta_{Total} + \beta_{Men} + \beta_{Women} + \beta_{ShareWomen} + \beta_{Majorcategory} + \epsilon$$

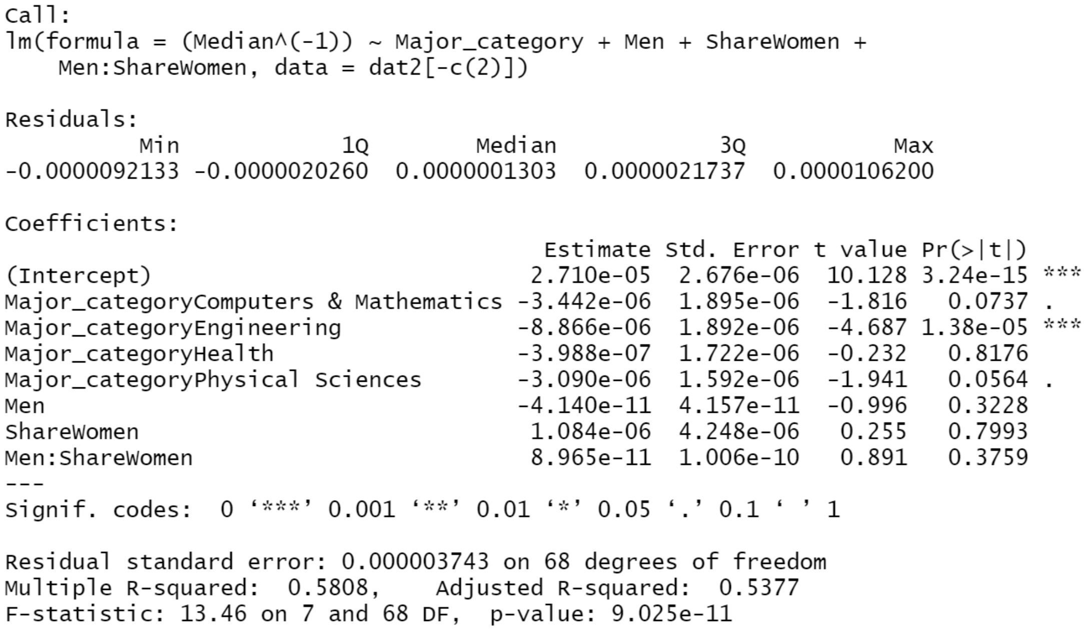
Running this model through the step-wise function using AIC as our criterion ended up removing too many predictors; thus, it was decided to check for interactions to see if this new model would help with this issue. Then, a step-wise process was run to reduce the model's AIC. This process aimed to remove the predictor women because the p-value was large (p = $0.7394 > \alpha = 0.05$).

The final reduced model is described by:

$$Y^{-1} = (2.71\times10^{-5}) - (3.44\times10^{-6})x_1 - (8.87\times10^{-6})x_2 - (3.99\times10^{-7})x_3 - (3.09\times10^{-6})x_4 - (4.14\times10^{-11})x_5 + (1.08\times10^{-6})x_6 + (8.98\times10^{-11})x_5 \cdot x_6$$

.

The results of the model fit can be found in the figure below.



```
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
```

Figure 5: Results of fitting the final model.

Looking at the summary of model fitting (see Figure 5), we can see that we have 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 can also see that the predictors `Men`, `ShareWomen`, and the interaction term `Men:ShareWomen` are not statistically significant at any significance level (given their p-values).

In addition, as noted earlier in the subsection **Box Cox Transformation** under the section **Data Description**, 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 (see Figure 6 below).

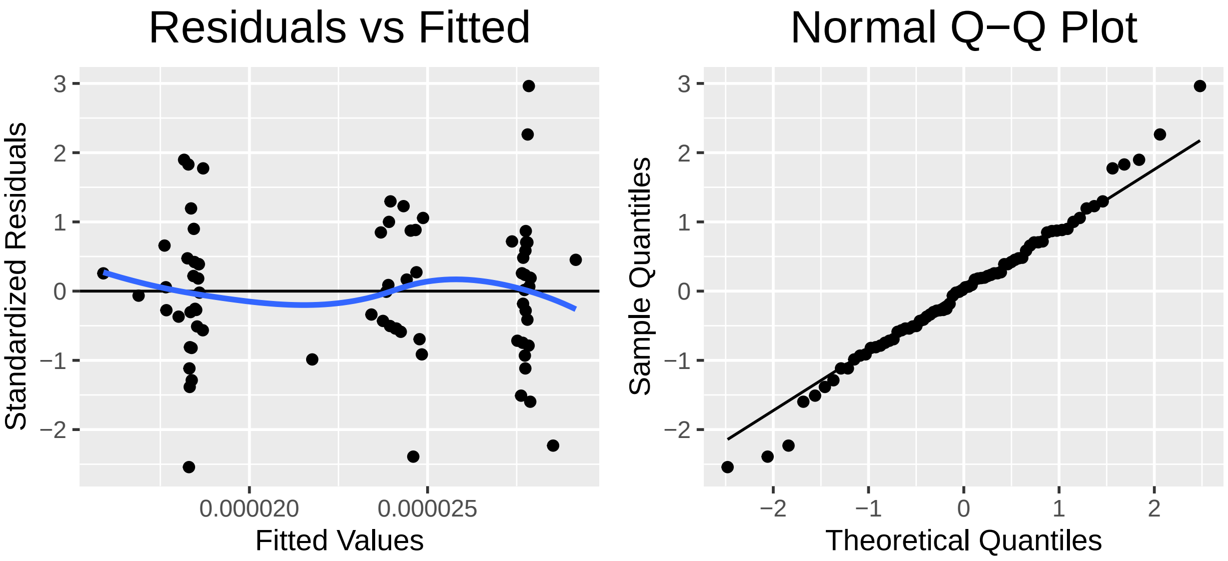


Figure 6: The residual plot (left) along with the Q-Q plot (right) for the final model.

Here, in Figure 6, 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 (which gives us p=0.8582 $>\alpha$ =0.05), it can be concluded that the assumption of constant variance for this data set holds up fairly well.

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-Wilk test was used, which gave this result - p=0.6165 $>\alpha$ =0.05. Therefore, it can be concluded that the standardized residuals generally follow a normal distribution so the normality assumption holds up here as well.

c. Model Prediction

To see if the goal of creating a predictive model for median wages was achieved, a prediction interval for $(\texttt{Median})^{-1}$ for Statistics and Decision Sciences was run. Also, to better understand the results the inverse was taken to see the answer in the original units (see Table 1).

Major	**Major Category**	**Men**	**Share Women**	**Median**
Statistics & Decision Science	Computers & Mathematics	2960	0.5265	45000

Table 1: The predicted median wage for Statistics and Decision Science is $45,000.

IV. Conclusion

a. Summary of Results

Based on the research done and the obtained results, here is a general summary of the key insights obtained from them:

- There is an association with gender and median wage of STEM majors.
- The median wage of STEM majors can be predicted based on the major category, total number of men in the major and total proportion of women in the major.
- Since Petroleum Engineering has the highest median salary in this data set (i.e. $110000) , potential students should consider majoring in this field if median salary is the only factor considered.

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. Thus, several potential opportunities for further research have been identified below:

- If the data set was sex-disaggregated for median wage, it could be used to see the difference in median wage by gender for each major.
- If time series data existed within this data set, analysis could be done to see how median wage changes with an influx of women and/or exodus of men from a given major.
- Since this project only looked at STEM majors, it would be interesting to see if these same variables (i.e. `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