MGT 6203 - Group Final Project

Trends and Explanation on the Gender Pay Gap in the U.S.A.

Team 79

Alex Atayev

Claudia Ezcurra

Taha Bekmez

Zhaowei Sun

TABLE OF CONTENTS

INTRODUCTION

It has been more than 50 years since President John F. Kennedy signed The Equal Pay Act of 1963, yet women in America still earn, on average, less than their male counterparts. Throughout the decades, the United States has implemented numerous laws to prevent employers from paying women less than men. Even with these efforts, the gender wage gap has barely narrowed, in fact, it has only changed by about a half a cent per year.

The issue of pay disparity between men and women has been studied extensively throughout the years and those studies have included legal, social, and economic factors. According to the studies: "there are two distinct numbers regarding pay gap: non-adjusted versus adjusted pay gap. The latter typically considers differences in hours worked, occupations were chosen, education, and job experience. In the United States, for example, the non-adjusted average female's annual salary is 79% of the average male salary, compared to 95% for the adjusted average salary." Gender pay gap (Wikipedia, n.d.)

Although gender pay inequality has decreased in the past few decades, the issue still exists. Despite all the different legislative actions and studies on the subject, it currently affects more than 55 million American women.

Problem Statement:

The purpose of this analysis is to gain insight into gender pay inequality in the U.S. from 1981 – 2011, examine its trends along with the primary factors affecting it, and determine possible reasons for its continued existence.

METHOD

Data

There are 2 data sets that were downloaded from the Kaggle website and are used in the project:

a) Panel Study of Income Dynamics (PSID),

b) Current Population Survey (CPS).

Based on our initial examination of the data, the PSID data set has 33,398 observations and 274 variables. The CPS data set has 344,287 observations and 234 variables. We decided to focus on men and women ages 25–64 who were full-time, nonfarm, wage, and salary workers and worked at least twenty-six weeks during the preceding year. We also excluded those in the military.

The age group, 25-64, has most likely left school, allowing us to avoid potential issues of combining work with attending school. Limiting the top of the age range to 64 helps avoid common retirement issues. In addition, the project is focused on those with a relatively market-labor solid commitment by narrowing the data to only those who worked full-time and had at least 26 weeks of work in the prior year.

Such a filter leads to a relatively homogeneous data sample regarding this commitment and allows us to draw more accurate conclusions about the pay men and women face in the labor market. Filtering out those who are self-employed and those in agriculture is needed because it is difficult to separate the labor income from the capital income for these population groups.

Filtered Data

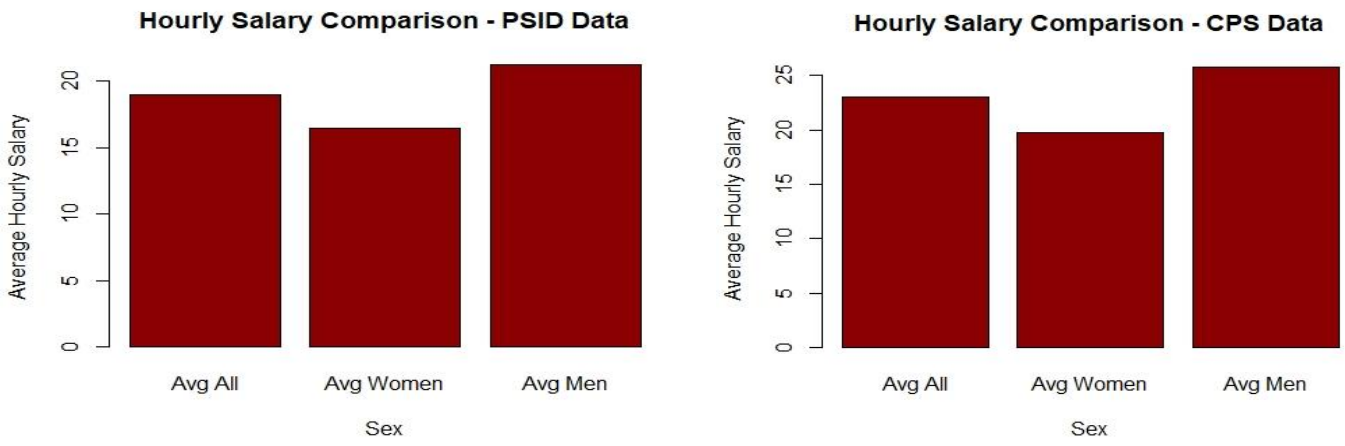
The summary of the filtered data is as follows: The filtered PSID dataset contains 27,795 observations, and the filtered CPS contains 218,752 observations.

We created a new variable to represent the Average Hourly Earnings by dividing the Annual Labor Income (annlabinc) by Annual Hours Worked (annhrs). The maximum Average Hourly Earnings of the filtered PSID data is 1928.571 while the minimum is 0.8928571. The maximum Average Hourly Earnings of the filtered CPS data is 612.2449 while the minimum is 1.587302.

According to the filtered PSID data, the mean of the Average Hourly Earnings of the entire survey population is 18.96083. The mean of the Average Hourly Earnings of the surveyed women is 16.4173, while for men, it is 21.21914. There is a difference of 4.80184.

According to the filtered CPS data, the mean of the Average Hourly Earnings of the entire survey population is 23.03214. The mean of the Average Hourly Earnings of the surveyed women is 19.79021, while for men, it is 25.75602. There is a difference of 5.96581.

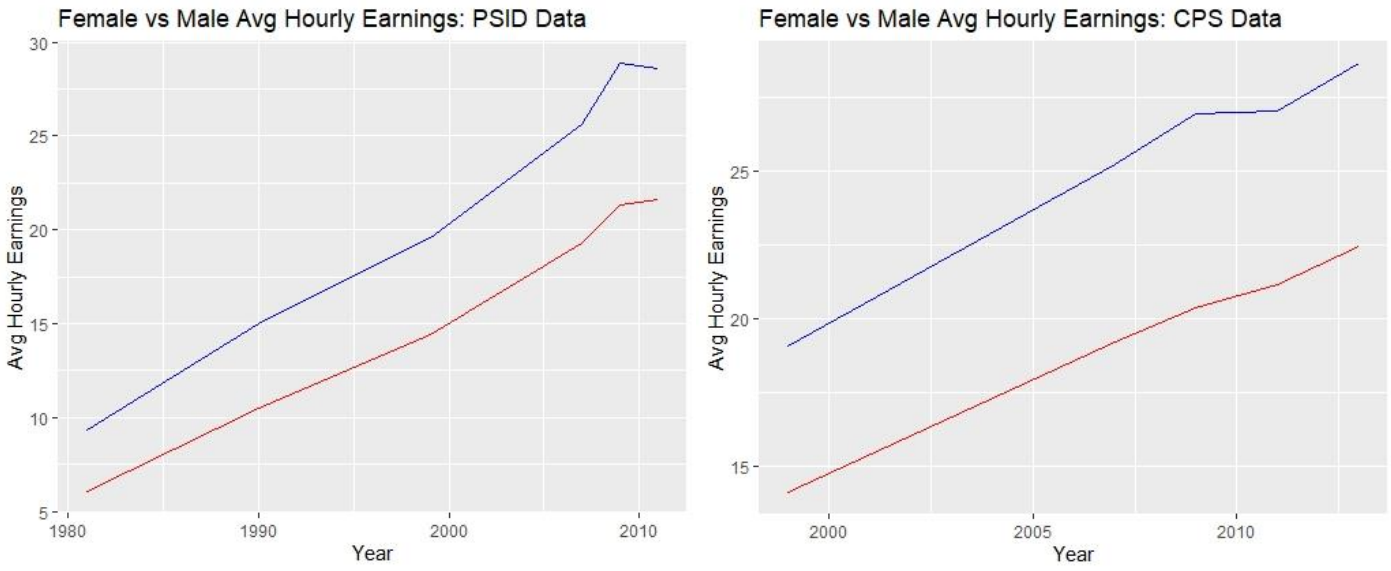
The above-mentioned differences in the Average Hourly Earnings between the PSID data and the CPS data are because these data sets have different years of data. In particular, the PSID data set contains the years 1981, 1990, 1999, 2007, 2009, and 011, while the CPS data set contains the years 1999, 2007, 2009, 2011 and 2013.



Hourly Salary Comparison (Population Mean vs. Women Mean vs. Men Mean)

Trends in Average Hourly Earnings and Female-to-Male Ratio

Because both data sets have several years of data, we wanted to see if there were any trends in the Average Hourly Earnings for males versus females over those years. The following two plots show the blue line for males and the red line for females. The plots reveal that despite the average hourly earnings for females were increasing over those years, so were the average hourly earnings for males. As a result, the pay gap was still present.



Trend in Average Hourly Earnings of Women vs. Men

To better see trends in the existing pay gap over years, we created another variable called the Female-to-Male Ratio which was calculated as the average hourly earnings for females divided by the average hourly earnings for males (presented as a percentage). The below plots show the Female-to-Male Ratio throughout the years. As the plots reveal, initially the ratio was improving for the females as it was raising from 65% in 1980 up to 76% -78% in 2007, but it stayed at the same level since then. To reiterate what the Female-to-Male Ratio means, 78% means that on every dollar that a male earns, a female only earns 78 cents.

Trend in Female-to-Male Ratio

The above-mentioned bar plots demonstrate that the ratio has been more consistent for the last 3 years. Therefore, to help explain the existing pay gap, further analysis of the variables has been narrowed to the most recent 3 years. We restricted the PSID data to include the years 2007, 2009, and 2011 and limited the CPS data to the years 2009, 2011, and 2013.

Training, Testing and Validation data sets

We are focusing on the most recent 3 years and assigning 70% of the data points to our Training set, and for the rest, 15% percent is allocated to the Testing set and 15% is allocated to Validation.

We did a splitting of the last 3 years of the PSID data (with 14726 observations) into the Training set (with 10308 observations), the Validation set (with 2209 observations), and the Testing set (with 2209 observations).

We did a splitting of the last 3 years of the CPS data (with 138028 observations) into the Training set (with 96619 observations), the Validation set (with 20704 observations), and the Testing set (with 20705 observations).

Union Coverage Models

Initially, we created models by taking the Union Coverage as one of the potential socio-economic factors that may explain the existing pay gap. We were trying to determine whether a Union Job makes the pay gap smaller.

The following are the frequency distribution by the 'female' and the frequency distribution by the 'unjob' in the last 3 years of the PSID data:

female	sex	n
<int>	<int>	<int>
0	1	7365
1	2	7361

unjob	n
<int>	<int>
0	12037
1	2689

While the CPS data set contains the 'female' variable, it does not have a binary variable 'unjob' and instead has the variable 'union' with the following values shown below in the frequency distribution for the last 3 years of the CPS data:

female	sex	n
<int>	<int>	<int>
0	1	74527
1	2	63501

union	n
<int>	<int>
0	115310
1	18910
2	3468
3	340

In the data dictionary of the CPS data, the following was described for the 'union' values: "(NIU=0, No union coverage=1, Member of labor union=2, Covered by union but not a member=3)". Therefore, in the CPS data we made a new binary variable (that has just 0 and 1 only) by following the CPS data dictionary description on the values in the 'union' variable. The frequency distribution by this new variable 'unjob' for the last 3 years is as follows:

unjob	union	n
<dbl>	<int>	<int>
0	0	115310
0	1	18910
1	2	3468
1	3	340

We first created a linear-linear regression model on both data sets where the dependent variable is the average hourly earnings, while the independent variables are the binary variables called “female” and “unjob” (Union Coverage).

The summary output of the linear-linear regression model on the PSID data shows the variable ‘unjob’ is not statistically significant. In addition, the Q-Q plots of the linear-linear regression models on both data sets suggest the relationships are most likely non-linear, and a log transformation is needed in both cases.

Please see below the summary outputs of the linear-linear regression models (for both PSID and CPS data) and their normal Q-Q plots.

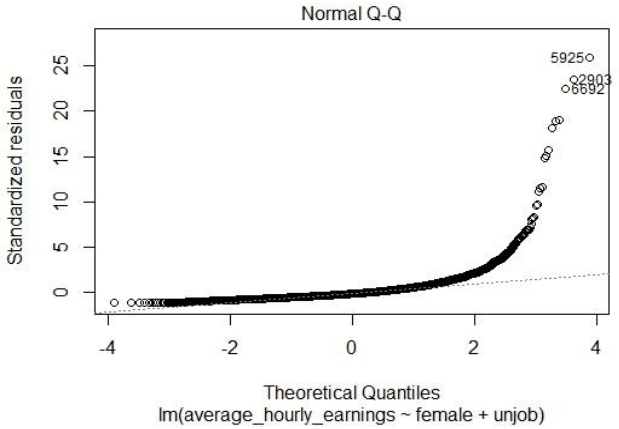
Summary Output of Union Coverage linear-linear regression model for PSID data and its Q-Q Plot:

```
Call:
lm(formula = average_hourly_earnings ~ female + unjob, data = PSID_data_training)

Residuals:
   Min     1Q Median     3Q    Max
-26.15 -10.17  -4.24   4.41 544.88

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  27.3674     0.3100  88.287   <2e-16 ***
female       -6.8657     0.4144 -16.570   <2e-16 ***
unjob         0.8587     0.5370   1.599     0.11
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 21.03 on 10305 degrees of freedom
Multiple R-squared:  0.02629,   Adjusted R-squared:  0.02611
F-statistic: 139.1 on 2 and 10305 DF,  p-value: < 2.2e-16
```



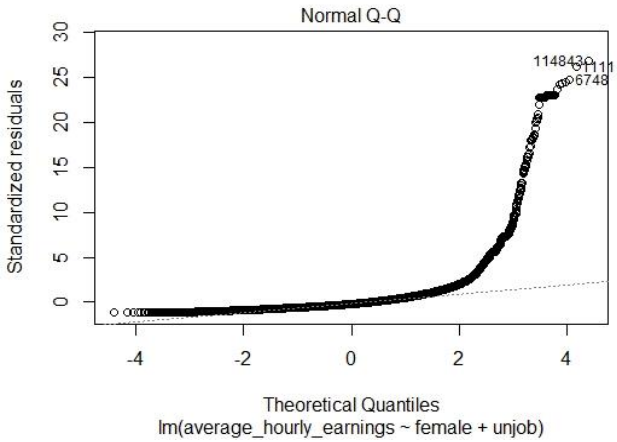
Summary Output of Union Coverage linear-linear regression model for CPS data and its Q-Q Plot:

```
Call:
lm(formula = average_hourly_earnings ~ female + unjob, data = CPS_data_training)

Residuals:
   Min     1Q Median     3Q    Max
-26.35 -10.75  -4.50   4.83 590.92

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 27.47344    0.09731 282.318   <2e-16 ***
female      -6.14752    0.14212 -43.255   <2e-16 ***
unjob        1.03790    0.43135   2.406   0.0161 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 22.02 on 96616 degrees of freedom
Multiple R-squared:  0.01906,   Adjusted R-squared:  0.01904
F-statistic: 938.7 on 2 and 96616 DF,  p-value: < 2.2e-16
```



For the log transformation, we used a natural log of the independent variable ‘average hourly earnings’ but kept the same dependent binary variables. After performing a log transformation on both datasets, we

noticed that the variable ' unjob' is statistically significant, and its coefficient is positive. This result means that if females are in a Union Job, their average hourly earnings are higher, and the gender pay gap is smaller.

The below log-linear models' Q-Q plots show a better linear relationship than the linear-linear models' Q-Q plots.

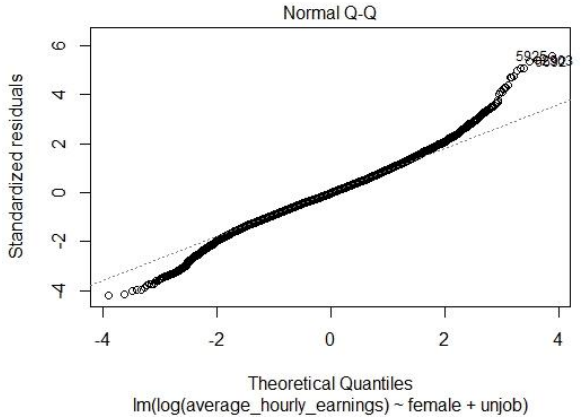
Summary Output of Union Coverage log-linear model for PSID data and its Q-Q Plot:

```
Call:
lm(formula = log(average_hourly_earnings) ~ female + unjob, data = PSID_data_training)

Residuals:
    Min      1Q  Median      3Q     Max
-2.4784 -0.3613 -0.0170  0.3490  3.2698

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.079819   0.008663 355.529   <2e-16 ***
female      -0.236128   0.011579 -20.392   <2e-16 ***
unjob        0.131197   0.015005   8.743   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5877 on 10305 degrees of freedom
Multiple R-squared:  0.04624,   Adjusted R-squared:  0.04606
F-statistic: 249.8 on 2 and 10305 DF,  p-value: < 2.2e-16
```



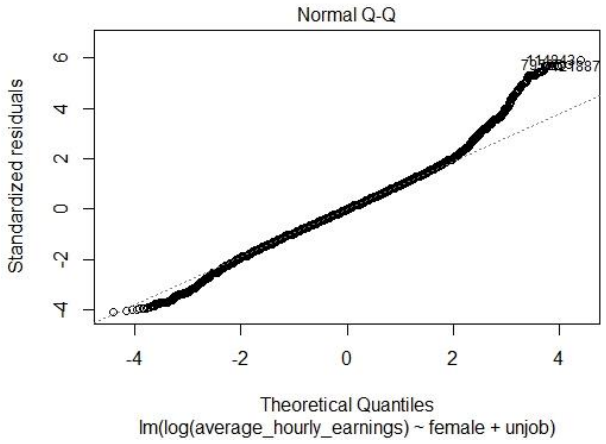
Summary Output of Union Coverage log-linear model for CPS data and its Q-Q Plot:

```
Call:
lm(formula = log(average_hourly_earnings) ~ female + unjob, data = CPS_data_training)

Residuals:
    Min      1Q  Median      3Q     Max
-2.4682 -0.3906 -0.0259  0.3780  3.5401

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.10054    0.00265 1170.14   <2e-16 ***
female      -0.22347    0.00387  -57.75   <2e-16 ***
unjob        0.13942    0.01175   11.87   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5996 on 96616 degrees of freedom
Multiple R-squared:  0.03477,   Adjusted R-squared:  0.03475
F-statistic:  1740 on 2 and 96616 DF,  p-value: < 2.2e-16
```



Upon further reviewing the data, we realized that if males or females were more present in the union jobs, it could potentially result in the females and unjob variables being highly correlated. Therefore, we needed to check for multicollinearity. We used the CAR package in R and ran the Variance Inflation Factor (VIF) function to calculate the VIF value of the variable 'unjob.' The VIF factor was less than 5, thus ensuring no multicollinearity.

The Union Coverage log-linear model was additionally applied to the validation and test data sets for additional R-squared estimations. The obtained predictions on the validation and test data sets had the R-squared consistent with the initial R-squared estimated on the training data set.

LASSO Models on Years of Experience and Years of Schooling

The PSID data contain numerous continuous variables associated with years of experience and years of schooling. To identify the variables that are strongly associated with the average_hourly_earnings variable, we used LASSO regression and selected these variables to conduct further analysis. However, we first scaled the continuous variables in the data.

In the PSID data, the following continuous variables were scaled and then used in LASSO regression along with the variable 'female':

1) Age # Age
2) sch # Highest Year of Schooling
3) schupd # updated years of schooling
4) yrsexp # Experience
5) yrsftexp # Full-Time Experience
6) yrsptexp # Part-Time Experience
7) yrsptexpsq # Part-Time Experience^2
8) yrsftexpsq # Full-Time Experience^2
9) yrsExpSq # Experience^2
10) yrsexpfz # Experience (filling in zeros)
11) yrsftexpfz # Full-time Experience (filling in zeros)
12) Yrsptexpfz # Years of Part-Time Experience (Filling in zeros)
13) yrsexpfzsq # Experience^2 (filling in zeros)
14) yrsftexpfzsq # Full-Time Experience^2 (filling in zeros)
15) potexp # potential experience (age-years of schooling-6) truncated to be between 0 and age-18
16) potexp2 # potential experience squared

We ran LASSO on the PSID scaled data and examined the selected variables' coefficients. Then, we created a new model using only the variables LASSO picked and ran VIF. The variable 'age' showed multicollinearity; therefore, it was excluded from the new model.

The summary output of the new linear-linear model shows that years of schooling and years of experience are statistically significant and have positive coefficients. Therefore, years of schooling and years of experience increase the females' average earnings and help narrow the gender pay gap.

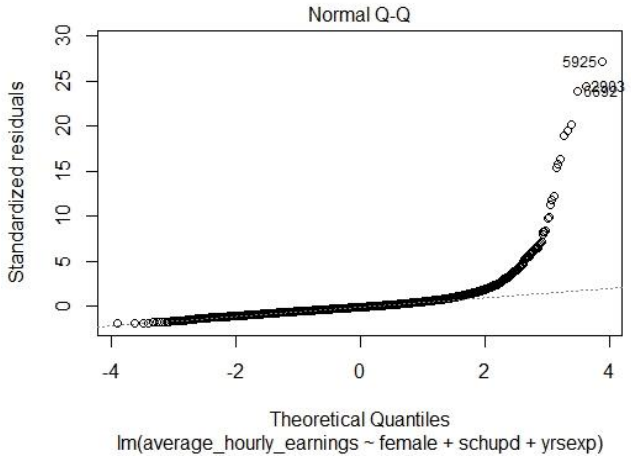
Summary Output of LASSO linear-linear model for PSID data and its Q-Q Plot:

```
Call:
lm(formula = average_hourly_earnings ~ female + schupd + yrsexp,
    data = PSID_scaled_data_comb)

Residuals:
   Min     1Q Median     3Q    Max
-38.42  -8.64  -2.46   4.90 526.16

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  27.8259     0.2704  102.92   <2e-16 ***
female       -7.4708     0.3831  -19.50   <2e-16 ***
schupd        7.6748     0.1917   40.03   <2e-16 ***
yrsexp        3.2721     0.1916   17.07   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 19.39 on 10304 degrees of freedom
Multiple R-squared:  0.1719,    Adjusted R-squared:  0.1717
F-statistic: 713.2 on 3 and 10304 DF,  p-value: < 2.2e-16
```



However, the above-presented Q-Q plot shows the relationship is most likely non-linear and will require a log transformation.

We created a new log-linear model on the PSID scaled data and used the variables LASSO selected. Its summary output shows that years of schooling and years of experience are statistically significant and have positive coefficients. The R-squared and Adjusted R-squared improved compared to the linear-linear model. The Q-Q plot also shows improvement.

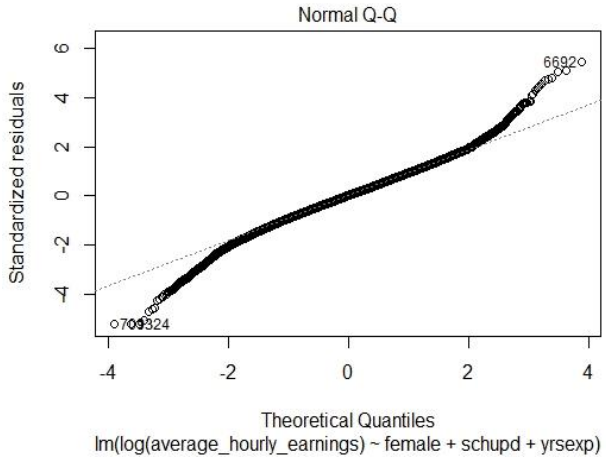
Summary Output of LASSO log-linear model for PSID data and its Q-Q Plot:

```
Call:
lm(formula = log(average_hourly_earnings) ~ female + schupd +
    yrsexp, data = PSID_scaled_data_comb)

Residuals:
     Min       1Q   Median       3Q      Max
-2.68744 -0.31457  0.01046  0.32472  2.79330

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.114990   0.007169  434.52   <2e-16 ***
female      -0.258742   0.010158  -25.47   <2e-16 ***
schupd       0.271349   0.005084   53.37   <2e-16 ***
yrsexp       0.119189   0.005081   23.46   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5142 on 10304 degrees of freedom
Multiple R-squared:  0.2698,    Adjusted R-squared:  0.2696
F-statistic:  1269 on 3 and 10304 DF,  p-value: < 2.2e-16
```



The LASSO log-linear model was additionally applied to the validation and test data sets for additional R-squared estimations. The obtained predictions on the validation and test data sets had the R- squared consistent with the initial R-squared estimated on the training data set.

Unlike PSID data, the CPS data set does not have numerous fields associated with years of experience: e.g., it only has the variables 'potexp' and 'potexp2'.

Therefore, the following continuous variables on years of education and years of experience were scaled and then used in LASSO along with the variable ‘female’:

1) age # Age
2) sch # Educational attainment
3) educ99 # Educational attainment for 1999 and later
4) potexp # potential experience (age-years of schooling-6)
5) potexp2 # potential experience squared

After we created LASSO for the CPS data, we examined the selected variables' coefficients and created a new model using only the variables LASSO picked and ran VIF. The variable 'age' showed multicollinearity; therefore, it was excluded from the new model.

The summary output of the new linear-linear model based on the CPS variables LASSO selected shows that years of schooling and years of experience are statistically significant and have positive coefficients. Therefore, years of schooling and years of experience increase the females' average earnings and help narrow the gender pay gap.

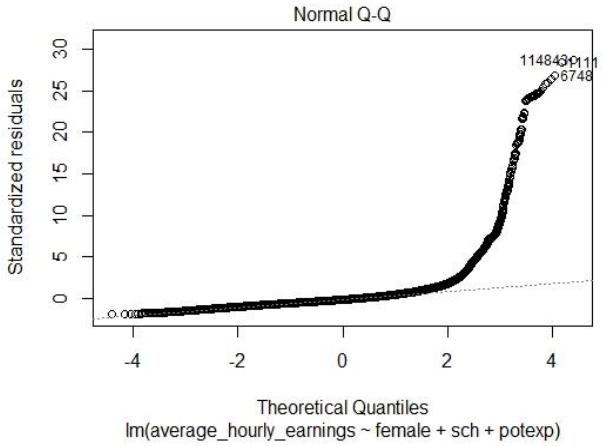
Summary Output of LASSO linear-linear model for CPS data and its Q-Q Plot:

```
Call:
lm(formula = average_hourly_earnings ~ female + sch + potexp,
    data = CPS_scaled_data_comb)

Residuals:
   Min     1Q Median     3Q    Max
-40.18  -9.06  -3.16   4.49 589.46

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 28.03741    0.09031  310.47   <2e-16 ***
female      -7.30713    0.13312  -54.89   <2e-16 ***
sch          7.99523    0.06769  118.11   <2e-16 ***
potexp       2.39696    0.06752   35.50   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 20.57 on 96615 degrees of freedom
Multiple R-squared:  0.144,     Adjusted R-squared:  0.1439
F-statistic:  5416 on 3 and 96615 DF,  p-value: < 2.2e-16
```



However, the above-presented Q-Q plot shows that the relationship is most likely to be non-linear which requires a log transformation.

We created a new log-linear model using the variables LASSO selected in the CPS scaled data. The summary output of the new log- linear model-based shows that years of schooling and years of experience are statistically significant and have positive coefficients. The R-squared and Adjusted R-squared improved compared to the linear-linear model. The Q-Q plot also shows improvement.

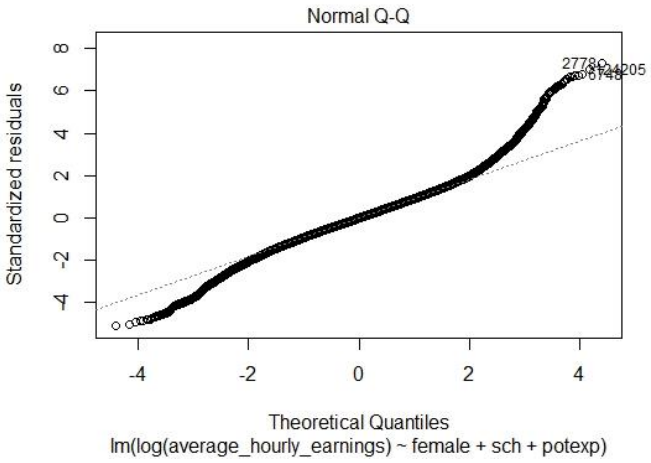
Summary Output of LASSO log-linear model for CPS data and its Q-Q Plot:

```
Call:
lm(formula = log(average_hourly_earnings) ~ female + sch + potexp,
    data = CPS_scaled_data_comb)

Residuals:
    Min      1Q  Median      3Q     Max
-2.6544 -0.3160 -0.0008  0.3183  3.7876

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.124928   0.002280 1370.55   <2e-16 ***
female      -0.267930   0.003361  -79.71   <2e-16 ***
sch          0.305994   0.001709  179.04   <2e-16 ***
potexp       0.088340   0.001705   51.82   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5194 on 96615 degrees of freedom
Multiple R-squared:  0.2757,    Adjusted R-squared:  0.2757
F-statistic: 1.226e+04 on 3 and 96615 DF,  p-value: < 2.2e-16
```



The LASSO log-linear model was additionally applied to the validation and test data sets for additional R-squared estimations. The obtained predictions on the validation and test data sets had the R- squared consistent with the initial R-squared estimated on the training data set.

Education Models

Based on our previous model analyses, the education factor seemed to be a significant predictor. There are three categorical (binary) variables in the PSID and CPS data that we used when creating the Education models to determine the importance of education types:

1) ba # Bachelor's Degree
2) adv # Advanced Degree
3) LEHS # High School or Less

The frequency distributions by those three variables for the last 3 years in both the PSID and CPS data are as follows:

PSID data:

female	ba	adv	LEHS	n
<int>	<int>	<int>	<int>	<int>
0	0	0	1	4969
0	0	1	0	745
0	1	0	0	1651
1	0	0	1	4784
1	0	1	0	937
1	1	0	0	1640

CPS data:

female	ba	adv	LEHS	n
<int>	<int>	<int>	<int>	<int>
0	0	0	1	47458
0	0	1	0	9877
0	1	0	0	17192
1	0	0	1	37811
1	0	1	0	9615
1	1	0	0	16075

We first prepared linear-linear regression models where the dependent variable is the average hourly earnings, while the independent variables are the binary variables called “female”, “ba”, and “adv”. Please see the summary outputs of the linear-linear regression models (for both PSID and CPS data) and their normal Q-Q plots below. The summary outputs show that the variables are statistically significant.

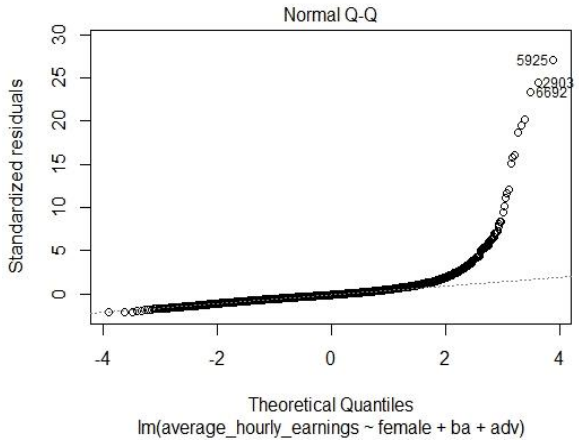
Summary Output of Education linear-linear model for PSID data and its Q-Q Plot:

```
Call:
lm(formula = average_hourly_earnings ~ female + ba + adv, data = PSID_data_training)

Residuals:
   Min     1Q Median     3Q    Max
-42.71  -8.56  -2.58   4.71 527.41

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  22.6008     0.3020   74.83   <2e-16 ***
female       -7.4170     0.3848  -19.27   <2e-16 ***
ba           12.0484     0.4724   25.50   <2e-16 ***
adv          22.2379     0.6161   36.09   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 19.52 on 10304 degrees of freedom
Multiple R-squared:  0.1613,    Adjusted R-squared:  0.1611
F-statistic: 660.8 on 3 and 10304 DF,  p-value: < 2.2e-16
```



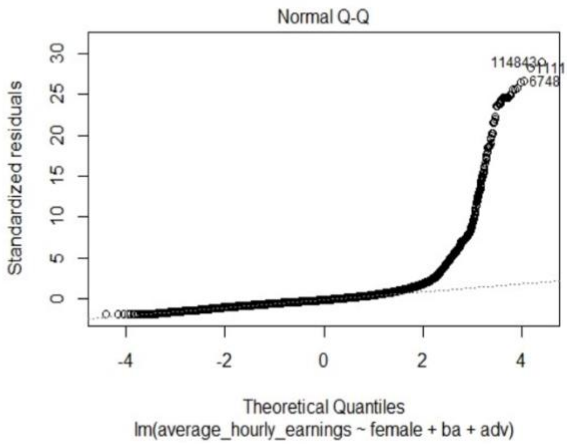
Summary Output of Education linear-linear regression model for CPS data and its Q-Q Plot:

```
Call:
lm(formula = average_hourly_earnings ~ female + ba + adv, data = CPS_data_training)

Residuals:
   Min     1Q Median     3Q    Max
-40.18  -9.02  -2.98   4.57 596.82

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  22.2094     0.1033  214.96   <2e-16 ***
female       -6.7834     0.1334  -50.86   <2e-16 ***
ba           11.0246     0.1594   69.16   <2e-16 ***
adv          20.7420     0.1964  105.64   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 20.65 on 96615 degrees of freedom
Multiple R-squared:  0.1376,    Adjusted R-squared:  0.1376
F-statistic:  5139 on 3 and 96615 DF,  p-value: < 2.2e-16
```



Because the above Q-Q plots showed that the relationships are most likely non-linear, we did a log transformation and created log-linear models for both the PSID and CPS data. To see if independent variables in the model are correlated, we performed VIF and did not find any multicollinearity.

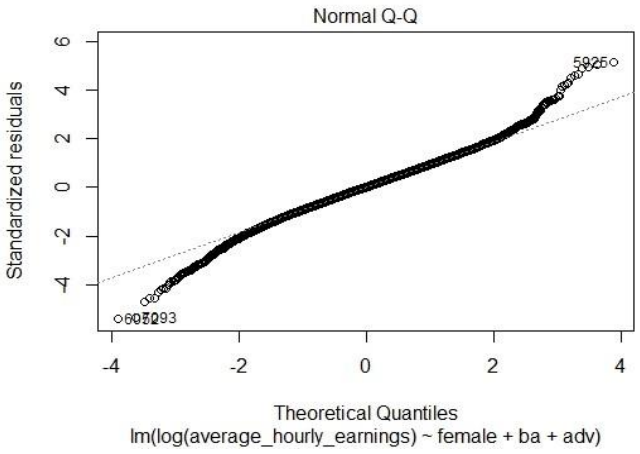
Summary Output of Education log-linear model for PSID data and its Q-Q Plot:

```
Call:
lm(formula = log(average_hourly_earnings) ~ female + ba + adv,
    data = PSID_data_training)

Residuals:
     Min       1Q   Median       3Q      Max
-2.87478 -0.32882  0.00788  0.33175  2.71976

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.935162   0.008196  358.12   <2e-16 ***
female      -0.254795   0.010442  -24.40   <2e-16 ***
ba           0.446569   0.012820   34.84   <2e-16 ***
adv          0.694645   0.016719   41.55   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5297 on 10304 degrees of freedom
Multiple R-squared:  0.2254,    Adjusted R-squared:  0.2252
F-statistic: 999.4 on 3 and 10304 DF,  p-value: < 2.2e-16
```



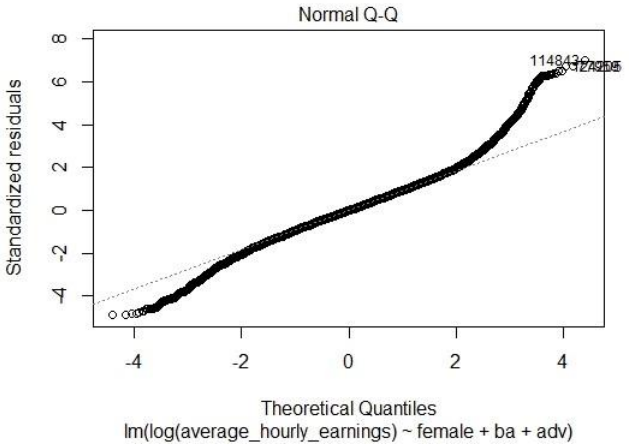
Summary Output of Education log-linear model for CPS data and its Q-Q Plot:

```
Call:
lm(formula = log(average_hourly_earnings) ~ female + ba + adv,
    data = CPS_data_training)

Residuals:
    Min      1Q  Median      3Q     Max
-2.6270 -0.3351  0.0027  0.3301  3.7510

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.912323   0.002685 1084.73   <2e-16 ***
female      -0.246221   0.003466  -71.04   <2e-16 ***
ba           0.433101   0.004143  104.55   <2e-16 ***
adv          0.694954   0.005102  136.20   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5366 on 96615 degrees of freedom
Multiple R-squared:  0.2271,    Adjusted R-squared:  0.2271
F-statistic:  9464 on 3 and 96615 DF,  p-value: < 2.2e-16
```



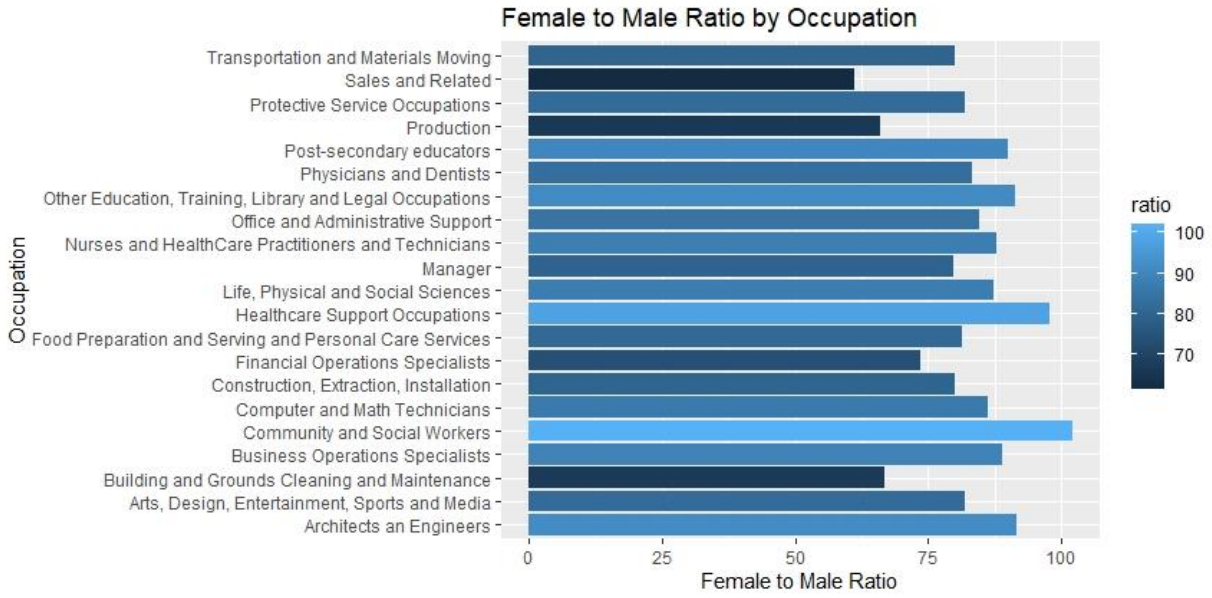
The Education log-linear model was additionally applied to the validation and test data sets for additional R-squared estimations. The obtained predictions on the validation and test data sets had the R- squared consistent with the initial R-squared estimated on the training data set.

CONCLUSIONS

1. Trend analysis showed that the Average Hourly Earnings was increasing for both males and females, the Female-to-Male Ratio was improving, which indicated that the gender pay gap was narrowing. However, for more recent years, the Female-to-Male Ratio has stayed on the same level.
2. Union Coverage log-linear models performed on both PSID and CPS data showed that Union Coverage increases the females' average earnings and narrows the gender pay gap
3. LASSO assessed several variables in both PSID and CPS data and narrowed them for further analysis. The log-linear models using the variables LASSO selected showed that Years of Experience and Years of Schooling increase the females' average earnings and narrow the gender pay gap
4. Education log-linear models performed on both PSID and CPS data showed that Bachelor's and Advanced degrees increase the females' average earnings and narrow the gender pay gap.

RECOMMENDATIONS FOR FUTURE ANALYSES

The following graph shows our preliminary examination of the data for the last 3 years. This revealed a significant variation in the Female-to-Male Ratio across different occupations.



Female-to-Male Ratios by Occupation

Since the PSID and CPS data sets contain binary variables showing which industry and occupation an employee belongs to, if given more time or resources, we would do a comparative analysis to reveal where the most significant gender pay gap still exists and we would also examine the reasons for such differences.

REFERENCES

1) Institute For Women's Policy Research. *The Economic Impact of Equal Pay by State*. https://statusofwomendata.org/featured/the-economic-impact-of-equal-pay-by-state/

2) Institute For Women's Policy Research. *Employment and Earnings.* *https://statusofwomendata.org/explore-the-data/employment-and-earnings/*

3) Bleiweis, R (2020, March 24). *Quick Facts About the Gender Wage Gap*. https://www.americanprogress.org/article/quick-facts-gender-wage-gap/

4) Wikipedia Contributors. Gender pay gap. (2022, September 11). In *Wikipedia, The Free Encyclopedia.* Retrieved October 8, 2022, from https://en.wikipedia.org/wiki/Gender_pay_gap#:~:text=The%20latter%20typically%20takes%20into,for%20the%20adjusted%20average%20salary.

5) fedesoriano. (January 2022). Gender Pay Gap Dataset. Retrieved (2022, September 11) from https://www.kaggle.com/fedesoriano/gender-pay-gap-dataset.