EPPS 6316 Paper Assignment

Nada Alasmi

Introduction

Through the years, a variety of researchers have found proof for an income disparity in the United States based on sex. Specifically, even when education is taken into account, women earn far less than men (Treiman & Terrell). The gap in income based on sex even applies to individuals with no work experience, and one publication has found that female college graduates find themselves paid less than their male counterparts (Corbett & Hill).

Because of the concerning nature of these findings, this paper will conduct further research into the issue of sex-based income disparity in the United States. The General Social Survey of 1996 will be used to determine if, one average, there is a difference between female and male income in the data. Because competing hypothesis suggest that income differences are impacted by marital status and age, I will control for these variables in my research.

Formal Presentation of Hypothesis

My research hypothesis is that there is a significant difference between income levels of males and females. This relationship holds even after controlling for age, years of education and marital status.

My null hypothesis is that there is no significant difference between the income level of males and females, even after controlling for age, years of education and marital status.

Research Hypothesis: $\beta_2 \neq 0$

Null Hypothesis: $\beta_2=0$

My regression model is a differential intercept model:

$$Y = \beta_1 + \beta_2 D_2 + \beta_3 D_3 + \beta_4 X_4 + \beta_5 X_5 + u_i$$

Y is real respondent income, B_2 is the dichotomous variable sex, B_3 is the dichotomous variable marital status (married/unmarried), B_4 is continuous variable years of education, B_5 is the continuous variable age. $D_2 = 1$ if respondent is female and =0 if respondent is male. D_3 is = 1 if respondent is married and =0 if respondent is unmarried.

Description of Data

The 1994 General Social Survey conducted responses from 2,992 individuals (General Social Survey, 1994). Below is a description of our variables.

Sex

"Sex" is a dichotomous variable defined as "respondent's sex" where the responses are male or female. Here I have coded "female" as the dummy variable where female = 1 and non-female, or male, = 0. A total 1,619 females answered this question, which comprises 55.75 percent of the total responses to this question. On the other hand, a total 1,285 males answered this question, which comprise 44.25 percent of the total responses to this question.

```
. tab female
```

RECODE of sex (respondent s sex)	Freq.	Percent	Cum.
0	1,285	44.25	44.25
1	1,619	55.75	100.00
Total	2,904	100.00	

The "sex" variable seems suitable for use in this analysis. First, total responses for this question, or N, is 2,904. This is a very large sample size. Second, out of a total 2,992 observations in the GSS, about 97 percent answered the question on sex and thus there are very few missing cases.

Married

"Married" is the second dichotomous variable defined as "martial status of person". Here I have coded "married" as the dummy variable = 1 and NOT married as = 0. Total responses, or N, is 2,903. We can see that 1,496 of the respondents to this question are married, which makes up 51.53 percent of total responses to this question. On the other hand, 1,407 of the respondents are unmarried, comprising 48.47 percent of total responses to this question.

The "married" variable has few missing cases. Out of a total 2,992 observations in the GSS, about 2903 answered the "married" variable.

```
. tab married
```

RECODE of mar1 (marital status of 1st person)	Freq.	Percent	Cum.
0	1,407	48.47	48.47
1	1,496	51.53	100.00
Total	2,903	100.00	

Education

"Educ" is a continuous variable defined as "highest year of school completed". Sample size is 2,895 respondents. Mean is 13.36477 years of education, standard deviation is 2.929417.

The "educ" variable has few missing cases. Out of a total 2,992 observations in the GSS, about 2895 answered the "educ" variable.

```
. sum educ
```

Variable	Obs	Mean	Std. Dev.	Min	Max
educ	2895	13.36477	2.929417	0	20

Age

"Age" is a continuous variable defined as "age of respondent". Sample size is 2898, mean age is 44.77709 and standard deviation is 16,8677. We have few missing cases. Out of a total 2,993 observations in the GSS, 2989 answered this question.

```
. sum age
```

Variable	Obs	Mean	Std. Dev.	Min	Max
age	2898	44.77709	16.8677	18	89

Suitablity of Data

As discussed above, our data does not have a large amount of missing cases, and all our variables were answered by around 95 percent of total respondents.

To determine initial impacts of collinearity in our data, we ran a pairwise test. At 0.5 and above, we may suspect collinearity to be a problem within our data. As indicated below, there is no correlation between independent variables that is above 0.5, thus collinearity is not a large enough problem in our data.

```
. pwcorr realrinc sex age educ married
```

	realrinc	sex	age	educ	married
realrinc	1.0000				
sex	-0.2555	1.0000			
age	0.2350	0.0301	1.0000		
educ	0.3539	-0.0586	-0.1655	1.0000	
married	0.1245	-0.0927	-0.0080	0.0609	1.0000

To test whether our data is normally distributed, we ran the sktest for normality within our data. As we can see from the results below, we reject the null that our data is not skewed for all the variables except for married. Further, we reject the null that our data does not suffer from kurtosis in all the variables. This means that kurtosis and skewness are inherent in our data, and thus our data is not normally distributed.

Skewness/Kurtosis tests for Normality

				joint	
Variable	Obs	Pr(Skewness)	Pr(Kurtosis)	adj chi2(2)	Prob>chi2
realrinc	1.9e+03	0.0000	0.0000	.	0.0000
sex	2.9e+03	0.0000	.	.	.
age	2.9e+03	0.0000	0.0000	.	0.0000
educ	2.9e+03	0.0013	0.0000	51.30	0.0000
married	2.9e+03	0.1765	.	.	.

To make a decision on the impact of skewnes and kurtosis on our data, we determined the exact numerical values of skewness and kurtosis. The further the value is from zero, the more skewed the data is. As we can see below, our skweness values are quite close to zero, which indicates that skewness is not a very large probme in our data. In terms of kurtosis, the ideal value is 3. As we can see below, our kurtosis values are very close to 3, except for realrinc. We decide that the degree of kurtosis and skewness is not large enough to impact our hypothesis tests later on.

```
. tabstat realrinc sex age educ married, stats(skewness, kurtosis)
```

stats	realrinc	sex	age	educ	married
skewness	2.173685	-.2315642	.6032969	-.1464862	-.0613447
kurtosis	8.46064	1.053622	2.581747	3.900781	1.003763

In sum, we decide that our data is suitable for use in the regression. We do not have many missing cases to impact our results. Our data has reasonable standard errors, and finally, the degree of skewness and kurtosis is not large enough to impact our hypothesis tests later on.

Results

After running a standard OLS regression on our data, we first notice that the number of observations is 1,941. This means that the regression only utilized about 67 percent of the total responses to the General Social Survey. While a large chunk of our initial data that has been ingored in the regression, we will still accept the further results due to how large the sample size of 1,941 is.

We also notice that each variable has a coefficient. The coefficent for "age" indicates that, holding every other variable constant, each year of age increases realrinc by $328.9466. The coefficient for "married" indicates that, holding all other variables constant, individuals who are married start off $3,010 richer. The coefficient for "sex," our primary coefficent of interest, indicatest that, holding everything else constant, females start out $9,664 poorer just for being females. This effect of being female on the starting off realrinc is larger than the effect of being married. The coefficient for "educ" indicates that, holding everything else constant, an individual receives $2,592 more in income for every year of education. Finally, holding everything else constant, everyone's income starts at -$2,44853.08 dollars.

```
. regress realrinc age married female educ
```

Source	SS	df	MS		
Model	1.8251e+11	4	4.5628e+10	Number of obs	= 1941
Residual	5.5178e+11	1936	285011545	F(4, 1936)	= 160.09
				Prob > F	= 0.0000
Total	7.3430e+11	1940	378502579	R-squared	= 0.2486
				Adj R-squared	= 0.2470
				Root MSE	= 16882

realrinc	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
age	328.9466	30.77489	10.69	0.000	268.5912	389.302
married	3010.477	775.8215	3.88	0.000	1488.943	4532.01
female	-9664.643	769.579	-12.56	0.000	-11173.93	-8155.352
educ	2592.921	143.4819	18.07	0.000	2311.526	2874.317
_cons	-24453.08	2396.686	-10.20	0.000	-29153.44	-19752.73

We notice that our R^2 is equal to .2486. This means that variablity in our independent variables predicts 24.86 percent of variablility in our dependent variable. This is a large percent, and indicates that we have a solid model. In addition, our F test probality is less than 0.05. This means that our overall regression is statistically significant, and at the 0.05 level of significance, we can reject the null hypothesis that our large model (whicn includes control variables) is due to samping error alone. We use the 0.05 level of significane as it is the standard in social sciences.

We look further to find whether we can reject the null hypothesis for each variable that we included as as control (age, married, educ) , and we find that their t-test probabilities are statistically significant. It is good that we included them in our model as they do have an impact on our dependent variable that is not due to random chance alone at the 0.05 level of significance.

Finally, on our main hypothseis of interest, we find that, holding age, married and education constant, females start their income at $9664 less then men. Our t-test statistic for female (12.56) was the second largest in our regression, supreceded by the t-test of education. Finally, our t-test result for female is statistially significant at the 0.05 level of significance. Thus we reject our main null hypothesis, and state that, holding education, marital status and age constant, the difference in income between males and females is statistically significant, i.e. not due to random error alone.

Post-Estimation Diagnostics

Our results above indicate that being female has a statistically significant impact on one's income. We will finally test for the degree of collinearity, heteroscedasticity and autocorrelation in our data to determine if these issues have an impact on our hypothesis testing.

Tests for Collinearity

We run the pwcorr test again and find that none of our correlations between our independent variables are equal to 0.5. While we do have coliniarity in our data, it is not large enough to have a significant impact on our regression results.

```
. pwcorr realrinc female age educ married

             | realrinc   female      age     educ  married
-------------+---------------------------------------------
    realrinc |   1.0000
      female |  -0.2555   1.0000
         age |   0.2350   0.0301   1.0000
        educ |   0.3539  -0.0586  -0.1655   1.0000
     married |   0.1245  -0.0927  -0.0080   0.0609   1.0000
```

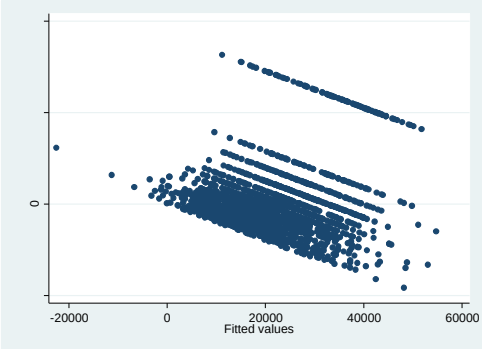
To further reassure us that collinearity is not a large enough problem in our data, we run the VIF test. If our VIF results are greater than 10, we have a collinearity problem. As we can see below, all our VIF results are approximately 1.0, not at all close to 10. We can thus be confident that collinearity does not have a large enough impact to distrub our regression results.

```
. estat vif
```

Variable	VIF	1/VIF
married	1.02	0.980714
age	1.01	0.986271
female	1.01	0.991726
educ	1.00	0.999104
Mean VIF	1.01	

Tests for Heteroscedasticity

We use two tests to determine the impact of heteroscedasticity on our results. The first is the table below which error terms with our fitted values. We see a slopng pattern in the error terms which indicates heteroscedasticity. We further conduct a Breush-Pagan test for our regression. The test rejects the null of homoskedasticity. Thus we are more aware that heteroskedasticity is a problem in our regression. Heteroskedasticity is unequal variance in error terms, which is imporatant because it reduces our ability to say that our regression coefficents are BLUE, and it impacts our hypothesis tests.




```
. estat hettest

Breusch-Pagan / Cook-Weisberg test for heteroskedasticity
         Ho: Constant variance
         Variables: fitted values of realrinc

         chi2(1)      =   526.35
         Prob > chi2  =   0.0000
```

Tests for Autocorrelation

Because our data does not have any inherent time or place order to it, instead all results are taken at one point in 1994, autocorrelation does not have a significant impact on our data and we do not test for it.

Fixing for Heteroscedasticity

Our tests above indicated that heteroscedasticity could have an impact on our regression results. It specificallly could impact our hypothesis test decisions and our ability to declare our regression coefficients as BLUE, or best linear unibased estimators. We run a robust standard regression. Coefficients are unchanged. Almost all our t-scores all the same, except for the t-score for education which increased by 4 units. Despite this, we still reject the null for the education variable, indicating that heteroscedasticity did not cause a type 1 or type 2 error. Specifically for our variable of interest "female", we find the same t-test value and same decision on the null: to reject. Finally, our F-test value decreased from 160.09 to 97.32. While this indicates that heteorskedasticity has an impact on our F-test value, it does not impact our decision on the null. In sum, by adjusting for it in the robust standard regression, we were able to see that the heteroscedasticity in our original OLS regression was not large enough to impact our decision on our null hypothesis of interest.

```
. regress realrinc educ married female age, robust

Linear regression                               Number of obs =    1941
                                                F(  4,  1936) =   97.32
                                                Prob > F      =  0.0000
                                                R-squared     =  0.2486
                                                Root MSE      =   16882
```

realrinc	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf.	Interval]
educ	2592.921	173.8794	14.91	0.000	2251.911	2933.932
married	3010.477	750.804	4.01	0.000	1538.007	4482.946
female	-9664.643	767.197	-12.60	0.000	-11169.26	-8160.024
age	328.9466	32.75718	10.04	0.000	264.7035	393.1896
_cons	-24453.08	2683.04	-9.11	0.000	-29715.04	-19191.13

Conclusion

Our paper was written to determine if there is a statistically significant impact of being female on one's income. We found that females make a $9,664 less then men, a difference that is statistically significant at the 0.05 level. The difference is not due to random chance alone. We found evidence to include the control variables that we did due to our large R^2 value. We did not find evidence for collinearity, and by controlling for heteroscedasticity in a robust regression, we found that it did not change our decision on the hypothesis tests. In sum, our paper is an addition to the research that indicates a sex-based disparity of income in the United States.

Works Cited

Corbett, C., & Hill, C. (2012). *Graduating to a Pay Gap: The Earnings of Women and Men One Year after College Graduation.* Washington DC: AAUW.

General Social Survey, 1994. (n.d.). Retrieved from The ARDA: http://www.thearda.com/Archive/Files/Codebooks/GSS1994_CB.asp

Treiman, D. J., & Terrell, K. (1975). Sex and The Process of Status Attainment: A Comparison of Working Women and Men. *American Sociological Review, 40 (April)*, 174-200.