**Final Project for APSTA-GE-2001**

**Yewei Sun**

Question 2:

summarize female, detail

tabstat (female), stats (n, mean, min, max) //Q2

See output results of descriptive statistics in Figure 1

Mean of *female*: 0.5316

SD of *female*: 0.4994

The mean of *female* represents the percentage of female participate in the study, since the value of 1 represents “yes” to the “Is R female” question.

tabstat weekly\_wage educ,stat(sd mean p25 p50 p75 skew min max)

文本

描述已自动生成

Question 3:

by female, sort: tabstat weekly\_wage educ, stat(sd mean p25 p50 p75 skew min max)

According to the output results (see in Figure 2), the primary conclusions we can draw were that on average, male participants (mean = 519.336) earn approximately $200 more than female participants (mean = 311.4533). The standard deviation of male group is larger than female group, indicating a larger range of distribution in income of male than income of female. The difference between the interquartile ranges (IQR) of male group and female group reflects a similar trend, the IQR of male group is larger than female group, represents a larger spread in income of male than female. The skewness of both male group and female group are very close to 1, which means that they are approximately symmetric. The minimal wage of both male (=4.98) and female participants (=5.466) is mostly identical, but the maximal wage reported by male participants (=1780.654) is much higher than female participants (=1325).

Question 4:

graph box (weekly\_wage), over (female) marker(1,mlabel( uniqid ))

graph box (educ), over (female) marker(1,mlabel( uniqid ))

by female, sort: tabstat weekly\_wage educ, stat(count min max p25 p50 p75)

See boxplot of weekly wages in Figure 3, boxplot of highest year of school completed in Figure 4, and detail results of five-number summary in Figure 5.

According to the boxplot of weekly wage (Figure 3), the range of male group is larger than the range of female group. The median and the interquartile ranges of male group are also larger than female group. The male and female group have similar number of outliners, but those of the male group spread more separately higher positioned than those of female group. The maximal number of the male group is larger than the female group.

As for highest year of school completed, the longest and shortest years of male and female group reported, and the interquartile ranges are identical. But the median of the male group (=14 years) is higher than female group (=13 years).

Question 5:

pwcorr female educ weekly\_wage, obs

See output as Figure 6.

Question 6:

According to Cohen’s rule, the correlation between the variable of *female* and *educ* is -0.0751 (N=570), indicating a very weak negative correlation between sexuality and highest year of school completed. The result of correlation means that female participants in this study has slightly lower years of education than male participants.

The correlation between the variable of *female* and *weekly\_wage* is -0.3593 (N=570), indicating a moderate negative correction between sexuality and weekly wage they earn. The result means that many female participants in this study earn lower wages than male participants.

The correlation between the variable of *educ* and *weekly\_wage* is 0.3163 (N=570), indicating a moderate positive correction between highest year of school completed and weekly wage they earn. The result means that those who spend longer years in school have a moderate chance to earn higher wages.

Question 7:

graph twoway (scatter educ female) (lfit educ female), ytitle (Education) xtitle (Gender) title(Education vs. Gender)

See result in Figure 7.

Question 8:

Reasons for the slope of the regression line to be reasonably flat:

1. The variable *female* is a dichotomous variable, and the variable *educ* is an interval variable. So the *r* coefficientis calculated by the point biserial correlation.
2. The slope of the regression line represents the strength of the correlation relationship between two variables, the stronger the relationship, the slope is closer to 1, and geometrically represents as a vertical regression line. The correlation coefficient between *female* and *educ* is -0.0751 (very close to 0), so the regression line appears to be negative, and very close to being a flat line.

Question 9:

graph twoway (scatter weekly\_wage educ ) (lfit weekly\_wage educ), ytitle (weekly wage) xtitle (Education) title(Wages vs. Education)

See result in Figure 8.

Question 10:

regress weekly\_wage educ, beta

See results of the regression in Figure 9.

Based on the results, the regression equation is: , Y=weekly\_wage, X=educ.

Question 11:

The p values of regression coefficient and intercept are both lower than 0.05, suggesting that they are accurate for predicting the Y value.

The slope of 40.1929 tells us that one year increase in education is associated with $40.1929 increase in weekly wage, on average. The intercept of -161.7683 is not meaningful because the minimal years of education in this dataset is 12 years instead of 0 year. R2 of this regression equation = 0.1, indicating that 10% of weekly wage can be explained by highest years in school.

The value = .3162779, suggesting that increase of one standard deviation in x variable *educ* will result in 0.3162779 standard deviation increase in y variable *weekly\_wage*.

Question 12:

regress weekly\_wage educ if female==1, beta //Q10 for female

regress weekly\_wage educ if female==0, beta //Q10 for male

See output results in Figure 10 and Figure 11.

From the results of the regression, we can see that, in general, the relationship between weekly wages and educ are similar.

**Female:**

The p values of regression coefficient and intercept are both lower than 0.05, suggesting that they are accurate for predicting the Y value.

The slope of 34.71371 tells us that one year increase in education is associated with $34.71371 increase in weekly wage of female, on average. The intercept of -203.8716 is not meaningful because the minimal years of education of female in this dataset is 12 years instead of 0 year. R2 of this regression equation = 0.15, indicating that 15% of female’s weekly wage can be explained by highest years in school.

The value = .3873419, suggesting that increase of one standard deviation in x variable *educ* will result in 0. 3873419 standard deviation increase in y variable *weekly\_wage*.

**Male:**

The p value of regression coefficient is lower than 0.05, suggesting that they are accurate for predicting the Y value. However, the p value of intercept is higher than 0.05, indicating that it may not be a precise predictor variable for this regression.

The slope of 37.23784 tells us that one year increase in education is associated with $37.23784 increase in weekly wage of male, on average. The intercept of -16.08 is not meaningful because the minimal years of education of male in this dataset is 12 years instead of 0 year. R2 of this regression equation = 0.0712, indicating that 7.12% of female’s weekly wage can be explained by highest years in school.

The value = .2669259, suggesting that increase of one standard deviation in x variable *educ* will result in 0. 2669259 standard deviation increase in y variable *weekly\_wage*.

Question 13:

graph twoway (scatter weekly\_wage educ if female == 0, ms(+)) ///

(lfit weekly\_wage educ if female == 0) ///

(scatter weekly\_wage educ if female == 1, ms(X)) ///

(lfit weekly\_wage educ if female == 1), ///

legend(row(1) order(2 4) label(2 "male +") label(4 "female X")) ///

ytitle(Weekly Wage)

See graphic output in Figure 12

Question 14:

The mean weekly wage for female = 499.2446 (Xmean=14.0363)

The mean weekly wage for male = 519.33609 (Xmean=14.37828)

Educmale = 12, Weekly Wagemale = 430.77408

Educmale = 16, Weekly Wagemale = 579.72544

Educmale = 20, Weekly Wagemale = 728.6768

From the regression lines, we can see that males get more wages than females at every education level, and the differences are fixed. More specifically speaking, those differences are floating around the difference of the y-intersection of the regression lines of male and female wages. From Figure 12, we can also see that most of the extreme values in each *highest years of school completed* (X) subgroup are reported by male participants. On average, males have higher weekly wages than females since there are more high rankers of males than females in each subgroup. The same conclusion can also be conducted by reviewing the relative position of the two regression lines. The regression lines of males and females are almost parallel to each other, and the regression line of male is located higher.

In short, the expected differences between estimated weekly wages for males and females are consistent with the graphical traits observed in Figure 12

Question 15:

generate college\_degree = 0

replace college\_degree = 1 if educ >=16 & educ!= .

Question 16:

tabulate college\_degree, freq

See output results in Figure 13.

Question 17:

robvar weekly\_wage, by(female), if college\_degree == 1

ttest weekly\_wage if college\_degree == 1, by (female)

Levene’s test:

The results of Levene’s test (see Figure 14, W0= 12.731852, df(1, 92), p= 0.00045403) demonstrate that the homogeneity of variance assumption of this dataset is not tenable.

Unnecessity of Levene’s test: Violations of the homogeneity of variance assumption may be essentially ignored as long as the samples being used have equal or approximately equal sizes. In this dataset, the female group has a simple size of 99 participants while the male group has 95, it’s approximately equal, so the results of Levene’s test can be ignored.

According to the results of independent sample t-test presented in Figure 15, we can state that among those who earned a college degree, male participants earn statistically significant higher weekly wages (t(1,192)=4.0711, p < 0.0001) than female participants.

Question 18:

robvar weekly\_wage, by(female), if college\_degree == 0

ttest weekly\_wage if college\_degree == 0, by (female)

The normality assumption is tenable because each sub-group has a population larger than 30.

The homogeneity of variance assumption is tested using Levene’s test. According to the results presented in Figure 16, we can see that according to Levene’s test (W0=36.298614, df(1, 374), p<0.0001), the homogeneity of variance assumption is not tenable. But since the samples being used have approximately equal sizes, the results of Levene’s test can be ignored.

Based on the results of independent sample t-test presented in Figure 17, we can state that among those who do not earn a college degree, male participants have statistically significant higher weekly wages (t(1,374)=9.0510, p<0.0001) than female participants.

Question 19:

The key finding of this research is that female earn less than male even with the same level of education. The present study starts with descriptive statistics (see results in Figure 2), finding that on average male participants included in this study earn approximately $200 more (mean = 519.336) than female participants (mean = 311.4533). The highest wage in this study is reported by male participants, and it is approximately $455 higher than highest wage reported by female. But at the same time, there’s no significant difference between years of education of male and female participants. The boxplot of weekly wage (Figure 3) carries out similar information, the weekly-wages of male group have a larger range than t female group. The median and the interquartile ranges of male’s weekly-wages are also larger than female. The male and female group have similar number of outliners, but those of the male group spread more separately higher positioned than those of female group. The maximal number of the male group is larger than the female group.

The correlation result of *female* and *weekly\_wage* (see Figure 6) further strengthens the primary results of the descriptive statistics. The correlation coefficient between the variable of *female* and *weekly\_wage* is -0.3593 (N=570), indicating a moderate negative correction between sexuality and weekly wage they earn. The result indicates that most of the female participants in this study earn lower wages than male participants. Similar results can also be observed from the linear regression (see Figure 12). The regression line shows that males get more wages than females at every education level, and the differences are fixed. More specifically speaking, those differences are floating around the difference of the y-intersection of the regression lines of male and female wages. Most of the extreme values in each highest year of school completed (X) subgroup are reported by male participants. On average, males have higher weekly wages than females since there are more high rankers of male than female in each subgroup. The same conclusion can also be conducted by reviewing the relative position of the two regression lines. The regression lines of males and females are almost parallel to each other, and the regression line of male is located higher.

The results of independent sample t-test (see Figure 15 & Figure 17) indicate that regardless the level of education, male participants earn statistically significant higher weekly wages (t(1,374)=9.0510, p<0.0001) than female participants(without college degree: t(1,374)=9.0510, p<0.0001; with a college degree or higher: t(1,192)=4.0711, p < 0.0001).

Conclusion: Based on those results discussed above, it is obvious that female participants in the current study earn lower wages than their male counterparts regardless of their education level. It seems to be a structural issue since this trend of unequal distribution appears in every level of education. So, more attention should be drawn to the unequal payment associated with sexuality.

Limitations: The sample size of the current study is only 570, making it difficult to generalize the conclusions of this study to a wider population. It also should be noted that the *income* variable relies on participants’ self-report, which will undoubtedly undermine the reliability of the current study.

**Appendix**

屏幕的截图

描述已自动生成

Figure 1

文本, 表格

描述已自动生成文本, 表格

描述已自动生成

Figure 2



Figure 3



Figure 4

屏幕上有字

描述已自动生成

Figure 5

图片包含 表格

描述已自动生成

Figure 6



Figure 7



Figure 8

屏幕上有字

描述已自动生成

Figure 9

屏幕上有字

描述已自动生成

Figure 10

屏幕上有字

描述已自动生成

Figure 11



Figure 12

表格

描述已自动生成

Figure 13

图片包含 日历

描述已自动生成

Figure 14

屏幕上有字

中度可信度描述已自动生成

Figure 15

文本

描述已自动生成

Figure 16

文本

描述已自动生成

Figure 17