Link to Github

https://github.com/Runshi-Yang/JSC270_HW2_2022_RunshiYang.git

Initial data exploration

1. age, fnlwgt, education-num, capital-gain, capital-loss and hours-per-week are described as continuous in this text file description of the data, they are expected to be of type float64 instead of int64. All of the other columns are the expected data types based on their descriptions.
2. There are 1836 missing values in column 'workclass', 1843 missing values in column 'occupation', 583 missing values in column 'native_country'. Other columns do not have missing values.
3. I think these variables should be transformed to categorical variables, since most of (more than 90%) of the capital gain or capital loss are 0, it is reasonable to group them into one group and let the others in one group.
4. I ploted the distribution of fnlwgt. The variable fnlwgt is not symmetrically distributed, but positively skewed. The distribution of this variable between men and women are very similar, both of them are positively skewed, and the modes are all at around 0.2*10^6. But there are about twice as many male as female with the same final weight. The outliers should be excluded, since there are 992 out of 32561 outiers, removing them will not affect the number of observasions greatly and will increase the accuracy of the models building on this dataset.

Correlation

1. education_num and hours_per_week appear to be correlated. I made my assessment since their correlation coefficient is much larger than 0.
2. The correlation between education_num and age is 0.148 and its significance is 0, which indicates strong evidence against the null hypothesis. The direction and significance of my finding are as expected since people with more education will find jobs more easily.
3. The correlation is 0.06 (and its significance is 0) between education_num and age for male and the correlation is -0.018 (and its significance is 0.063) between education_num and age for female. This is as expected since the older men are, the earlier they are educated and the more education they receive. And because of the prevalence of gender equality, women living in the 21st century have become more educated, so younger women receive more education.
4. The covariance between education_num and hours_per_week is 4.7. So the education_num and hours_per_week tend to show similar behavior (greater values of one variable mainly correspond with the greater values of the other variable).

Regression

a. Yes, since β1=6.0117 indicates that men works 6.0117 hours more than women on average.

b. The trend in hours worked by men vs women remain the same and the coefficient for education_num is statistically significant. The 95% confidence interval is [5.697, 6.245] for sex and [0.647, 0.748] for education_num.

c.

(i) The coefficient for sex is the estimated expected difference in number of working hours per week compare male with female (positive value indicates that male work for more hours).

(ii) The coefficient for sex is the estimated expected difference in number of working hours per week compare male with female holding education_num constant. (positive value indicates that male work for more hours).

(iii) The coefficient for sex is the estimated expected difference in number of working hours per week compare male with female holding education_num and gross_income_group constant. (positive value indicates that male work for more hours).

Statistics R-squared and adjusted R-squared can help to decide which model is the “best”. The R-squared and adjusted R-squared for model (i) (0.053) is less than that of model (ii) (0.074) and less than that of model (iii) (0.094), so (iii) fits the data better than (ii) better than (i).

Bonus: The slope parameter in simple linear regression is

$$\begin{aligned}
\beta_1 &= \frac{\sum(x-\overline{x})(y-\overline{y})}{\sum(x-\overline{x})^2} \\
&= \frac{\sum(xy-\overline{x}y-x\overline{y}+\overline{xy})}{\sum(x^2-2x\overline{x}+\overline{x}^2)} \\
&= \frac{\sum xy-\overline{x}\sum y-\overline{y}\sum x+n\overline{xy}}{\sum x^2-2\overline{x}\sum x+n\overline{x}^2} \\
&= \frac{\sum xy-2n\overline{xy}+n\overline{xy}}{\sum x^2-2n\overline{x}^2+n\overline{x}^2} \\
&= \frac{\sum xy-n\overline{xy}}{\sum x^2-n\overline{x}^2} \\
&= \frac{\sum xy-\frac{n\overline{x}n\overline{y}}{n}}{\sum x^2-\frac{n\overline{x}n\overline{x}}{n}} \\
&= \frac{\sum xy-\frac{(\sum x)(\sum y)}{n}}{\sum x^2-\frac{(\sum x)^2}{n}} \\
&= \frac{n\sum xy-\sum x\sum y}{n\sum x^2-(\sum x)^2}
\end{aligned}$$

The sample correlation coefficient is:

$$r_{xy} = \frac{n\sum xy-\sum x\sum y}{\sqrt{n\sum x^2-(\sum x)^2}\sqrt{n\sum y^2-(\sum y)^2}}$$

So

$$\begin{aligned}
\beta_1 &= r_{xy}\times\frac{\sqrt{n\sum y^2-(\sum y)^2}}{\sqrt{n\sum x^2-(\sum x)^2}} \\
&= r_{xy}\frac{SD(Y)}{SD(X)}
\end{aligned}$$