JSC270 – Assignment 2

Cindy (Shih-Ting) Huang

Github

https://github.com/cindyellow/JSC270_Assg2

Initial Data Exploration

1. Yes, the columns have the corresponding expected data types. The continuous variables have type int64 while the categorical variables have type object in the given data.
2. Missing values are represented by "?" in this data. *Workclass* has 1836 missing values, *occupation* has 1843, and *native_country* has 583.
3. These variables should be transformed because both are heavily skewed to the left. Taking the log or transforming them into binary variables will create a more even distribution.
4. The data for fnlwgt is not symmetrically distributed; there are more values between 0-0.2*1e6 than 0.4-0.6*1e6. Looking at both the boxplot and histogram for women and men's final weight, we see that they share approximately the same mean at around 0.2*1e6. There seems to be more variance for men's final weight as there are more data points lying in and beyond the third quartile. I don't think outliers should be excluded because there seem to be a significant amount of them, suggesting that perhaps there is another category of the population unaccounted for.

Correlation

1.
    a. There doesn't seem to be a strong correlation between these variables. The strongest correlation occurs between *education_num* and *hours_per_week* with a correlation coefficient of 0.148123.
    b. We test the variable pair *education_num* and *hours_per_week.* By fitting a linear regression model setting *education_num* as the independent variable and *hours_per_week* as dependent, we can see from the summary table that there is a t-statistic value of 27.026 and P>|t| value of 0.00. From this, we can reject the null hypothesis that there is no correlation between the two variables. Our coefficient 0.7109 indicates that with 1 unit change in *education_num*, *hours_per_week* changes by 0.7109 units. This seems reasonable since individuals with higher level of education tend to have jobs that are more challenging, thus increasing hours worked per week.
    c. The correlation between *age* and *education_num* is -0.0179 for female and 0.06049 for males. Considering the time period, this is expected. There's a negative correlation for females possibly because older women were less likely to have pursued higher level of education. This is due to past social limitations, such as gender roles constraining them to being housewives. On the other hand, younger women during 1994

had more opportunities to enter post-secondary institutes.
For men, there is a slightly positive correlation because as age increases, men are more likely to have completed more years of education. For instance, an 18-year-old boy would not have been able to achieve 13 years of education in comparison with a 30-year-old.

d. By multiplying *fnlwgt* with *education_num* and *hours_per_week* respectively, we observe that the weighted variance and covariance of the latter two variables have drastically increased. Since the unweighted values are small, this means that the samples with small values for *education_num* and *hours_per_week* in the original data are overrepresented.

Regression

1.

a. Since the intercept = 36.4104 is the average mean of hours worked per week for women and 36.4104+6.0177 for men, yes men do tend to work more hours.
b. Although the difference is lowered, the general trend that men work more hours still holds with this additional variable. *Education_num* is significant with a coefficient of 0.6975, indicating that weekly work hours increase with it.
c. We see that *gross_income_group* also has a strong positive relationship with *hours_per_week.* The coefficient for intercept lies between its respective values in model 1 & 2, whereas that for *sex* (male linear relationship) has decreased. The coefficient for *education_num* also decreased. To decide which model is the best, we look at R-squared, which measures the proportion of variability in the dependent variable that can be explained with the independent variables. Out of the three models, model 3 has the highest R-squared value and a reasonable standard error, so it is potentially suitable for our data. We can re-do this with a model fitting procedure by automating the process of adding independent variables and keeping them if the R-squared value increases and standard error is still below a certain threshold.

Bonus Question

- The correlation coefficient represents the strength of relationship between the change in two variables. The closer the estimator of regression slope coefficient is to 0, the less linearly related X & Y are, corresponding to a correlation coefficient that is also close to 0.