There are 15 questions worth 4 points each for a total of 60 points. You must answer all questions using whichever software you want, and submit your answers together with the rest of the quiz on the day of the quiz. It is acceptable to consult with other students, but each student should run their own regressions and submit their own individual report.

Download the dataset regarding the "Economic Value of a Year of Education" available following this link. The dataset contains cross-sectional data for approximately 3,000 individuals for the following variables:

\[ a_{hei} = \text{average hourly wage for the } i^{th} \text{ individual} \]
\[ yrseduc = \text{years of schooling for the } i^{th} \text{ individual} \]
\[ a_{sex} = \text{gender of the } i^{th} \text{ (1 for male, 2 for female)} \]
\[ a_{age} = \text{age of the } i^{th} \text{ individual} \]

1. Plot an histogram of the distribution of hourly wages.

![Figure (1)](image)

Distribution of hourly wages

2. Create a new categorical variable \( \text{deg}_i \) that takes the value 0 for individuals who did not finish high school (less than 12 years of schooling), the value 1 for who finished high school but not college (at least 12 and less than 16 years of schooling), and the value 2 for individuals who completed college (at least 16 years of schooling).

3. What proportion of individuals falls into each of these categories?

4. What is the average hourly wage for individuals in each of these categories?

5. Make a box-plot showing the distribution of hourly wages for these different categories of individuals. What can you tell from this box-plot?

<table>
<thead>
<tr>
<th>Maximum degree</th>
<th>Some HS</th>
<th>Completed HS</th>
<th>College or more</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number</td>
<td>193</td>
<td>1792</td>
<td>1004</td>
</tr>
<tr>
<td>Proportion</td>
<td>6.46%</td>
<td>59.95%</td>
<td>33.59%</td>
</tr>
<tr>
<td>Average wage</td>
<td>11.459</td>
<td>16.496</td>
<td>24.306</td>
</tr>
</tbody>
</table>
6. What is the proportion of females in the sample?
7. What is the average hourly wage for males and females?
8. Make a box-plot showing the distribution of hourly wages for males and females. What can you tell from this box-plot?

<table>
<thead>
<tr>
<th>Gender</th>
<th>Female</th>
<th>Male</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number</td>
<td>1331</td>
<td>1658</td>
</tr>
<tr>
<td>Proportion</td>
<td>44.53%</td>
<td>55.47%</td>
</tr>
<tr>
<td>Average wage</td>
<td>17.105</td>
<td>20.150</td>
</tr>
</tbody>
</table>

9. Plot a scatterplot of years of schooling vs. hourly wages.
10. Use OLS to estimate the following model. You must report both the estimated coefficients and the corresponding standard errors.

\[
\hat{ahe}_i = \beta_0 + \beta_1 \text{yrseduc}_i + \beta_2 \text{a_sex}_i + \beta_3 \text{a_age}_i + \varepsilon_i
\]

\[
\hat{ahe}_i = -30.16 + 1.885 \cdot \text{yrseduc}_i - 4.312 \cdot \text{a_sex}_i + 0.991 \cdot \text{a_age}_i
\]

11. With 95% confidence, which regressors have a statistically significant effect on hourly wages? [justify your answer]
   
   For all coefficients we have \( t = \hat{\beta}/SE(\hat{\beta}) > 2.90 \), which results in \( p \)-values well below 0.05. Therefore, all regressors appear to be significant.

12. What is the value of the adjusted \( R^2 \) coefficient? What does this number tell you for this particular model?

   \( R^2 = 0.202 \). Which means that we can explain approximately 20% or the heterogeneity in income by looking at differences in gender, age and years of schooling.

13. What is the estimated average effect of years of schooling on hourly wages?

   \( \hat{\beta}_1 = 1.885 \).

14. According to the estimated model, what would be the average hourly wage of a 30 years old female who completed a 2-years master degree after finishing college?
\[ \hat{a}_{he} = -30.16 + 1.885 \times 18 - 4.312 \times 1 + 0.991 \times 30 = 29.188 \]

15. According to the estimated model, what is the average effect over hourly wages of attending college?

\[ \Delta \hat{a}_{he} = \hat{\beta}_1 \times \Delta \text{yrseduc} = 1.885 \times 4 \approx 7.54 \]
The STATA commands that I used:

```stata
use data.dta
hist ahe

gen deg = 0
replace deg = 1 if(yrseduc>=12)
replace deg = 2 if(yrseduc>=16)

sum ahe if(deg=0)
sum ahe if(deg=1)
sum ahe if(deg=2)
graph box ahe, over(deg)

sum ahe if(a_sex=1)
sum ahe if(a_sex=2)
graph box ahe, over(a_sex)

scatter ahe yrseduc
reg ahe yrseduc a_sex a_age
```

The output from the regression that I got is:

```
               | Coef.  Std. Err.   t    P>|t|    [95% Conf. Interval]
---------------+-------------------+-------------------+-----+-------------------+-------------------+-------------------+
 ahe           |                   |                   |     |                   |                   |                   |
 yrseduc       | -2.681508         | 1.393991          | -1.92| 0.055             | -5.424424         | .0614073           |
 a_sex         | -3.702419         | .154011           | -24.04| 0.000             | -4.005491         | -3.399348          |
 a_age         | .1086104          | .090719           | 1.20 | 0.232             | -.067947          | .2871154           |
 _cons         | 906.7392          | 28.25605          | 32.08| 0.000             | 851.1228          | 962.3555           |
---------------+-------------------+-------------------+-----+-------------------+-------------------+-------------------+
```

```
Predictor | Coef  SE Coef | T    P
Constant   | -30.16 10.03  | -3.01| 0.003
a_age      | 0.9911 0.3383 | 2.93 | 0.003
a_sex      | -4.3118 0.3437| -12.54| 0.000
yrseduc    | 1.88468 0.07276| 25.90| 0.000

S = 9.24512  R-Sq = 20.3%  R-Sq(adj) = 20.2%
```