Sample Midterm Exam

The dataset WAGE2.dta contains wage (monthly earnings) information for a random sample of 935 individuals. Please clearly mark/number your answers while answering the following questions.

1. Setup the working environment as follows:
   1.1. Add your name and last name (commented out) in the beginning of your code
   1.2. Define global variables for where you save the "data" and where you save your "output" (or log file). Note that these two globals might point to the same location/path on your computer if you prefer to do so.
   1.3. Start a log file in the "output" path.
   1.4. Open the dataset. Review the list of available variables and their labels.
2. Variables *urban* and *married* are dummy variables. For example, *urban* takes a value of 1 if the individual lives in an urban area, but a value of 0 otherwise. If you randomly selected one person from this sample, what is the probability of drawing someone who is married conditional on having drawn someone who lives in an urban area?
3. Test the null hypothesis that average monthly earnings is equal to $1,000 at the 1% significance level (alpha) with the alternative that it is smaller than $1,000 (i.e., one-sided hypothesis testing). Insert a comment with your conclusion. Your conclusion can depend on the t-statistic, the p-value, or the confidence interval in the Stata output.
4. Calculate the 95 percent confidence interval for the average monthly earnings (*wage*). The Stata command for calculating confidence interval is "ci." Add a comment on the meaning of this confidence interval.
5. Test if the monthly earnings in urban areas are statistically different from monthly earnings in non-urban areas at the 1% significance level. Add a comment with your conclusion.
6. Calculate the hourly rate from monthly earnings (*wage*) and average weekly hours (*hours*) assuming that there are exactly 4 work weeks in a month. Call this variable *hourly_rate*.
7. Create a new variable called *lhourly_rate* and set it equal to (natural) log of *hourly_rate*.
8. The data include information on years of education (*educ*). What are the two most common values for this variable? Do they correspond to anything meaningful? Add a comment with your answer.
9. Estimate the following regression equation:

$$lhourly\_rate = \beta_0 + \beta_1 educ + u$$

10. Comment on the statistical significance of each coefficient at the 5% significance level using the t-statistics.

11. Interpret the meaning of your coefficient estimate for *educ*. Note that your dependent variable is in logs.

12. How much of the variation in *lhourly_rate* is explained by *educ*? If the amount of variation explained by the model is small, what does that mean?

13. Estimate the following regression equation:

$$lhourly\_rate = \beta_0 + \beta_1 educ + \beta_5 urban + u$$

14. Is the coefficient estimate for *urban* statistically significant at the 5% significance level? Use the confidence interval in your answer/explanation.
15. Create a scalar with the **F-statistic** from the regression output. Add a comment about what it tells us (i.e., null hypothesis and the conclusion).
16. Conduct the **F-test** (using STATA commands) for the null hypothesis that the *urban* has no effect on hourly earnings using 1% significance level. Confirm that the square root of the F-statistic is (approximately) equal to the t-statistic for *urban* from the regression output.

17. If the estimate for the coefficient on *educ* in Question 13 and 9 are $\hat{\beta}_1$ and $\tilde{\beta}_1$, respectively, how do they compare? Why do we see a difference in these two estimates? Show the relationship between the two using appropriate regressions and their coefficient estimates.